



Universitat de Girona

FUNCIONS DENSITAT I SEMBLANÇA MOLECULAR QUÀNTICA: NOUS DESENVOLUPAMENTS I APLICACIONS

Xavier GIRONÉS TORRENT

ISBN: 84-688-1293-5

Dipòsit legal: GI-425-2002

<http://hdl.handle.net/10803/8023>

ADVERTIMENT. L'accés als continguts d'aquesta tesi doctoral i la seva utilització ha de respectar els drets de la persona autora. Pot ser utilitzada per a consulta o estudi personal, així com en activitats o materials d'investigació i docència en els termes establerts a l'art. 32 del Text Refós de la Llei de Propietat Intel·lectual (RDL 1/1996). Per altres utilitzacions es requereix l'autorització prèvia i expressa de la persona autora. En qualsevol cas, en la utilització dels seus continguts caldrà indicar de forma clara el nom i cognoms de la persona autora i el títol de la tesi doctoral. No s'autoritza la seva reproducció o altres formes d'explotació efectuades amb finalitats de lucre ni la seva comunicació pública des d'un lloc aliè al servei TDX. Tampoc s'autoritza la presentació del seu contingut en una finestra o marc aliè a TDX (framing). Aquesta reserva de drets afecta tant als continguts de la tesi com als seus resums i índexs.

ADVERTENCIA. El acceso a los contenidos de esta tesis doctoral y su utilización debe respetar los derechos de la persona autora. Puede ser utilizada para consulta o estudio personal, así como en actividades o materiales de investigación y docencia en los términos establecidos en el art. 32 del Texto Refundido de la Ley de Propiedad Intelectual (RDL 1/1996). Para otros usos se requiere la autorización previa y expresa de la persona autora. En cualquier caso, en la utilización de sus contenidos se deberá indicar de forma clara el nombre y apellidos de la persona autora y el título de la tesis doctoral. No se autoriza su reproducción u otras formas de explotación efectuadas con fines lucrativos ni su comunicación pública desde un sitio ajeno al servicio TDR. Tampoco se autoriza la presentación de su contenido en una ventana o marco ajeno a TDR (framing). Esta reserva de derechos afecta tanto al contenido de la tesis como a sus resúmenes e índices.

WARNING. Access to the contents of this doctoral thesis and its use must respect the rights of the author. It can be used for reference or private study, as well as research and learning activities or materials in the terms established by the 32nd article of the Spanish Consolidated Copyright Act (RDL 1/1996). Express and previous authorization of the author is required for any other uses. In any case, when using its content, full name of the author and title of the thesis must be clearly indicated. Reproduction or other forms of for profit use or public communication from outside TDX service is not allowed. Presentation of its content in a window or frame external to TDX (framing) is not authorized either. These rights affect both the content of the thesis and its abstracts and indexes.

Memòria per a la defensa i obtenció del
grau de Doctor de:

Xavier Gironés Torrent

**Funcions Densitat i Semblança
Molecular Quàntica: Nous
Desenvolupaments i Aplicacions**



Universitat de Girona

Febrer 2002

Els sotasignats Prof. Ramon Carbó-Dorca Carré, Catedràtic d'Universitat del Departament de Química de la Universitat de Girona, i Dr. Emili Besalú LLorà, Titular d'Universitat del Departament de Química de la Universitat de Girona

CERTIFIQUEN:

Que en Xavier Gironés, llicenciat en Ciències, secció de Químiques, ha realitzat sota la nostra direcció, a l'Institut de Química Computacional i al Departament de Química d'aquesta Universitat, el treball titulat "Funcions Densitat i Semblança Molecular Quàntica: Nous Desenvolupaments i Aplicacions" que es troba recollit en aquesta memòria i que es presenta en pública defensa per optar al grau de Doctor en Ciències Químiques.

I perquè consti als efectes legals escaients, signem aquest certificat a Girona, a 1 de febrer de 2002.

R. Carbó-Dorca

E. Besalú

If the stars should appear one night in
a thousand years, how would men believe
and adore, and preserve for many generations
the remembrance of the city of God?

Emerson (de *The Complete Stories II*,
Asimov, Harper Collins Publishers).

A Carla Solà
(1976-2000)

Índex de Continguts

1	Agraïments	1
2	Introducció General i Objectius	7
2.1	Creació de Contorns Moleculars de IsoDensitat Electrònica	8
2.2	Sobreposició Molecular	9
2.3	Relacions Quantitatives Estructura-Propietat (QSAR)	10
2.4	Estructura del present treball	12
3	Isocontorns de Densitat Electrònica Molecular	15
3.1	Introducció	15
3.2	Obtenció de Funcions Densitat: Metodologies	22
3.2.1	<i>Mecànica Molecular</i>	23
3.2.2	<i>Mecànica Quàntica</i>	24
3.2.3	<i>Funcions Densitat Ajustades: ASA i ASA Promolecular</i>	26
3.2.4	<i>Funcions Densitat Ajustades i Semblança Molecular Quàntica</i>	29
3.3	Obtenció de MIDCOs	32
3.3.1	<i>Obtenció de les coordenades moleculars</i>	33
3.3.2	<i>Creació de la funció densitat</i>	33
3.3.3	<i>Creació del Grid</i>	34
3.3.4	<i>Càlcul del valor de la funció</i>	35
3.3.5	<i>Reconstrucció i representació de la superfície: El MCA</i>	37
3.4	Exemples Gràfics	38
3.4.1	<i>Comparativa de MIDCOs</i>	38
3.4.2	<i>Isocontorns PASA per a molècules grans</i>	45
3.4.3	<i>Isocontorns PASA per a macromolècules</i>	47
3.4.4	<i>Isocontorns PASA per a sistemes interaccionants</i>	49
3.5	Refinaments del mètode	51
3.5.1	<i>Particionament de l'espai</i>	51
3.5.2	<i>Definició de variables-àtom</i>	54
3.5.3	<i>Càlcul de la densitat</i>	56
3.5.4	<i>Exemple gràfic</i>	57
3.5.5	<i>Exemples macromoleculars</i>	58
3.6	Conclusions	63
3.7	Contribucions i Articles Adjunts	63
3.8	Referències	65
3.9	Llistat d'Articles Adjunts	69

4	Relacions Quantitatives Estructura-Activitat (QSAR)	95
4.1	Importància del QSAR	95
4.2	Del Segle XIX al XXI, breu cronologia del QSAR	96
4.2.1	<i>Els Inicis</i>	96
4.2.2	<i>Les Relacions d'Energia Lliure</i>	97
4.2.3	<i>Anàlisi de Hansch</i>	102
4.2.4	<i>Estudis Topològics</i>	104
4.2.5	<i>Mètodes Espaials</i>	105
4.2.6	<i>El disseny molecular assistit per ordinador</i>	105
4.3	Sobreposició en QSAR	107
4.4	Tècniques estadístiques d'ajust	107
4.4.1	<i>Introducció</i>	108
4.4.2	<i>Regressió multilíneal</i>	109
4.4.3	<i>Anàlisi de Composants Principals</i>	109
4.4.4	<i>Mínims quadrats parcials</i>	111
4.4.5	<i>Xarxes Neurals Artificials</i>	113
4.4.6	<i>Altres metodologies</i>	114
4.5	Validació d'un model QSAR	115
4.5.1	<i>Validació interna</i>	115
4.5.2	<i>Validació externa</i>	120
4.6	Tria de Descriptors en els models QSAR	121
4.7	Referències	123
5	Algorisme de Sobreposició Topo-Geomètrica (TGSA)	129
5.1	Introducció	129
5.2	Metodologies de sobreposició	131
5.2.1	<i>Aproximacions manuals</i>	131
5.2.2	<i>Algorismes genètics</i>	132
5.2.3	<i>Tècniques atom-based</i>	133
5.2.4	<i>Tècniques field-based</i>	134
5.2.5	<i>Altres tècniques</i>	135
5.3	Sobreposició topo-geomètrica	138
5.3.1	<i>Introducció</i>	138
5.3.2	<i>Metodologia</i>	140
5.3.3	<i>Exemple il·lustrat: el programa TGSA2001</i>	142
5.3.4	<i>Altres exemples gràfics</i>	149
5.4	Conclusions	153
5.5	Contribucions i articles adjunts	154
5.6	Referències	155
5.7	Article Adjunt	159
6	La Semblança Molecular Quàntica	171
6.1	El concepte de semblança	171

6.2 Semblança Molecular Quàntica	172
6.3 Formulació matemàtica de la semblança molecular quàntica	173
6.3.1 <i>Funcions densitat</i>	173
6.3.2 <i>Definició general</i>	175
6.3.3 <i>Tipus de mesures de semblança molecular quàntica</i>	175
6.3.4 <i>Normalització de les mesures de semblança</i>	178
6.3.5 <i>Dependència de la sobreposició molecular en les mesures de semblança</i>	180
6.4 Àmbits d'aplicabilitat	181
6.5 Semblança Molecular Quàntica en QSAR	182
6.6 Exemples d'aplicació I: matrius de semblança	183
6.6.1 <i>Modelatge Molecular</i>	184
6.6.2 <i>Càlculs de MQSM i Exemples</i>	184
6.7 Exemples d'aplicació II: V_{ee} com a descriptor molecular	191
6.7.1 <i>Introducció</i>	192
6.7.2 <i>V_{ee} i semblança molecular quàntica</i>	192
6.7.3 <i>Construcció de models lineals senzills amb V_{ee}</i>	193
6.7.4 <i>Modelatge molecular</i>	194
6.7.5 <i>Interpretació de V_{ee}</i>	194
6.7.6 <i>Exemple: V_{ee} com a descriptor de la refractivitat molar</i>	196
6.7.7 <i>Exemple: V_{ee} com a descriptor de la inhibició de creixement</i>	199
6.7.8 <i>Exemple: V_{ee} com a descriptor de toxicitat</i>	201
6.8 Exemple d'aplicació III: Semblances sobre densitats de Forats de Fermi	203
6.8.1 <i>Funcions Densitat de Forats de Fermi</i>	203
6.8.2 <i>Mesures de Semblança sobre Densitats de Forats de Fermi</i>	206
6.8.3 <i>Exemple d'aplicació: Àcids Aromàtics</i>	207
6.9 Conclusions	210
6.10 Contribucions i articles adjunts	211
6.11 Referències	213
6.12 Taules Adjunes	223
6.13 Llistat d'Articles Adjunts	231
7 Consideracions finals	327
7.1 Conclusions Generals	327
7.2 Altres Aspectes Estudiats	328
7.3 Objectius de Futur	329
7.4 Llistat Total de Publicacions	330

1 Agraïments

Basant-me en el precedent que va *assentar* el meu company Josep Nadal en el seu treball de recerca, posaré els agraïments al principi d'aquest treball ja que, i cito textualment, *acostumen a ser les úniques pàgines llegides*. I per què li dedico un capítol sencer? Doncs perquè crec que dóna idea de que una tesi doctoral, tot i ser el colofó de la carrera acadèmica d'un estudiant, no pot ser de cap de les maneres el treball d'una única persona.

Resulta molt difícil recollir, i sobretot expressar, en un treball com aquest el conjunt de fets i persones a qui agrair. Ja no només la gent en sí (ai! si em deixo a algú), sinó en l'ordre en que els cites. D'altra banda, també pot resultar costós agrair per compromís o per tradició a aquells que tan sols han tingut una influència trivial, si no negativa, en el desenvolupament d'aquest treball. Sigui com sigui, intentaré com sempre *fer el millor de mi mateix*, i citaré a tots aquells a qui crec que han fet possible o ajudat a que aquest treball no tan sols comencés, sinó que arribés a la seva fi.

D'altra banda, cal comentar que durant aquest període no només m'he dedicat a la química computacional, sinó que també he expandit el meu coneixement a d'altres àrees (segons com es miri) properes, no exactament reflectides en aquest treball, com pot ser la configuració i reparació d'ordinadors, creació de pàgines web, tractament d'imatge i d'altres; i algunes d'aquestes especialitats m'han estat en alguns casos també útils¹. Totes aquelles persones, sovint anònimes, que m'han donat un cop de mà en aquests aspectes es mereixen, a falta de res millor, una menció.

Bé doncs, deixant ja aquestes petites consideracions preliminars enrere, permeteu-me que expressi el meu més sincer agraïment a:

¹ Valgui dir que en alguns casos també m'ha resultat ser una càrrega feixuga.

- **Els meus pares i família en general.** De fet, de no ser per ells el treball ja no s'hagués dut a terme mai. No els vull agrair el fet de que em portessin al món, sinó el seu amor, suport moral i econòmic i, en molts casos, paciència sobrehumana que han mostrat vers mi, sobretot en termes de *política educativa*, que ara passaré a explicar, que en el fons més els agraeixo.

Poc després d'acabar la EGB, jo ja volia ser químic, però més enllà de voler tenir una carrera, volia ser doctor. El dia que vaig comunicar-ho als meus pares, amb la corresponent càrrega econòmica que suggeria fer una carrera i un doctorat de nou anys en total, la resposta va ser tant ràpida com senzilla: 'tu fes el que consideris que hakis de fer, nosaltres hi farem el que podem'. Aquesta frase, senzilla en forma, però farcida de sentiment, no tan sols fou dita, sinó que fou duta a la pràctica fins a les últimes instàncies, excedint-se 'con creces'. Amb això vull dir que no només em van finançar, a fons perdut, els estudis i material estrictament necessari, sinó ordinadors, llibres complementaris, cotxe per a transportar-me a Girona (amb els corresponents costos afegits de gas-oil, canvis d'oli, revisions, ...). Al demanar, la resposta també era senzilla: 'Si et cal ...', i tal dit, tal fet.

Aquest comportament, que fins i tot va arribar-me a resultar obvi per costum, va resultar no ser tant comú (fins i tot rar), en companys del meu entorn. Llevat dels casos en que les possibilitats econòmiques de les famílies d'aquests companys eren limitades, he pogut veure que en ocasions la família no recolza el progrés acadèmic/educatiu, limitant així al company. Frases del tipus: 'per què necessites un ordinador nou? Que el que tens ara (un 286 de 6 anys) ja no funciona?' o bé 'Sí molt compraràs un llibre de 15000 pessetes!' les he sentit diverses vegades. No negaré que en el meu cas particular, ha resultat difícil d'entendre als meus pares la necessitat de canviar un Pentium per un PentiumIII, o en el seu moment el 386 pel Pentium anterior, però finalment han pogut entendre que un ordinador no és com una televisió o un microones, els quals funcionen bé o no funcionen, sinó que són màquines que amb el temps esdevenen obsoletes per les noves possibilitats. El mateix raonament es pot aplicar a la compra d'altre material.

I com no, recordar que mai m'han mancat estímuls. Tan els meus pares com els meus germans m'han portat des de ben petit a llocs on, de forma divertida, podia aprendre. Això inclou museus, ruïnes, llibres antics d'història i, entre d'altres, in comptables anades al cine que tant han desenvolupat el meu gust personal cap a la ciència ficció. Aquesta última particularitat, lluny de ser trivial, la cito pels intents (no sempre correctes, però al final valoro molt més la intenció i la voluntat que no l'encert) que van fer els meus pares, i els meus germans, per a explicar a l'infant que era les preguntes que em sortien. Aquestes preguntes em són recordades quan estic amb els meus tres nebots, Ricard, Sergi i Jordi, als quals, la curiositat innata dels

infants els fa preguntar-se mil coses, les quals intento respondre² de la manera més fàcil, però concreta i veritable a la vegada, que puc.

Finalment, i per acabar amb la unitat familiar, agrair si més no, l'interès que altres parents m'han demostrat pel meu progrés.

- **Als companys de l'Institut de Química Computacional (IQC) i facultat en general.** La millor, i única manera viable de començar aquest apartat, és agrair als meus co-directors de Tesi, Emili Besalú (per fi, ja titular!) i Ramon Carbó-Dorca, el seu interès i entusiasme en la realització d'aquest treball. Sense les seves idees, aquesta tesi no hagués ni començat, i són aquestes la guia que m'han permès desenvolupar les meves pròpies. A tots dos, gràcies per un treball ben fet.

Però l'IQC no es limita a nosaltres tres. Hi ha molta més gent a qui dono les gràcies per resoldre'm aquells dubtes de feina o formals que m'han anat sortint. Però no voldria anomenar-los per això (ai altra vegada si me'n deixo algun!), sinó per altres motius que passaré a detallar ara mateix: Lluís Amat, per la seva constant 'tutela', tant per aspectes estructuralistes de programació com per disponibilitat del seu temps; Miquel Lobato, per introduir-me en el món del Fortran 90, i sobretot en aspectes en l'ús *eficient* d'Internet, i per donar-me tants bons consells que, per desgràcia, tant poc he seguit. David Robert, qui em va obrir els ulls a una manera pulcra i meticulosa a l'hora de treballar, així com per ser un, ja no company, sinó amic, amb qui sempre puc comptar. El qualificatiu d'amic també pot aplicar-se directament a Pedro Salvador (ara ja Dr. Pedro), amb qui tant hem compartit, parlat, i sobretot ser un dels pocs que ha sabut escoltar-me i *fotre'm canya* quan convé. La mateixa mesura pot aplicar-se a Josep M^a Luís, que també sap escoltar-me, i al igual que en Miquel Lobato, m'ha donat molts consells poc seguits; Sílvia Simon, amb qui tants farts de riure ens hem fet; Pilar del Acebo, amb qui acostumava a començar al matí a discutir (de bon rotllo) sobre qualsevol tema, i per excedir-se en les seves funcions de (ex-)secretària; Xavi Fradera, gràcies a qui molts anys he recordat el meu sant, i sobretot per descobrir-me el meravellós món del Linux; Montse Cases, pel e-mails més càlids que rebia quan estava en *terres fredes*, i com no, per a posar al límit la meva capacitat degustativa amb els seus pastissos i trufes (no existeix paraula gastronòmica prou elevada per a descriure-les); Miquel Duran, per aquells dubtes capciosos de pàgines web que tinc, i per aquells *enigmes informàtics* que em proposa; Miquel Solà, per haver-se deixat interrompre tant, i sovint reunit, per estar disposat aclarir-me qualsevol concepte; a Joan Miró, bàsicament li agraeixo ser tal com és, amb qui puc parlar des d'Star Trek fins a relativitat, des de lluites medievals fins al mètode de codificació de la màquina enigma, i des d'esferes Dyson a alineacions planetàries (si algú ha de substituir la frase de Sócrates 'només sé que no sé res'

² Citant a Einstein: 'No entens veritablement un fenomen fins que no ets capaç de fer-li entendre a la teva àvia'.

amb quelcom semblant a 'ja sé bastant de tot', jo voto per a tu); a la Carme, pel seu bon humor de bon matí (fins i tot en temps difícils); a en Pep Martí, per les discussions (també de bon rotllo) que teníem de tant en tant; a l'Anna Gallegos, per donar aquell toc d'*alegria femenina* al despatx (que durant 3 anys només vam ser nois!); a en Jordi Poater, pel seu recolzament en un moment difícil i per fer de *company de cafè suplent* quan cal; a en Miquel Torrent, per ajudar-me a completar la col·lecció mp3-os, i a qui sincerament recomano escurçar la longitud dels e-mails; a en David Hugas, per obrir-me els ulls al meravellós món del DivX; i a l'Albert Poater, per posar sempre a prova la meua paciència :-)

No voldria deixar-me de, com a mínim, mencionar a d'altres companys que han anat passant per l'IQC, com són Jordi Mestes, (gran consell: Estigués sempre bé amb la secretària, ja que d'ella en depens!), Marta Forés (qui *ojalà* tingui un nen ben sa), Albert Bach, Mariazel Torrent, Carles Martin, Sara Insa i Pere Constants. Així com visitants amb qui he tingut alguna relació, com Olav Foosgard, Leonel Vera, Mathias Biquelhaupt, Stephen Sauer (aquest va posar a prova el meu alemany!), Carlos Bunge, Jesús Vicente de Julián i la Greet Boon.

Però com que la vida acadèmica no es limita tampoc a l'IQC, he pogut comptar també amb la inestimable ajuda de companys d'altres departaments, molt sovint per a dubtes de formulació de compostos. A molts de vosaltres, moltes gràcies.

Una menció especial mereixen els *operates informàtics* de la facultat, l'Albert, Pep, Pere i Xavi, que, tot i que la seva feina és vetllar pel bon funcionament del maquinari, puc jurar que amb mi potser se'ls hauria d'haver pagat hores extres per la quantitat de dubtes i problemes, molts d'ells excedint la seva obligació, que m'han resolt, i fins i tot, solucionat per anticipat.

- **Professors on he fet estades.** Com no agrair a aquells que, estant jo fora de casa, em van tractar com si hi fos. En especial al prof. Robert Ponec, de la República Txeca, el qual va minimitzar els efectes de sentir-me analfabet en un país en el que realment ho era; deixant de banda els coneixements que vaig adquirir i la productivitat que va resultar d'aquella estada. També gràcies a Paul G. Mezey, que segurament va fer el millor que va poder durant la meua estada a Canadà (i qui fa el que pot ...).

- **La 'colla' de Santa Coloma.** La feina no hi ha manera de fer-la bé si no es desconnecta de tant en tant, i per això tinc els meus amics de Santa Coloma. Amb ells hem passat bones estones, i fins i tot, en algun cas, m'han ajudat en la recerca perquè quan s'intenta explicar un dubte a una persona *no entesa en a matèria*, un s'obliga a simplificar les explicacions fins a un punt en que la resposta sembla evident a qui m'escolta, o apareix per si sola al enfocar-ho des d'un altre punt de vista. Com no, anomenar a la Rosa i en Miki, la Berta i en Ciso, la Sandra i en Sala (a veure si ho arreglem), la Neus i en Joan, la M^a Àngels i en Leti, la Carla (per desgràcia ja no físicament amb nosaltres), la Marga, la Susana i l'Abel, la Marta i l'altre Abel, l'Eva i en

Berto, l'altra Sandra i l'Albert, en Putxi, en Tito, la Berta (no d'en Ciso, sinó germana de l'Albert), entre de bastants d'altres amb qui ara (no m'ho tireu en cara, que va de bona fe) no he caigut.

- **Altres 'ajudes'.** En aquest calaix dels mals endreços voldria anomenar a tots aquells a qui m'han ajudat, de manera sovint anònima, en el meu treball, bé sigui directament o indirecta. Per exemple, les respostes rebudes del *Computational Chemistry List*, o del *alt.prog.fortran*; o aquella gent (a qui encara no entenc) que permet que programari no accessible a la butxaca d'un becari pugui emprar-se sense limitacions :-)

- **Fonts de Finançament.** En aquest 'calaix, voldria citar aquelles entitats, les quals bé amb pessetes, €'s, material divers o hores de càlcul gratuïtes, han contribuït en aquesta recerca. Primer de tot la *Universitat de Girona* per concedir-me una beca pre-doctoral, així com finançar-me part de les estades de recerca. La *Fundació M^a Francisca de Roviralta*, per les seves generosos donacions a l'Institut en forma de material informàtic d'última generació. Al CESCO-CEPBA per les hores de CPU consumides. I finalment als organismes públics per les seves aportacions als projectes SAF 96-0158, ENV4-CT97-0508, SAF2000-0223-C03-01.

2 Introducció General i Objectius

Aquesta tesi doctoral, que porta per títol ‘Funcions Densitat i Semblança Molecular Quàntica: Nous Desenvolupaments i Aplicacions’, pretén ser el resum del treball realitzat durant quatre anys i mig a l’Institut de Química Computational de la Universitat de Girona, sota la co-direcció del Dr. Emili Besalú i el Prof. Dr. Ramon Carbó-Dorca, així com una estreta relació amb els companys de grup Lluís Amat, David Robert i Anna Gallegos, i per descomptat, també amb la resta de companys de l’Institut.

Aquest capítol introductori pretén, sense voler entrar en gaires detalls i pressuposant alguns conceptes, ser la guia per on es desenvoluparan la resta de capítols, dedicats exclusivament als temes tractats durant el temps de recerca a l’Institut, i en els quals s’expressarà detalladament la metodologia, així com els conceptes i teoria necessaris per a la total comprensió d’aquest treball.

Des de que vaig començar a treballar a l’Institut l’any 1997, s’han tractat diversos temes relacionats sempre amb l’eix principal de la recerca del grup de treball: les Mesures de Semblança Molecular Quàntica (MQSM, de l’anglès *Molecular Quantum Similarity Measures*).

Dins del gruix de la teoria de les MQSM, la meua recerca s’ha centrat majoritàriament en les derivacions definides en els següents tres àmbits:

- La creació de Contorns Moleculars de IsoDensitat Electrònica (MIDCOs, de l’anglès *Molecular IsoDensity COntours*) a partir de densitats electròniques ajustades.
- El desenvolupament d’un mètode de sobreposició molecular, alternatiu a la regla de la màxima semblança.
- Relacions Quantitatives Estructura-Activitat (QSAR, de l’anglès *Quantitative Structure-Activity Relationships*).

2.1 Creació de Contorns Moleculars de IsoDensitat Electrònica

La creació de MIDCOs fou inspirada pel Prof. Paul G. Mezey, amb qui posteriorment vaig realitzar una estada a la Universitat de Saskatchewan (Canadà) l'hivern de l'any 2000, en una visita que va fer l'any 1998. El seu treball consistia en representar gràficament la densitat electrònica de diferents sistemes moleculars a partir de fragments moleculars sobre els quals ja es disposava de la seva funció densitat a nivell *ab initio*; així com també s'havia calculat el valor d'aquesta funció en el seu entorn. El seu mètode, anomenat MEDLA (Molecular Electron Density Lego Approach, o bé Aproximació a la Densitat Electrònica per Blocs), permetia ajuntar aquestes peces pre-calculades presents en un sistema de majors dimensions, i obtenir la representació gràfica desitjada. Aquest mètode permetia obtenir figures de densitat electrònica per a sistemes de grans dimensions amb recursos computacionals modestos.

El nostre objectiu en aquest camp fou l'aplicació de funcions densitat ajustades, ideades inicialment per a abaratir els càlculs de MQSM, per a l'obtenció de MIDCOs. Així, es va realitzar un estudi gràfic comparatiu entre diferents funcions densitat ajustades a diferents bases amb densitats obtingudes de càlculs duts a terme a nivells més alts (*ab initio*). D'aquesta manera, l'analogia visual entre les funcions ajustades i les *ab initio* obtinguda en el ventall de representacions de densitat obtingudes, i juntament amb els valors de les mesures de semblança obtinguts prèviament, els quals eren totalment comparables a qualsevol nivell, es va poder acabar de fonamentar l'ús d'aquestes funcions densitat ajustades. A més a més, es va implementar programari, no només pel simple càlcul de la densitat, sinó per a la reconstrucció de les superfícies d'interès a partir de la xarxa de punts prèviament calculada, així com programari específic de representació de superfícies.

Ja més enllà del propòsit inicial, i conjuntament amb el Prof. Mezey durant l'estada a Canadà, es van poder realitzar dos estudis complementaris a la simple representació de densitats, i són l'anàlisi de curvatura i l'extensió a macromolècules. La primera observació correspon a comprovar no només la semblança dels MIDCOs, sinó la coherència del seu comportament a nivell de curvatura, podent-se així observar punts d'inflexió en la representació de densitats i veure gràficament aquelles zones on la densitat és còncava o convexa. Aquest primer estudi va revelar, com en part era d'esperar, que tant les densitats ajustades com les calculades a nivell *ab initio* es comportaven de manera totalment anàloga. En la segona part d'aquest treball, i aprofitant l'experiència de Mezey en representacions gràfiques d'isocontorns, es va poder estendre el mètode a molècules més grans, de fins aquell moment uns 2500 àtoms. Afegint el fet de que alguna de les molècules tractades havia estat estudiada prèviament per ell, es va poder realitzar una altra comparativa visual entre els resultats obtinguts amb densitats

ajustades i aquells resultants d'aplicar el seu mètode MEDLA. Aparentment, els contorns emprant el MEDLA poden ser obtinguts més ràpidament, degut a que el mètode pren contorns prèviament calculats i els reordena en l'espai per a formar el contorn final. Tanmateix, aquest mètode requereix que prèviament s'hagin calculat i emmagatzemat tots els fragments necessaris, fet que implica que la molècula ha de ser minuciosament examinada per a evitar emprar fragments que distorsionin la geometria molecular inicial. Fent servir funcions ajustades, l'única preocupació consisteix en tenir paràmetres per a tots els àtoms que intervinguin en el càlcul, els més comuns dels quals es troben ja calculats en diferents bases, i cas de trobar-se amb un no calculat, dur a terme l'ajust. D'aquesta manera, ambdues formes d'obtenir funcions densitat esdevenien, en termes de temps necessari per a finalitzar els càlculs, equivalents per a molècules petites, però l'ús de funcions ajustades disminuïa considerablement el temps global, incloent temps d'estudi previ, d'obtenció dels contorns.

Finalment, es va aplicar part de la filosofia del MEDLA en aquest camp. Sabent que la densitat electrònica decau ràpidament al allunyar-se dels nuclis, el càlcul d'aquesta pot ser obviat a distàncies grans d'aquests. D'aquesta manera es va proposar particionar l'espai, i calcular tan sols les funcions ajustades de cada àtom tan sols en una regió petita, la qual envolta l'àtom en qüestió. Dient a terme el procés per a cada àtom¹, no només es disminuïa el temps de càlcul, sinó que aquest esdevenia un procés lineal del nombre d'àtoms presents en la molècula tractada enlloc de ser un procés a la tercera potència.

2.2 Sobreposició Molecular

Conjuntament als estudis de creació de MIDCOs, un problema que s'anava fent patent en el marc de la semblança era el de la sobreposició molecular. En tractar series de molècules homogènies, les sobreposicions obtingudes eren prou coherents en el sentit de intuïció química, tanmateix, en incrementar la diversitat, o bé en tenir presents àtoms pesats en les estructures, el procés donava sobreposicions no intuïtives. El mètode emprat era la regla de la màxima semblança, la qual es pot enunciar com 'aquella sobreposició òptima serà aquella que proporcioni la màxima semblança entre ambdues molècules'. Tot i que aquest mètode complia rigorosament el seu propòsit buscant exhaustivament, i de manera computacionalment costosa, aquella orientació relativa que donava aquest màxim, els resultats quan es donaven les anteriorment dites circumstàncies eren poc coherents des del punt de vista intuïtiu.

Per això, el següent objectiu fou crear un algorisme, així com la seva implementació en forma de programa, batejat TGSA, d'un mètode que proporcionés aquells alineaments que hom esperaria si tal tasca es pogués realitzar manualment. El resultat d'aquesta recerca fou un programa informàtic, codificat en Fortran 90, el qual alineava les molècules per parelles considerant tan sols nombres i distàncies atòmiques. La total absència de paràmetres teòrics va permetre desenvolupar un mètode de sobreposició molecular general, tot i que fins a la data (i al meu coneixement) tan sols s'ha emprat en semblança (però no tots en el nostre grup de recerca), que proporcionés una sobreposició intuïtiva, i també de forma rellevant, de manera ràpida i amb poca intervenció de l'usuari.

Tot i que en continua revisió per a la inclusió de nous conceptes i millora de la interfície gràfica, absent en la versió original, el programa no tan sols és capaç de sobreposar conjunts de molècules per parelles, sinó de prendre'n una de plantilla² i sobreposar la resta del conjunt a estudiar sobre aquesta. A partir d'aquí, es va afegir al programa la capacitat de realitzar càlculs de semblança, ja que la sobreposició ja estava donada, sobre funcions densitat ajustades, les quals es llegeixen de fitxers prèviament preparats. Cas de no voler realitzar aquests càlculs amb funcions ajustades, o ni tan sols càlculs de semblança, el programa treu com a sortida les pròpies sobreposicions en forma de fitxers d'estructures múltiples, les quals es poden emprar per a diverses aplicacions, relacionades o no amb la semblança. Tanmateix, l'ús màxim del TGSA s'ha dedicat a calcular semblances per al seu ús posterior en QSAR, les quals majoritàriament no corresponen al valor que s'obtindria d'emprar el mètode previ, sobretot si hi ha àtoms pesats en joc.

2.3 *Relacions Quantitatives estructura-Activitat (QSAR)*

Ja l'últim tema que es tractarà al llarg del present treball és el desenvolupament, tant des del punt de vista teòric com computacional, i aplicació de la semblança molecular quàntica com a descriptors moleculars en estudis QSAR.

Des d'un barroc punt de vista de *mida*, entesa pel nombre de descriptors o focus on aquests van definits, s'ha contribuït en tres aspectes diferents:

- *Ús de matrius de semblança*. Aquí intervé l'anomenada matriu de semblança, calculada a partir de les semblances per parelles d'entre un conjunt de

¹ Es pot pensar que es podria haver desat cada càlcul atòmic en disc i recuperar-lo quan fos necessari, però aquest procediment va esdevenir més lent per raons de maquinari: era més ràpid recalculer cada cop l'entorn d'un àtom que no llegir els mateixos valors de disc.

molècules. Aquesta matriu és emprada posteriorment, degudament tractada, com a font de descriptors moleculars per a estudis QSAR. Dins d'aquest àmbit s'han fet diversos estudis de correlació d'interès farmacològic, toxicològic, així com de diverses propietats físiques.

- *Aplicació de l'energia d'interacció electró-electró, assimilat com a una forma d'autosemblança.* Aquesta modesta contribució consisteix breument en prendre el valor d'aquesta magnitud, i per analogia amb la notació de l'autosemblança molecular quàntica, assimilar-la com a cas particular de d'aquesta mesura. Aquesta energia d'interacció s'obté fàcilment a partir de programari mecanoquàntic, i esdevé ideal per a fer un primer estudi preliminar de correlació, on s'utilitza aquesta magnitud com a únic descriptor. Lluny de voler presentar-se com un gran descriptor, es poden obtenir bones correlacions per a diferents propietats en sistemes homogenis, tan sols aprofitant el valor d'aquesta energia en acabat de modelar el conjunt molecular d'interès.
- *Càlcul d'autosemblances, on la densitat ha estat modificada per a augmentar el paper d'un substituent.* Les contribucions presentades en aquest aspecte venen inspirades per dues bandes: treballs previs en càlculs de semblança sobre fragment moleculars i l'estada de recerca amb el Prof. Robert Ponec (República Txeca), i els seus treballs amb densitats de Forats de Fermi. Aquests treballs amb fragments, tot i donar molt bons resultats, manquen de cert rigor conceptual en aïllar un fragment, suposadament responsable de l'activitat molecular, de la totalitat de l'estructura molecular, tot i que les densitats associades a aquest fragment ja difereixen degut a pertànyer a esquelets amb diferents substitucions. Un procediment per a omplir aquest buit que deixa la simple separació del fragment, considerant així la totalitat de la molècula (calcular-ne l'autosemblança), però evitant al mateix temps valors d'autosemblança no desitjats provocats per àtoms pesats, és l'ús de densitats de Forats de fermi, els quals es troben definits al voltant del fragment d'interès. Aquest procediment modifica la densitat de manera que es troba majoritàriament concentrada a la regió d'interès, però alhora permet obtenir una funció densitat, la qual es comporta matemàticament igual que la densitat electrònica regular, podent-se així incorporar dins del marc de la semblança molecular. Les autosemblances calculades amb aquesta metodologia han portat a bones correlacions amb àcids aromàtics substituïts, podent així donar una explicació al seu comportament.

² Realment, la sobreposició emprant una plantilla consisteix en fer sobreposicions per parelles entre la plantilla i cadascuna de les molècules restants. En cap cas es fan sobreposicions simultànies.

Des d'un altre punt de vista, també s'han fet contribucions conceptuals. En el marc de les mesures de semblança i la seva aplicació a QSAR, s'ha desenvolupat la ja comentada energia d'interacció electró-electró com a senzill descriptor molecular per a sèries homogènies. També s'ha implementat una nova mesura de semblança, la d'energia cinètica, la qual consisteix en prendre la recentment desenvolupada funció densitat d'energia cinètica, la qual al comportar-se matemàticament igual a les densitats electròniques regulars, s'ha incorporat en el marc de la semblança. A partir d'aquesta mesura s'han obtingut models QSAR satisfactoris per diferents conjunts moleculars. Només mencionar el desenvolupament de les ja comentades mesures d'autosemblança sobre densitats de forats de Fermi.

Dins de l'aspecte del tractament de les matrius de semblança s'ha implementat l'anomenada transformació estocàstica com a alternativa a l'ús de l'índex Carbó. Aquesta transformació de la matriu de semblança permet obtenir una nova matriu no simètrica, la qual pot ser posteriorment tractada per a construir models QSAR. Així mateix, com ja s'ha comentat en l'apartat 2.2, la implementació del TGSA ha portat al càlcul ràpid i a l'obtenció de mesures de semblança molecular, en principi sobre densitats ajustades, basades no en el valor màxim de la mesura per aquelles parelles de molècules estudiades, sinó a una mesura restringida a l'esquelet molecular comú. Finalment, comentar que s'ha treballat amb diferents protocols estadístics, com són la regressió multilineal (de l'anglès *Multilinear Regression*, o MLR) l'anàlisi de components principals (de l'anglès *Principal Component Analysis*, o PCA), i mínims quadrats parcials (de l'anglès *Partial Least-Squares*, o PLS). Aquests protocols s'han implementat en diferent programari desenvolupat a l'Institut.

2.4 Estructura del present treball

L'estructura d'aquest treball s'ha intentat organitzar des d'un punt de vista cronològic, vist des de l'inici en que cada problema va presentar-se i tractar-se. Dels tres temes comentats en els apartats anteriors, es comença amb els *Isocontorns de Densitat Electrònica Molecular* al tercer capítol. En aquest capítol es tractaran diversos conceptes sobre les funcions densitat ajustades, la justificació del seu ús en semblança quàntica, i finalment la seva aplicació en la representació de MIDCOs, així com les posteriors ampliacions del mètode. Com a apèndix en aquest capítol, s'inclouen les tres contribucions que s'han realitzat en aquest camp en forma d'article, o manuscrit original, ja que en el moment de redacció del present treball, l'article es troba en premsa.

En el capítol 4 es fa una introducció en el camp del QSAR, intentant cobrir tots els seus aspectes. Des del seu desenvolupament històric, fins a les seves etapes, bé siguin generació i tipus de descriptors, escalat d'aquests, protocols estadístics i validació de resultats. En aquest capítol no s'annexona cap contribució ja que pretén ser una introducció als dos següents capítols, on es tracta el desenvolupament del TGSA i el gruix de la semblança quàntica amb les seves corresponents aplicacions.

En el cinquè capítol, s'aprofundeix en el concepte d'alineament molecular, metodologies de sobreposició i aplicacions. Així mateix, i lligant amb l'apartat 2.2, es tracta el desenvolupament del programa informàtic TGSA, tant des del punt de vista algorísmic com la seva implementació pràctica, així com exemples d'aplicació. Com a apèndix, s'adjunta la publicació derivada del treball presentat.

El sisè capítol tracta les Mesures de Semblança Molecular Quàntica en la seva totalitat. Es comença amb un desenvolupament històric, la seva formulació matemàtica i tipus de mesures, seguit de les múltiples aplicacions que s'han derivat de la idea inicial. El text continua amb exemples d'aplicació de la semblança en QSAR, en els quals es lliga amb els capítols anteriors, justificant-se l'ús de densitats ajustades enlloc de les *ab initio* i la necessitat d'un nou algorisme d'alineament molecular. Els exemples d'aplicació mostrats, complementats amb les contribucions publicades annexades al final del capítol, cobreixen diferents àmbits d'interès, com són el farmacològic, toxicitat o predicció de propietats físiques.

Finalment, el setè capítol engloba les consideracions finals, així com rendeix comptes d'altres temes tractats amb menys profunditat al llarg del treball global, o en els que s'han realitzat contribucions menors. També es tracten probables objectius de futur a tractar en qualsevol dels àmbits de recerca que s'han tractat. Per acabar, es dóna un llistat total de les publicacions on s'ha contribuït amb el seu corresponent resum (de l'anglès *abstract*).

3. Isocontorns de Densitat Electrònica Molecular

- **Resum:** En aquest capítol, es revisaran les metodologies emprades per a la visualització de superfícies moleculars des d'una perspectiva històrica fins a l'actualitat, on l'ús de computadors i de gràfics tridimensionals ha esdevingut quotidià. Seguidament, es centrarà més en l'ús de la Funció Densitat Electrònica com a punt de partida de la construcció de les superfícies moleculars, detallant algunes de les metodologies més emprades per a la seva obtenció. Finalment, s'exposa el protocol seguit per a la construcció de superfícies de isodensitat electrònica, així com diferents refinaments aportats.
- **Paraules Clau:** DF (Funció Densitat), MIDCO (IsoCOntorn de Densitat Molecular)

3.1. Introducció

Els models físics moleculars han emprat, i malgrat la nova era dels computadors, es continuen utilitzant per a la descripció i l'estudi d'estructures moleculars. Aquests models físics, en el que es fonamenten els moderns gràfics per ordinador, es construeixen a partir de diferents materials, com són plàstic, goma, metall, fusta, paper, tints o pintures. La classificació bàsica d'aquests models és:

- *Models d'àtoms i enllaços*, emprats sobretot a química orgànica.
- *Matrius regulars polièdriques i patrons repetitius*, utilitzats en camps on la simetria és important, tal com la cristal·lografia o química inorgànica.
- *Representacions de plegament de cadenes*, vitals en l'estudi de polímers i proteïnes.
- *Models orbitalaris*, importants en química quàntica.

Els models d'àtoms i enllaços es poden descomposar al llarg d'un espectre més ampli entre l'èmfasi que es dóna a les representacions atòmiques i a la descripció assignada als enllaços. Els models de 'Space-Filling' (espai farcit) són els més emprats. En aquest model, els àtoms es representen per esferes de radi proporcional al seu respectiu radi de van der Waals. Dins d'aquest tipus de model, el més conegut és el de Corey-Pauling-Koltun, normalment abreujat per CPK, del que se'n mostra un exemple a la Figura 1. Els models originals de Corey i Pauling [1] eren

fets de fusta, els quals permetien un escalat d'una polsada per angstrom, o bé de plàstic, assolint una proporció de mitja polsada per angstrom. Els models de fusta estaven interconnectats amb claus metàl·lics, mentre que els de plàstic ho estaven amb cinta adhesiva. Al 1960, en la secció d'estudis de biofísica i química biofísica del NIH, es va dissenyar una millora d'aquests models, portant a una nova representació, que tot i basada en el model de Corey i Pauling, que tenia millors connectors (elaborats per Koltun, d'aquí el nom CPK), n'alleugeria el pes. Tot i les millores, la construcció d'un model CPK per a una macromolècula continuava essent una tasca llarga i tediosa. Per exemple, Yankeeelov i Coggins [2] van construir un model CPK per a la mioglobina a un ritme de 50 aminoàcids per setmana.

A l'altre extrem es poden trobar els models d'esquelet, on els enllaços es representen per cilindres metàl·lics, i els àtoms són tan sols les juntes o punts d'unió d'aquests cilindres (p.e. Figura 2). Amb aquest sistema, Kendrew i col·laboradors [3] van crear un model de la mioglobina, l'escala del qual era de dos centímetres per angstrom.

Entremig dels models basats únicament en àtom o enllaços, es troben models híbrids com els de Rupley i Bruice [4] o Clarke [5]. Així mateix, també s'inventaren sistemes més variats per a àcids nucleics, com el model de Barret [6].

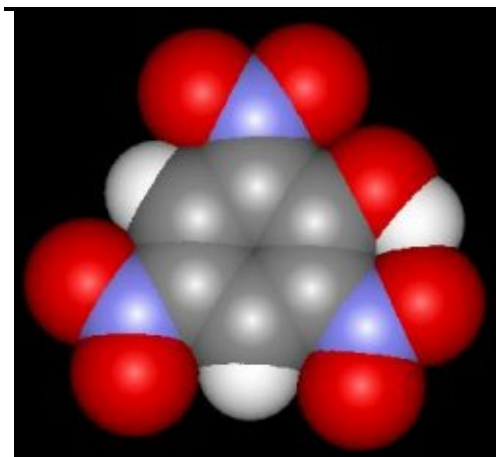


Figura 1.- Representació CPK del 2,4,6-trinitrofenol.

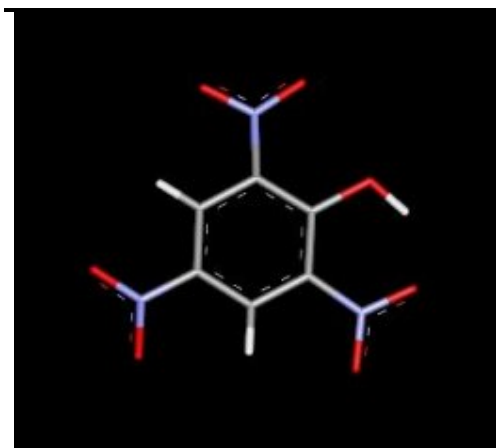


Figura 2.- Representació esquelètica del 2,4,6-trinitrofenol.

Les *Matrius regulars polièdriques i els patrons repetitius* són destacats a la literatura. A títol d'exemple, Wells [7] mostra com la química inorgànica podria ser ensenyada a partir de cinc tipus de models: polièdrics, de xarxa, d'empaquetaments esfèrics i d'estructures octaèdriques i tetraèdriques. El llibre de Walton [8], on també es discuteixen diversos models físics moleculars (així com representacions orbitalàries), tracta també models

polièdrics, incloent una llarga llista de proveïdors comercials. El llibre de Bassow [9] dóna una pauta d'ensenyament de química orgànica i cristal·lografia a partir de la construcció de models. Finalment, i per donar més èmfasi a la importància d'aquests models moleculars, el *Journal of Chemical Education* ha publicat diversos treballs sobre aquestes representacions (veure per exemple les referències [10] i [11]). La figura 3 mostra un exemple d'aquest tipus de representació.

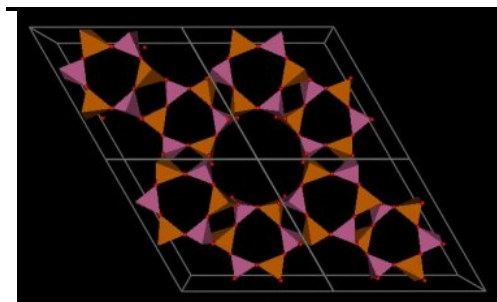


Figura 3.- Esquema de repetició d'una zeolita.

Els *models de plegament de cadenes*, sobretot emprats en el cas concret de proteïnes, estan basats en la simplificació o idealització de les cadenes polipeptídiques com a una sèrie d'enllaços virtuals que connecten els carbonis alfa de la proteïna estudiada (p.e. Figura 4). Els angles d'enllaç de cada carboni alfa i els angles de torsió al llarg dels enllaços virtuals especifiquen el plegament proteic. Per a la construcció d'aquests models, es construïren dos aparells: el Rubin's wire-bender [12] i el Blackwell Molecular Models [13,14].

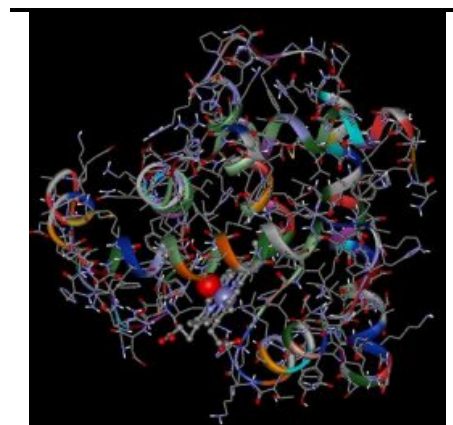


Figura 4.- Representació de les cadenes de l'hemoglobina.

Els *models orbitalaris* començaren a partir de la representació de densitat electrònica obtinguda a partir de dades cristal·logràfiques. L'any 1960 s'elucidà per primer cop una estructura tridimensional d'una proteïna, la mioglobina [15]. La densitat electrònica associada a la qual, fou representada per models metàl·lics, a una escala de cinc centímetres per angstrom. El model estava construït a partir d'un munt de barets metàl·liques, a les que s'adjuntaren clips colorejats que representaven diferents valors de la densitat. Degut al gran nombre de barets, aproximadament unes

2500, aquestes dificultaven la visió del model molecular, així com impediéren fer-hi ajustos fàcilment.

L'any 1968, s'introduí el comparador òptic [16], el qual esdevingué la manera estàndard de construir models moleculars a contorns de densitat electrònica. Els contorns es dibuixaven sobre fulles primes d'acetat muntades sobre fulles de Perspex de 36''×36'', i ambdós el model i la densitat eren il·luminats per bulbs de llum, els quals estaven fets a la mateixa escala, normalment dos centímetres per angstrom. El model s'aguantava per bares metàl·liques i cordes de piano. La col·locació d'un mirall semi-platejat¹ feia possible sobreposar imatges de contorns de densitat electrònica. Tot i la gran acceptació que van rebre aquestes representacions, la possibilitat d'emprar gràfics per ordinador desplaçaren aquest model ràpidament a partir del 1985.

Les grans avantatges que comporta l'ús de sistemes informàtics sobre la construcció de models físics són evidents: rapidesa, cura, control, mida de la molècula que pot ser tractada i costos, ja que un sol ordinador pot generar, emmagatzemar, visualitzar i realitzar diferents anàlisis de diferents models.

Els primers treballs per a modelar computacionalment molècules i densitat electròniques es van dur a terme dins del projecte MAC (Mathematics And Computations) al Massachusetts Institute of Technology (MIT) [17-19]. Tot i que amb el temps, part de l'equip va disgregar-se a diferents institucions, el treball de modelització va continuar [20-24]. Dins d'aquesta primera etapa, i degut a les clares limitacions del maquinari de l'època, tan sols es podien modelar i representar gràficament molècules petites, mentre que el treball amb proteïnes quedaria aparcat per a un futur no tant llunyà.

Amb l'adveniment de la projecció dels gràfics rasteritzats i renderitzats, els models usualment representats per simples línies foren complementats per esferes sòlides del tipus CPK. Força mètodes s'han desenvolupat per aquests models [25-30], així com una quantitat elevada de programari destinat a aquestes representacions, com per exemple:

- RasMol [31], capaç de dibuixar esferes sòlides amb ombres.
- Flex [32].
- Sculpt [33], que realitza fins i tot càlculs senzills de canvis energètics degut a conformacions proteiques.

L'aparició i l'establiment de noves tecnologies informàtiques també ha afavorit en gran mesura el camp de les representacions moleculars. Aquestes nous recursos, com són Internet i

¹ Val a dir que aquesta tècnica de miralls s'havia emprat anteriorment en trucs d'il·lusionisme.

alguns dels llenguatges de programació que s'han derivat de la xarxa, tal com el VRML² [34] (Virtual Reality Modeling Language, de l'anglès llenguatge de modelat en realitat virtual) i el llenguatge Java [35], han permès la difusió, visualització i interacció amb el model creat a qualsevol usuari que disposi d'una connexió a Internet i un navegador Web. La gran difusió d'aquestes eines es deu, en part, a la excel·lent capacitat d'execució en diferents plataformes informàtiques (PC-Win/Linux, MacOS, Dec, Sgi,...), així com a la gratuïtat del programari i de diversos entorns per al seu ús.

Finalment, per acabar amb la part de gràfics per ordinador, destacar les llibreries *OpenGL* [36] (propietat de Silicon Graphics) i les seves clòniques *Mesa* [37] (llicència GPL). Aquestes llibreries, portades a diferents plataformes i llenguatges de programació, permeten la creació de gràfics tridimensionals de manera fàcil i senzilla amb una qualitat molt elevada, tot i els elevats requeriments de maquinari.

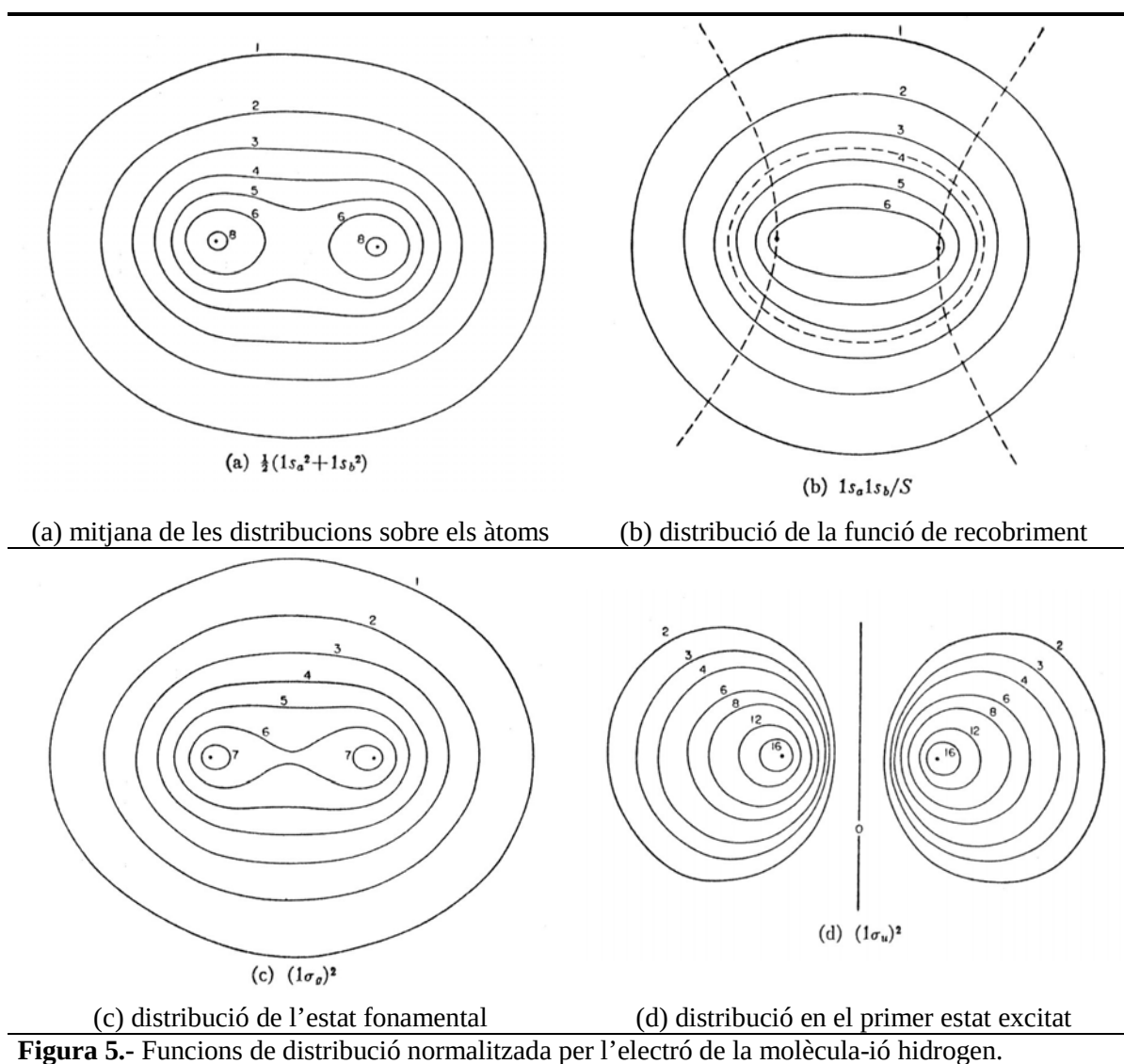
Tornant al tema de les representacions moleculars, i com ja s'havia comentat de bon principi, la visualització, així com la manipulació de la forma molecular, constitueix una poderosa eina per tant l'elucidació dels sistemes molecular com per l'enteniment del seu comportament químic o de la resposta biològica que comporta al interaccionar amb matèria viva. Aquesta capacitat ha portat, tal i com s'ha citat anteriorment a la introducció de diversos models per a la descripció.

Dins del camp de la química quàntica i pel que fa a les primeres descripcions de formes locals, les representacions orbitalàries de Mulliken [38-42] donaven un primer 'borrador' de la forma molecular, així com de les interaccions presents dins de l'estructura electrònica molecular. La figura de la següent pàgina reproduïx algunes de les representacions gràfiques per a la molècula-ió hidrogen.

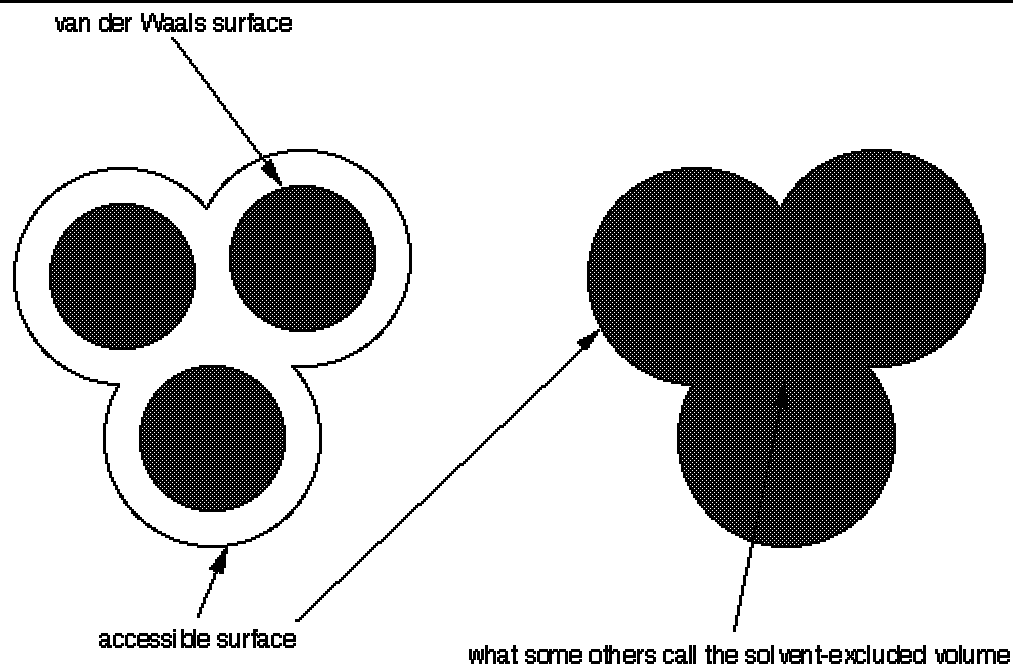
La química quàntica proporciona, en general, les representacions més acurades de l'anatomia molecular [43]. Els mètodes anteriorment citats, com són els esqueletals o de esferes i cilindres, proporcionen una visió discreta de la forma molecular, emprant els àtoms i els enllaços com a blocs de construcció primaris. Com a alternativa, hi ha les representacions amb 'cintes', utilitzades principalment en bioquímica per a representar proteïnes. Tanmateix, aquests diagrames

² Cal tenir en compte que els fitxers generats en VRML i Java no són directament programes executables en el sentit tradicional, sinó que són llenguatges interpretats, no compilats. Això significa que un programa, que existeix prèviament en l'ordinador de l'usuari, s'encarrega de 'compilar' i executar en temps real aquest fitxer. En el cas del VRML, és el propi navegador web que interpreta el fitxer, tal i com es fa amb el llenguatge HTML. El sistema utilitzat per Java és crear codi objecte semi-compilat, interpretable per una màquina virtual, reconeixible per la versió adient a cada plataforma i/o sistema operatiu. D'aquí la seva fàcil portabilitat.

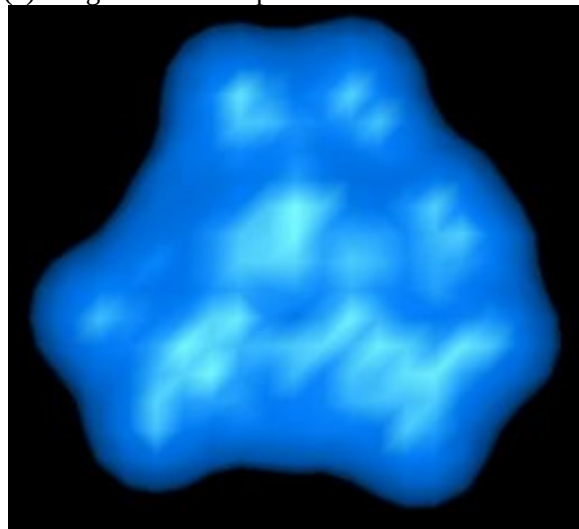
no són prou apropiats per a una representació de la superfície molecular, ja que no són capaços de donar una correcta impressió de l'espai realment ocupat en l'entorn molecular.



Com a remei, es pot recórrer a les representacions CPK anteriorment citades, o a d'altres procediments per a la creació de superfícies moleculars, com són les superfícies accessibles al solvent [44] (de l'anglès *solvent-accessible*), o bé les superfícies de Connolly [45]. El primer tipus correspon a aquella superfície que recobreix els àtoms, la qual és accessible per a una esfera d'un radi predeterminat (que simula el solvent). Exemples de superfícies accessibles al solvent, es mostren a la figura 6.



(a) Diagrama de la superfície accessible al solvent



(b) Superfície accessible al solvent per a la molècula de 2,4,6-trinitrofenol

Figura 6.- Diagrama i exemple de superfície accessible al solvent.

La superfície de Connolly està composta per dos tipus de contribucions: la part de la superfície de van der Waals de cada àtom que és accessible a una esfera de radi donat (normalment $1.4 \text{ \AA}$, corresponent al volum que inclou la majoria de la molècula d'aigua³), i una part de la superfície interior d'aquesta esfera quan està en contacte simultani amb dos o més àtoms. Un diagrama d'aquesta superfície es mostra a la figura 7.

³ En termes estrictament mecanoquàntics, la molècula d'aigua és infinita. Tanmateix, el volum indicat conté la major part de la seva funció densitat.

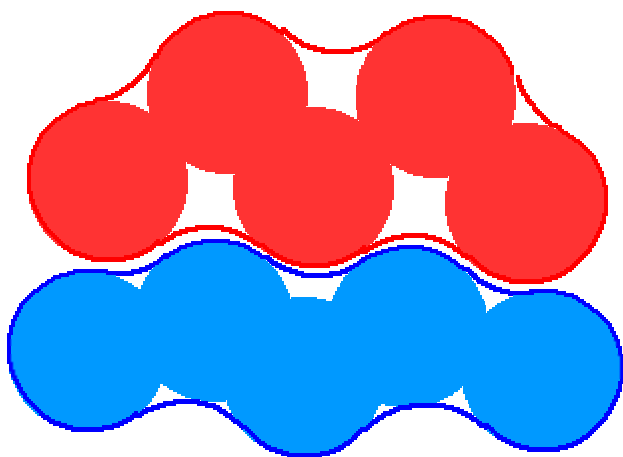


Figura 7.- Diagrama d'una superfície de Connolly.

Tanmateix, als mètodes anteriorment citats els manca realisme alhora de representar la forma molecular, ja que no tenen en compte les interaccions existents dins del sistema, i el seu caràcter difús; d'aquí que l'ús de Funció Densitat Electrònica proporcioni una representació més apropiada i verosímil de la forma molecular.

Per a l'obtenció de Funcions Densitat existeixen diferents metodologies. En el següent apartat, es descriuran aquests procediments, tenint en compte la seva cura i costos associats; els quals constitueixen l'esquelet principal de la química computacional.

3.2. Obtenció de Funcions Densitat: Metodologies

Per a l'obtenció de models moleculars teòrics es disposa de dues metodologies bàsiques diferents, a fi de predir i interpretar-ne les seves propietats. Aquests dos tipus de modelatge els constitueixen la mecànica molecular i la mecànica quàntica.

- Mecànica molecular: es descriu la molècula des d'un punt de vista clàssic mecànic com a un conjunt d'enllaços elàstics. Els mètodes basats en mecànica molecular proporcionen prediccions de geometries moleculars i calors de formació.
- Mecànica quàntica: un model mecanoquàntic modelitza l'estructura electrònica d'una molècula, el que significa resoldre l'equació d'Schrödinger. La mecànica quàntica s'empra per a predir aquestes propietats electròniques, tal com moments dipolars i

dades espectroscòpiques. Els procediments mecanoquàntics es desdoblen, segons la seva complexitat, en *ab initio*, funcional de la densitat i semiempírics.

3.2.1. Mecànica Molecular

La mecànica molecular és un model mecànic clàssic que representa la molècula com a un conjunt d'àtoms lligats entre ells per enllaços elàstics. Esquemàticament, la mecànica molecular considera aquests enllaços com *molles*, les quals poden ser allongades, comprimides, torçades als angles d'enllaç, i distorsionades als angles de torsió; tot i que també es tenen en compte diferents interaccions entre àtoms no enllaçats directament. La suma de totes aquestes forces s'anomena *camp de forces* (de l'angles *force field*) de la molècula. Aquests camps de forces es construeixen i parametritzen per comparació amb un grup de molècules d'una mateixa família, per exemple un conjunt d'alcanes. El camp de forces resultant pot ser llavors utilitzat per a altres molècules similars a aquelles prèviament parametritzades. Per a dur a terme un càlcul de mecànica molecular cal fixar uns paràmetres estructurals i un camp de forces adequat. Llavors, la geometria s'optimitza en base a la minimització de la tensió a partir de canvis incrementals a l'estructura. Aquesta minimització és ordres de magnitud més ràpida que no el càlcul mecanoquàntic més senzill en qualsevol molècula equivalent.

La mecànica molecular és una eina valuosa per a la predicció de geometries i calors de formació de molècules, sempre que existeixi un camp de forces adequat a aquestes. D'aquí un dels punts febles dels mètodes basats en mecànica molecular: al basar-se en camps de forces parametrizats per a molècules corrents, l'estudi de nous compostos pot ser impossible (o sotmès a un error imprevisible) per la inexistència de camps de forces adequats. Una segona mancança és la consideració de les molècules com a conjunt de cordes elàstiques: al no tenir en compte les estructures electròniques, les propietats basades en aquestes no podran ser tractades, com per exemple dissociacions o formacions d'enllaços, havent de recórrer a la mecànica quàntica.

Tanmateix, tot i no ser una metodologia destinada a l'obtenció de funcions densitat, si en pot ser un bon punt de partida per a l'obtenció d'una geometria inicial, per a després ser refinada per un càlcul més precís, estalviant així recursos computacionals.

3.2.2. Mecànica Quàntica

La realització d'un model mecanoquàntic de l'estructura electrònica d'una molècula comporta la resolució de l'equació d'Schrödinger:

$$\mathbf{H}\Psi = E\Psi \quad (1)$$

L'operador Hamiltonià, $\mathbf{H}$, depèn dels operadors energia cinètica i energia potencial aplicats sobre els nuclis i electrons dels àtoms presents a la molècula. La funció d'ona, Ψ , es proporcionarà informació sobre la probabilitat de trobar els electrons a diferents localitzacions de la molècula a partir del seu mòdul al quadrat, donant la funció densitat $\rho = |\Psi|^2$. L'energia total del sistema, E , s'obté resolent l'equació.

La resolució de l'equació d'Schrödinger consisteix un problema molt complex, irresoluble per a sistemes polieletrònics sense considerar aproximacions. Les aproximacions utilitzades són:

- *L'aproximació de Born-Oppenheimer*: les posicions dels nuclis són fixades, mantenint així les distàncies interatòmiques constants. Tal aproximació és força subtil, ja que els nuclis són considerablement més pesants que no els electrons, podent-se considerar com a pràcticament immòbils respecte els electrons. Triant així una disposició atòmica, es pot solucionar de manera aproximada l'equació per aquella geometria donada, i cas de voler dur a terme una optimització de geometria, es poden anar provant diferents disposicions fins a trobar la que proporciona la mínima energia, existint també mètodes analítics derivats específicament per a aquesta tasca.
- *L'aproximació d'electrons independents*: la presència de més d'un electró impossibilita la resolució exacte de l'equació d'Schrödinger degut al terme interelectrònic al Hamiltonià. En el senzill cas de la molècula d' H_2 , la formulació del hamiltonià és:

$$H = -\frac{\hbar^2}{2m}\nabla_1^2 - \frac{\hbar^2}{2m}\nabla_2^2 - \frac{e^2}{4\pi\epsilon_0}\left(\frac{1}{r_{a1}} + \frac{1}{r_{a2}} + \frac{1}{r_{b1}} + \frac{1}{r_{b2}} - \frac{1}{R} - \frac{1}{r_{12}}\right)$$

Els primers dos termes representen a l'energia cinètica i els sis restants expressen l'energia potencial d'un sistema de 4 partícules. L'últim terme, el que té en compte la repulsió interelectrònica, fa impossible la resolució exacta. Així, s'assumeix que la Ψ pot ser descrita com a una combinació lineal de productes de funcions monoelectròniques, anomenades orbitals moleculars, l'equivalent molecular als orbitals atòmics. Cada orbital molecular s'expressa com a una combinació d'orbitals atòmics dels àtoms que constitueixen la molècula. En el cas anterior de la molècula d' H_2 ,

l'orbital molecular més senzill pot escriure's com a $c_1 1S_1 + c_2 1S_2$, on les $1S_i$ són orbitals $1S$ de l'àtom d'H i els c_i són paràmetres.

A partir d'aquests principis, apareixen diferents mètodes de resolució d'acord a les diferents aproximacions matemàtiques que es poden dur a terme. Aquests mètodes són:

- *ab initio*: tal i com el nom indica, aquests mètodes no parteixen de cap informació prèvia excepte aquella proporcionada per les lleis de la mecànica quàntica i una sèrie de constants físiques, com són la velocitat de la llum, la constant de Plank i les masses dels nuclis i de l'electró. Els mètodes *ab initio* són adequats a qualsevol tipus de sistema, tanmateix el seu límit principal està el la realització pràctica del càlcul, ja que degut a la gran quantitat d'integrals a resoldre, la quantitat necessària de memòria per al seu emmagatzement i els processos iteratius que cal convergir, superar una estructura de 100-150 àtoms⁴ pot resultar en una tasca computacionalment molt costosa segons el tipus concret de càlcul i conjunt de funcions de base a emprar. Alguns dels mètodes *ab initio*, de menys a més computacionalment costós⁵, són:
 - Hartree-Fock
 - Teoria de perturbació de Møller-Plesset
 - Coupled Cluster
 - Interacció de configuracions
- *Funcional de la densitat*: basats en el teorema de Hohenberg-Kohn, on es demostra l'existència d'un únic funcional capaç de descriure l'estat fonamental i la densitat de primer ordre de manera exacta (tot i que la forma explícita de tal funcional no es dona). La gran avantatge, i causa de la creixent popularitat, del mètode és la inclusió de correlació electrònica amb requeriments computacionals equivalents als de Hartree-Fock, el mètode *ab initio* menys costós.
- *Semiempírics*: aquests mètodes estan dissenyats per a aproximar, o fins i tot ignorar, el càlcul de diverses integrals que es realitzen en els càlculs *ab initio*, en base a diversos paràmetres preestablerts, que provenen de dades experimentals o de càlculs *ab initio* previs. De manera semblant a la mecànica molecular, els resultats són bons quan

⁴ Amb l'excepció dels sistemes periòdics, com per exemple algunes cadenes polimèriques o xarxes cristal·lines.

⁵ Val a dir que el cost relatiu de les diferents metodologies *ab initio* depèn de les condicions particulars del càlcul a realitzar. D'aquesta manera, per exemple, un càlcul d'interacció de configuracions truncat (no complet) amb una base petita pot esdevenir menys costós que un Hartree-Fock amb una base gran sobre el mateix sistema molecular.

existeixen paràmetres disponibles per les molècules a tractar, però poden fallar imprevisiblement quan aquests manquen.

Més detalls sobre les diverses metodologies exposades es poden trobar a les referències 46 a 54. Així mateix també existeix una gran diversitat de programari disponible per a realitzar càlculs a tots els nivells exposats, des de programes dedicats a un sol mètode fins a programes que unifiquen força dels procediments llistats, incloent interfícies i visualitzadors gràfics de resultats. Algunes pàgines WWW cataloguen, de manera més o menys actualitzada, el programari disponible, per exemple, la referència 55 en mostra un petit llistat.

3.2.3. Funcions Densitat Ajustades: l'Aproximació a Capes Atòmiques (ASA) i ASA Promolecular

L'Aproximació a Capes Atòmiques (ASA, de l'anglès *Atomic Shell Approximation*) [56-59] expressa la densitat electrònica molecular de primer ordre com a una combinació lineal de funcions gaussianes 1S. Aquesta metodologia es fonamenta sobre els càlculs ab initio, ja que es pren la densitat electrònica, prèviament calculada a qualsevol nivell computacional i amb qualsevol conjunt de funcions de base, i es realitza un procés d'ajust de la densitat ab initio a les citades funcions 1S, de tal manera que la densitat, dins de l'ASA, resta expressada com:

$$\rho_A^{ASA}(\mathbf{r}) = \sum_{i \in A} w_i |S_i(\mathbf{r})|^2, \quad (2)$$

on les $S_i(\mathbf{r})$ són funcions esfèriques normalitzades, tal que $\int |S_i(\mathbf{r})|^2 d\mathbf{r} = 1$, i els coeficients w_i són restringits a complir les condicions de convexitat, expressades a les següents equacions (3) i (4):

$$w_i \in \mathbf{R}^+, \forall i \in \mathbf{A}, \quad (3)$$

per a assegurar que la funció densitat sigui definida positiva en tot el seu domini, i

$$\sum_{i \in A} w_i = N_e, \quad (4)$$

per a garantir la normalització de la funció ASA al nombre d'electrons (N_e), preservant així el sentit estadístic de funció densitat de càrrega de l'equació (2).

Tot i que l'equació (2) deriva una expressió molt simplificada i, que degut a la única presència de funcions 1S, els càlculs resultats, fonamentalment de semblança quàntica que es comentaran al següent apartat, seran conceptualment senzills i relativament molt més ràpids que els

respectius ab initio, el fet de la funció densitat ASA provingui d'un càlcul previ ab initio comporta un conjunt de seriosos desavantatges:

- El nombre d'àtoms que pot abarcar pràcticament està acotat per el propi límit del càlcul ab initio. Al necessitar aquest càlcul previ, tot i que els càlculs derivats serien molt més ràpids, no es poden considerar, per exemple, estudis de macromolècules.
- El procés d'ajust de la funció densitat ab initio a la ASA [56] és computacionalment molt costós, encarint notablement el temps de càlcul a mesura que la mida molecular, o la grandària del conjunt de funcions de base original, augmenta. Veure la Taula 1 per a un conjunt d'exemples.
- Els càlculs no són reutilitzables, sinó que cal fer un ajust complet per a cada nova molècula que es vulgui estudiar o fer intervenir en un càlcul.

Taula 1.- Temps requerit (en segons) per ajustar una DF ASA a partir d'una ab initio, obtinguda d'un càlcul HF/3-21G [57].

Aigua	13
Metanol	300
Etanol	1675
Alcohol al·lílic	1478
Benzè	4815
Glicina	1750

Nota: Els càlculs es feren en un Intel Pentium 200MHz amb 64 Mb de RAM sota W95.

Per a poder donar una solució a aquests inconvenients, però basant-se en les mateixes idees, es va formular la ASA Promolecular [58-60] (PASA, de l'anglès *Promolecular Atomic Shell Approximation*). En aquesta aproximació, la funció densitat molecular es construeix a partir de contribucions atòmiques discretes, les quals, i a l'igual que en l'ASA, es construeixen a partir de combinacions lineals de funcions 1S. Matemàticament, les funcions densitat ASA són expressades:

$$\rho_A^{PASA}(\mathbf{r}) = \sum_{a \in A} P_a \rho_a^{ASA}(\mathbf{r} - \mathbf{R}_a) \quad (5)$$

on P_a és el nombre atòmic de l'àtom a -èssim. D'aquesta manera, la densitat molecular és una senzilla addició de densitats atòmiques, ρ_a^{ASA} , centrades a les posicions nuclears $\mathbf{R}_a$. Aquestes densitats atòmiques són expressades com a (2), excepte que la condició de concavitat (3) passa a

$$\sum_{i \in A} w_i = 1 \quad (6)$$

per a mantenir $\int \rho_A^{PASA}(\mathbf{r}) = N_e$, ja que cada ρ_a^{ASA} ja està ponderada per P_a , mantenint-se així altra vegada vàlida la definició de funció de distribució de càrrega.

Així, i a efectes pràctics, cal dur a terme una sèrie de càlculs ab initio per a sistemes atòmics, els quals són ajustats per aquest procediment a funcions 1S amb els corresponents coeficients w_i . Val a dir que els càlculs sobre sistemes atòmics, per complex que sigui el mètode i gran que sigui el conjunt de base, seran sempre assequibles en maquinari, tot i que modest, actual. Els resultats de cada ajust són guardats, i la construcció d'una funció densitat PASA correspondrà a obtenir unes coordenades moleculars i a l'addició, segons correspongui, de funcions i coeficients, podent-se així reciclar els càlculs i ampliant abastament la mida màxima tractable. A títol d'exemple, la Taula 2 presenta coeficients i exponents de funcions 1S ajustats de càlculs HF/6-311G [60], per alguns dels àtoms més utilitzats.

Taula 2.- Coeficients i exponents de funcions 1S, ajustats de càlculs HF/6-311G [60], per a alguns àtoms d'ús comú.

Àtom	# funcions	Coeficients	Exponents
H	3	0.027596800839	0.390860160770
		0.108544030597	1.459212872680
		0.117081861100	8.104816659580
C	4	0.328822039966	0.577860648504
		15.583115815951	16.55857680144
		42.939229093834	60.89248951420
		44.267106467827	331.7374865580
N	4	0.702871539536	0.830955781502
		25.121782328105	22.77798472040
		70.406085400475	84.24628587740
		71.127462144409	466.7310065480
O	4	1.286587067288	1.107483845410
		34.139655718380	28.49567506560
		104.966167997914	103.2029729188
		110.521366850651	564.7182462400
Cl	5	0.597525259315	0.551763872506
		42.336372823651	10.24479791954
		472.343731771283	165.5913610356
		1106.160702352482	592.5897432320
		1041.087589805639	3241.403113780

Nota: els coeficients ja es troben normalitzats.

Ajustos sobre més àtoms, així com per a diferents bases, es poden descarregar del web de l'Institut de Química Computacional [61].

En aquest punt cal remarcar la semblança conceptual entre la PASA i els mètodes semiempírics. Mentre que aquests últims aprofiten paràmetres prèviament calculats per a simplificar els propis, la PASA pren aquests ajustos previs per a construir les funcions densitat. D'aquesta manera, es pot considerar la PASA com a un mètode semiempíric que proporciona funcions densitats ajustades.

3.2.4. Funcions Densitat Ajustades i Semblança Molecular Quàntica

Tot i que el gruix de la teoria de la Semblança Molecular Quàntica no es tractarà amb profunditat fins a un capítol posterior, és un bon moment per a veure la diferència de les funcions densitat ajustades front les ab initio en aquests tipus de càlculs, sobretot en els aspectes de complexitat i velocitat de càlcul.

En termes molt genèrics, la semblança quàntica mesura la similitud entre dos objectes quàntics en base a les seves distribucions electròniques. Matemàticament, la semblança quàntica s'expressa com [62]:

$$Z_{AB} = \iint \rho_A(\mathbf{r}_1) \Omega(\mathbf{r}_1, \mathbf{r}_2) \rho_B(\mathbf{r}_2) d\mathbf{r}_1 d\mathbf{r}_2 \quad (7)$$

on Z_{AB} és la corresponent mesura de semblança entre els objectes A i B , ponderada per un operador definit positiu Ω . Cal tenir en compte que donades dues molècules, la mesura de semblança no és única, sinó que depèn en gran mesura de les posicions relatives d'ambdues molècules en l'espai, pel que si es desitja trobar la màxima semblança, caldrà fer una busca exhaustiva, fet que comportarà diverses repeticions del càlcul en diverses orientacions moleculars.

En el cas que s'agafés una funció densitat provinent d'un càlcul Hartree-Fock, l'expressió genèrica d'aquesta seria amb l'estructura de:

$$\rho_A(\mathbf{r}) = \sum_{\mu\nu} D_{\mu\nu} \chi_\mu^*(\mathbf{r}) \chi_\nu(\mathbf{r}), \quad (8)$$

on $D_{\mu\nu}$ representa la matriu de les càrregues i els ordres d'ordres d'enllaç i $\{\chi_\mu\}$ és el conjunt de funcions de base d'orbitals atòmics. Si es vol realitzar una mesura de semblança molecular quàntica de dos sistemes moleculars calculats a aquest nivell, l'expressió que cal avaluar es pot formular com:

$$Z_{AB} = \sum_{\mu \in A} \sum_{\nu \in A} \sum_{\lambda \in B} \sum_{\sigma \in B} D_{\mu\nu} D_{\lambda\sigma} \iint \chi_\mu^*(\mathbf{r}_1) \chi_\nu(\mathbf{r}_1) \Omega(\mathbf{r}_1, \mathbf{r}_2) \chi_\lambda^*(\mathbf{r}_2) \chi_\sigma(\mathbf{r}_2) d\mathbf{r}_1 d\mathbf{r}_2, \quad (9)$$

on com es pot veure, cal calcular integrals a quatre centres, les quals, si es té en compte la presència de funcions de moment angular elevat, són força costoses en termes de temps computacional.

Aquesta desavantatge es tradueix en una limitació encara més acusada en la mida dels sistemes tractables dins d'un punt de vista pràctic⁶.

D'altra banda, si es formula la funció densitat d'un sistema emprant les equacions (2) o (5), recordant però que l'equació (2) requereix un càlcul ab initio previ, l'expressió que pren la mesura de semblança quàntica es redueix a:

$$Z_{AB} = \sum_{i \in A} \sum_{j \in B} w_i w_j \iint |S_i(\mathbf{r}_1)|^2 \Omega(\mathbf{r}_1, \mathbf{r}_2) |S_j(\mathbf{r}_2)|^2 d\mathbf{r}_1 d\mathbf{r}_2, \quad (10)$$

on es simplifica notablement la complexitat de l'expressió com el temps de càlcul requerit. Si es té en compte que el les mesures de semblança avaluades emparant ASA es desvien un 2% de les ab initio en el pitjor dels casos, el seu ús esdevé justificable.

Tot seguit, i per a corroborar les afirmacions anteriors, es presenta una taula comparativa de valors d'autosemblança molecular quàntica per a les molècules presentades a la Taula 1. L'autosemblança molecular quàntica, expressada a partir de l'equació (7) considerant $A=B$ i l'operador $\Omega(\mathbf{r}_1, \mathbf{r}_2) = \delta(\mathbf{r}_1 - \mathbf{r}_2)$, també anomenat operador funció delta de Dirac. La mesura resultant, anomenada *Recobrimet* (de l'anglès *Overlap*) degut a que s'avalua l'espai efectivament compartit entre ambdues funcions, s'expressa com a:

$$Z_{AB} = \int \rho_A(\mathbf{r}) \rho_B(\mathbf{r}) d\mathbf{r}, \quad (11)$$

i dóna una idea de la concentració dels electrons dins la molècula, augmentant en base a la incorporació d'elements més pesants i a la disminució de les distàncies interatòmiques.

La comparativa es fa sobre tres sistemes de funcions densitat ASA i tres més provinents de càlculs Hartree-Fock amb diferents conjunts de funcions de base. Les funcions densitat ASA són les següents:

- PASA simple (Pm. Simple), la qual prové d'ajustos atòmics emprant poques funcions 1S en el procés. En aquest cas, tan sols un funció sobre l'hidrogen i 3 funcions per a la resta d'elements del segon període.
- PASA límit (Pm. Límit), provinent també d'ajustos atòmics, incrementant en aquest cas el nombre de funcions gaussianes 1S fins al punt en que addició de més funcions no en millorava els resultats, encarint per contra el temps de càlcul. En aquest ajust,

⁶ El terme sistema (o molècula) petit o gran va lligat a la capacitat i als requeriments de càlcul del moment. Sistemes inimaginables fa pocs anys són ara tractables a nivells de teoria complexos, tanmateix, i recorrent a una cita apareguda en el Computational Chemistry List (<http://www.ccl.net>), si aparegués l'Octium XII a 2000 OHZ amb gigues de RAM i teres de disc, els químics teòrics estariem contents 15 dies.

l'hidrogen s'ajusta amb 7 funcions, 8 funcions pel carboni, i 9 per la resta d'àtoms del segon període.

- ASA molecular (Mol. 3-21G), provinent d'un ajust directe de la densitat calculada ab initio. En aquest cas s'ha triat el mètode HF amb la base 3-21G.

En el cas de les PASA, les densitats atòmiques han estat ajustades també a partir de càlculs ab initio resultants del mètode Hartree-Fock amb la base 3-21G.

Dins de les mesures ab initio, s'ha utilitzat sempre el mètode de Hartree-Fock amb tres bases diferents:

- STO-3G, per a veure els resultats amb una base petita.
- 3-21G, que permet comparar directament amb els resultats ASA.
- 6-31⁺⁺G**, per a poder veure els resultats amb una base considerablement gran.

En tots els casos, ASA i ab initio, les geometries emprades han estat les mateixes, corresponents a una optimització de geometria al nivell Hartree-Fock amb la base 6-31⁺⁺G**. Els resultats de les mesures sobre els diferents sistemes moleculars són:

Taula 3.- Comparativa entre mesures d'autosemblança de tipus recobriment per a sis sistemes moleculars amb diferents metodologies.

Molècula	ASA			ab initio (HF)		
	Pm. Simple	Pm. Límit	Mol. 3-21G	STO-3G	3-21G	6-31 ⁺⁺ G**
Aigua	80.86	80.52	80.25	79.41	80.24	81.08
Metanol	112.56	112.25	107.89	107.02	108.03	107.74
Etanol	144.32	144.00	141.05	139.66	141.18	142.24
Alcohol al·lílic	175.94	175.60	172.47	170.10	171.61	176.28
Benzè	190.09	190.02	179.59	179.43	180.12	189.81
Glicina	277.16	276.53	273.33	266.53	272.53	274.30

Tal i com es pot veure a la taula anterior, no apareixen diferències significatives entre les mesures d'autosemblança calculades dins de l'aproximació ASA, i aquelles a nivell ab initio en les bases 3-21 i 6-31⁺⁺G**. Un fet que pot resultar curiós, però que realment no és d'estranyar, és la similitud entre les mesures realitzades dins de l'ASA molecular i les ab initio amb la base 3-21G. Aquesta coincidència prové de que l'ASA molecular ha estat ajustada en base a aquestes densitats ab initio, per tant aquest paral·lelisme de valors era esperat. Al mateix temps, destacar les diferències importants en les mesures ab initio amb la base STO-3G comparades amb la resta, les quals resalten la pobre descripció que aquesta base ofereix.

Vista la validesa de la metodologia ASA en l'entorn de la semblança quàntica, val la pena comentar els costos computacionals associats a les mesures dins de cada metodologia. Per a dur a

terme aquest estudi, es va enregistrar el temps de càlcul necessari per a completar el càlcul. Els resultats són:

Taula 4.- Comparativa entre els temps de càlcul⁷ (en segons) per a obtenir mesures d'autosemblança de tipus recobriment per a sis sistemes moleculars amb diferents metodologies.

Molècula	ASA			ab initio (HF)		
	Pm. Simple	Pm. Límit	Mol. 3-21G	STO-3G	3-21G	6-31 ⁺⁺ G**
Aigua	≈0	≈0	≈0	1	1	1
Metanol	≈0	≈0	0.06	1	2	21
Etanol	≈0	0.17	0.33	3	4	125
Alcohol al·lílic	≈0	0.22	≈0	7	8	286
Benzè	≈0	0.22	0.33	16	17	576
Glicina	≈0	0.44	0.38	12	13	484

Vista aquesta taula, es corroboren les afirmacions anteriorment exposades sobre la simplificació que suposa l'ús de densitats ajustades (eq. (10)) en front de les ab initio (eq. (9)). Els temps de càlcul difereixen i augmenten considerablement amb la complexitat del mètode emprat, tanmateix aquest augment de precisió alhora de calcular la funció densitat, no esdevé un factor significatiu en la precisió de la mesura resultant. Tal i com s'ha comentat abans, si es té en compte que aquesta mesura pot repetir-se diverses vegades si es busca el màxim de semblança entre dos sistemes moleculars diferents, depenent de la mida de les molècules i del mètode escollit per a representar la funció densitat, el temps de càlcul de la mesura pot arribar a eternitzar-se.

3.3. Obtenció de MIDCOs

En aquest apartat es descriurà els diferents procediments conceptuals i computacionals utilitzats per a la creació de contorns d'isodensitat molecular (MIDCOs), així com la presentació de codis font en Fortran 90 específics d'algun dels passos del procediment. Primer es presenta l'algorisme general, per a ser detallat, punt per punt, en els següents sub-apartats.

L'algorisme general per a la creació de MIDCOs és:

- Obtenció de les coordenades moleculars.

⁷ Els temps de càlcul reportats han de ser presos des d'un punt de vista semi-quantitatiu. Com es pot veure hi ha temps que són zero, i una incoherència en l'alcohol al·lílic dins de l'ASA molecular. Els temps zero es justifiquen en base a que el rellotge intern del PC (Intel Pentium 200, 64Mb, Windows 95) s'actualitza cada 1/18 segons, permetent així que un càlcul de duració menor no enregistri el vertader temps. La incoherència pot provenir de l'execució paral·lela de programes o controladors del sistema operatiu.

- Construcció de la funció densitat.
- Creació d'una xarxa de punts (hom acostuma a referir-se a aquest tipus de xarxes amb la paraula anglesa *grid*), on es calcularà la funció densitat, prou gran com per abarcar-ne tots els punts significatius, i prou tupida a fi d'obtenir-ne una representació contínua.
- Càlcul del valor de la funció a cada punt de la xarxa.
- Reconstrucció de la superfície i representació d'aquesta per a un valor donat de la funció.

3.3.1. Obtenció de les coordenades moleculars

El conjunt de coordenades moleculars és un dels factors que determina principalment la forma i aspecte dels MIDCOs, ja que en base a les localitzacions atòmiques es construirà la funció densitat. Així, si es vol que els MIDCOs a obtenir representin de forma realista la distribució de la densitat d'una molècula, aquestes han de ser triades amb certa cura, dins de les pròpies restriccions de recursos i temps disponibles.

Dins de les diverses fonts que hi ha per a l'obtenció de coordenades, les més utilitzades són:

- Dades cristal·logràfiques: proporcionen geometries determinades experimentalment amb mètodes de difracció de raig X (o tècniques semblants), donant coordenades usualment força precises. Aquesta font és la més utilitzada quan es tracten sistemes moleculars de mida considerable, o interaccions substrat-enzim o àcid nucleic. Veure per exemple [63-65]
- Càlculs teòrics d'optimització de geometria: proporcionen geometries, la qualitat de les quals depèn en gran mesura del mètode emprat. Com ja s'ha comentat abans, els mètodes basats en mecànica molecular i els semiempírics donen bons resultats en sistemes dels quals existeixen paràmetres calculats en temps raonablement curts. Els mètodes *ab initio* i DFT proporcionen bones geometries a nivells computacionals (o funcionals) elevats amb bases grans, amb l'impediment del temps de càlcul [50].

3.3.2. Creació de la funció densitat

Com que les metodologies existents per a la creació de funcions densitat ja s'ha comentat en l'apartat anterior, només es remarcarà la diferència entre els temps de càlcul entre elles. En el cas de la molècula de 2,4,6-trinitrofenol, presentada a la figura 1, construir-ne la funció promolecular PASA, emprant coeficients i exponents prèviament parametritzats a la base 6-311G [66], és

virtualment instantani, ja que únicament cal realitzar una assignació *àtom* $\leftrightarrow$ *conjunt de funcions* 1S. Al emprar el mètode Hartree-Fock [57] amb la mateixa base, el càlcul va trigar uns 35 minuts a convergir⁸.

3.3.3. Creació del Grid

Per a la creació del grid cal tenir en compte dos factors:

- La mida de la molècula.
- la resolució del gràfic.

Si es vol tenir una representació acurada de la funció densitat, o en general d'altra funció molecular, la mida del grid ha d'encabir el màxim de valors representatius d'aquesta. En el cas concret de la funció densitat, la qual decau seguint una pauta $\exp(-r)$, calculada dins d'una caixa cúbica, no cal desplaçar-se gaire més lluny de l'àtom més allunyat en aquella direcció més 5 unitats atòmiques (au, de l'anglès *atomic units*) [43], distància a la qual la majoria dels valors de densitat cauen per sota de la mil·lèsima de e^-/au^3 . Augmentar gaire més aquests límits comporta un increment lineal del temps de càlcul, així com quadrat i cúbic si s'augmenten dos i tres dimensions respectivament. Si els límits es rebaixen, llavors tan sols es podran estudiar valors relativament alts de la funció densitat, ja que les representacions de MIDCOs de valors baixos poden veure's truncades o deformades als extrems per manca de valors.

La resolució és la quantitat de punts en els que s'avalua la funció per unitat de volum. Aquest factor determina la qualitat de la representació, ja que la majoria de programes de representació gràfica o de reconstrucció de superfícies realitza interpolacions dins del conjunt de punts per representar el MIDCO concret per a un valor pre-establert. Malauradament, qualitat i temps de càlcul estan relacionats cúbicament, en altres paraules, duplicar la resolució incrementa

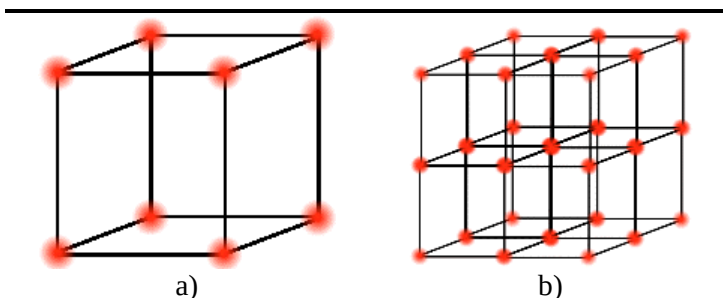


Figura 8.- Exemple de nombre de a) punts a avaluar (color vermell) en un cas senzill, i b) duplicant-ne la resolució.

vuit vegades el temps de càlcul, al haver-se de avaluar 8 vegades més punts, tal i com es mostra gràficament a la figura 8.

⁸ En un DEC Alpha 3000 amb 64 Mb de RAM.

Per a finalitzar amb aquest subapartat, es mostra un exemple de codi en Fortran 90 sobre com definir els límits del grid de càlcul en base unes coordenades moleculars:

```
! na, nombre d'àtoms (entrada)
! xyz, coordenades moleculars: na×3 (entrada, en u.a.)
! doxlim, doylim, dozlim: límits en els eixos XYZ (sortida)
subroutine limits (na,xyz,doxlim,doylim,dozlim)
  implicit none
  integer, intent(in) :: na
  real, intent(in) :: xyz(na,3)
  real, intent(out) : doxlim,doylim,dozlim
  ! mirant quins són els valors extrems de coordenades
  doxlim=minval(xyz(:,1)) ; upxlim=maxval(xyz(:,1))
  doylim=minval(xyz(:,2)) ; upylim=maxval(xyz(:,2))
  dozlim=minval(xyz(:,3)) ; upzlim=maxval(xyz(:,3))
  ! ampliant els límits amb un marge prudencial
  doxlim=-INT(max(ABS(doxlim),upxlim)+5)
  doylim=-INT(max(ABS(doylim),upylim)+5)
  dozlim=-INT(max(ABS(dozlim),upzlim)+5)
end subroutine limits
```

Tall de Codi 1.- Exemple de definició de grid de càlcul. La molècula es suposa centrada al seu centre de càrregues, amb origen a (0,0,0).

3.3.4. Càlcul del valor de la funció

Un cop creat el grid, cal avaluar-ne a cada punt el valor de la funció densitat i emmagatzemar-lo en disc per al seu posterior ús. Seguidament es presenta un segon tall de codi, en el qual com es calcula la funció densitat PASA dins d'un grid pre-establert.

```
! doxlim,doylim,dozlim són els límits del grid
! aug és el factor de resolució triat
! tx,ty,tz indiquen la posició (x,y,z) on es calcularà la DF
npuntsx = int(abs(doxlim*2.0d0/aug)) + 1
npuntsy = int(abs(doylim*2.0d0/aug)) + 1
npuntsz = int(abs(dozlim*2.0d0/aug)) + 1
tx = doxlim ; ty = doylim ; tz = dozlim
do i=1,npuntsx
  do j=1,npuntsy
    do k=1,npuntsz
      call densit(na,nf,xyz,tx,ty,tz,c,ex,lcen,den) ! Càlcul
      write (fileout,fmt="(1x,d10.4)") den ! emmagatzemar
      tz = tz + aug
    end do
    tz = dozlim ; ty = ty + aug
  end do
  ty = doylim ; tx = tx + aug
end do
! =====
! Subrutina del Càlcul de Densitat
! na: nombre d'àtoms
! nf: nombre de funcions 1S
! xyz: coordenades atòmiques, dim na×3
```

```

!   tx,ty,tz: coordenades del punt a avaluar
!   c,ex: coeficients i exponents de les funcions, dim nf
!   lcen: correspondències funció-àtom, dim nf
!   den: valor de la funció en tx,ty,tz
subroutine DENSIT(na,nf,xyz,tx,ty,tz,c,ex,lcen,den)
  implicit none
  real, intent(in) :: xyz(na,3),c(nf),ex(nf),tx,ty,tz
  integer, intent(in) :: na,nf,lcen(nf)
  integer :: i
  real, intent (in,out) :: den
  real, allocatable :: rad(:)
  den = 0.0d0
  allocate (radis(na))
  ! càlcul de les distàncies punt-àtom
  do i=1,na
    rad(i) = (xyz(i,1)-tx)**2+(xyz(i,2)-ty)**2+(xyz(i,3)-tz)**2
  end do
  do i=1,nf
    den = den + c(i)*exp(-ex(i)*rad (lcen(i)))
  end do
  deallocate (rad)
end subroutine densit

```

Tall de Codi 2.- Càlcul de la funció densitat PASA dins d'un grid.

En alguns programes de càlcul aquesta funció pot trobar-se ja inclosa dins del paquet de programari. En el cas concret de Gaussian 98 [57], el càlcul de la funció densitat, o d'altres propietats, és accessible des de la ruta de comandes del fitxer d'entrada, o bé fent servir la utilitat inclosa *Cubgen*. En qualsevol dels dos casos, el resultat és un fitxer *cube* que conté els valors de la funció. Un exemple de com accedir a aquesta funció des de la ruta de comandes, per a la molècula d'hidrogen, es mostra a continuació:

```

%Mem=100Mb
#P rhf/sto-3g geom=cartesian Cube=(Density,Coarse)

Densitat de l'H2

0 1
H 0.0 0.0 0.0
H 0.0 0.0 0.7

```

density.cube

Exemple d'entrada de Gaussian 98 per a obtenir un grid de densitat. Càlcul sobre la molècula d'hidrogen, en coordenades cartesianes, a nivell Hartree-Fock amb base mínima STO-3G. El fitxer de sortida és el *density.cube*.

Un cop calculada la funció densitat, aquesta s'ha de guardar en disc, si cal, en format adaptat per a ser l'entrada de programari reconstrucció de superfícies o de representació gràfica, els quals es tractaran en el següent subapartat.

3.3.5. Reconstrucció i representació de la superfície: El Marching Cubes Algorithm (MCA)

Actualment, l'usuari poc s'ha de preocupar per la reconstrucció de superfícies, ja que aquesta funció es troba ja implementada en diversos conjunts de programari, els quals a partir del fitxer d'entrada, adequadament formatat, i amb la introducció del valor de l'isocontorn que es desitja visualitzar, el programa el recrearà i representarà internament, de manera totalment transparent, excepte per diferents aspectes més trivials, com són selecció de colors, aspecte o textures. Alguns programes que inclouen aquesta funció són:

- GaussView [67].
- OrbDraw [68].
- GOpenMol [69].
- AVS [70].

Tanmateix, per diferents raons, aquest tipus de programari pot no estar disponible, i és en aquests casos on cal fer dur a terme com a un pas més la reconstrucció de la superfície a representar.

Un dels algorismes més emprats, entre d'altres existents (veure per exemple [71-78]), en la reconstrucció de superfícies és el dels Marching Cubes (MCA) [79]. Aquest algorisme s'utilitza en *renderització*⁹ volumètrica per a construir isosuperfícies a partir de camps tridimensionals de valors. Una analogia en el pla 2D consistiria en prendre una imatge, i per a cada píxel, fixar el color a negre si el valor d'una funció preestablerta cau sota d'un llindar, i a color blanc en cas contrari. Llavors cal delimitar la superfície a partir d'ajuntar els punts negres exteriors amb línies. L'MCA avalua els valors de vèrtexs de cubs (anomenats *voxels*) dins del camp escalar i els cataloga segons si superen o no arriben al llindar predeterminat. Aquest procés porta a un conjunt de cubs amb les cantonades classificades. Com que cada cub té 8 costats, les quals poden estar per sota o per sobre el llindar, això porta a 256 diferents possibilitats per a cada cub. Llavors, per a cada cub, cal delimitar les superfícies segons escau a cada classificació. La següent figura mostra una sèrie de cubs, en 2D, i les seves superfícies associades:

⁹ La renderització (de l'anglès *rendering*) és una tècnica de construcció d'imatges fotorealístiques basada en com un objecte és vist donades la posició d'observació i la disposició de la il·luminació.

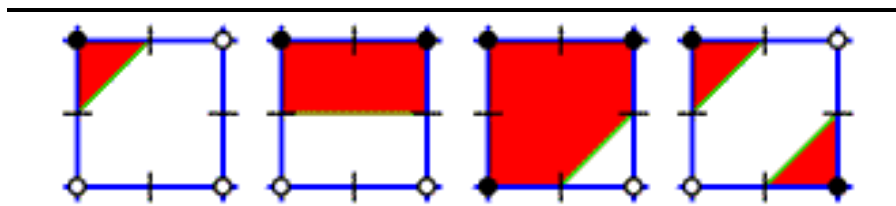


Figura 9.- Exemples bidimensionals de cubs amb les seves superfícies associades. En negre estan els punts que cauen per sota la superfície, i en blanc els que la superen. Els punts intermitjos corresponen, per interpolació, al valor aproximat del llindar. Per unió d'aquests punts, es crea la superfície.

El resultat de l'MCA és una superfície tassellada (construïda a partir de múltiples triangles), la qual aproxima l'isosuperfície de valor constant per a un llindar donat.

3.4. Exemples gràfics

En aquests apartat, es mostraran diferents exemples de MIDCOs, obtinguts a partir del protocol mostrat anteriorment. Primer de tot es comença amb una comparativa d'isocontorns calculats a partir de diferents metodologies de càlcul de funcions densitat, exposant també diferents exemples d'anàlisi de curvatura dels isocontorns. En segon lloc, es mostraran MIDCOs per a molècules de mida més gran a nivell PASA. Finalment, es presenten isocontorns per a proteïnes, obtinguts fàcilment a partir de l'ús de funcions densitat PASA.

3.4.1. Comparativa de MIDCOs

En aquest subapartat, les molècules que s'han utilitzat per a comparar els valors d'autosemblança, així com el temps necessari per al seu càlcul, s'utilitzen per a comparar els seus respectius isocontorns de densitat a diferents nivells. La tria d'aquestes molècules es va fer en base a Mezey [43], permetent una font addicional per a la comparació.

Els resultats gràfics, emprant una resolució de 0.25 unitats atòmiques, es poden veure de la figura 9 a la 14:

Nivell	ASA			ab initio (HF)		
	Pm. Simple	Pm. Límit	Mol. 3-21G	STO-3G	3-21G	6-31 ⁺⁺ G ^{**}
0.01						
0.15						
0.22						
0.35						
0.45						

Figura 9.- Comparativa de MIDCOs per a la molècula d'aigua.

Nivell	ASA			ab initio (HF)		
	Pm. Simple	Pm. Límit	Mol. 3-21G	STO-3G	3-21G	6-31 ⁺⁺ G ^{**}
0.01						
0.15						
0.22						
0.35						
0.45						

Figura 10.- Comparativa de MIDCOs per a la molècula de metanol.

Nivell	ASA			ab initio (HF)		
	Pm. Simple	Pm. Límit	Mol. 3-21G	STO-3G	3-21G	6-31 ⁺⁺ G ^{**}
0.01						
0.15						
0.22						
0.35						
0.45						

Figura 11.- Comparativa de MIDCOs per a la molècula d’etanol.

Nivell	ASA			ab initio (HF)		
	Pm. Simple	Pm. Límit	Mol. 3-21G	STO-3G	3-21G	6-31 ⁺⁺ G ^{**}
0.01						
0.15						
0.22						
0.35						
0.45						

Figura 12.- Comparativa de MIDCOs per a la molècula d’alcohol al·líc.

Nivell	ASA			ab initio (HF)		
	Pm. Simple	Pm. Límit	Mol. 3-21G	STO-3G	3-21G	6-31 ⁺⁺ G ^{**}
0.01						
0.15						
0.22						
0.35						
0.45						

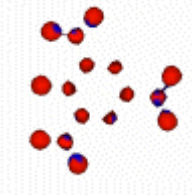
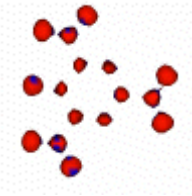
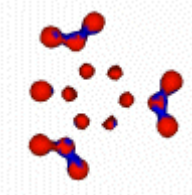
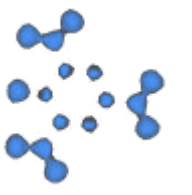
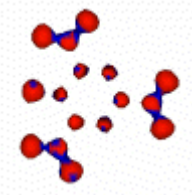
Figura 13.- Comparativa de MIDCOs per a la molècula de benzè.

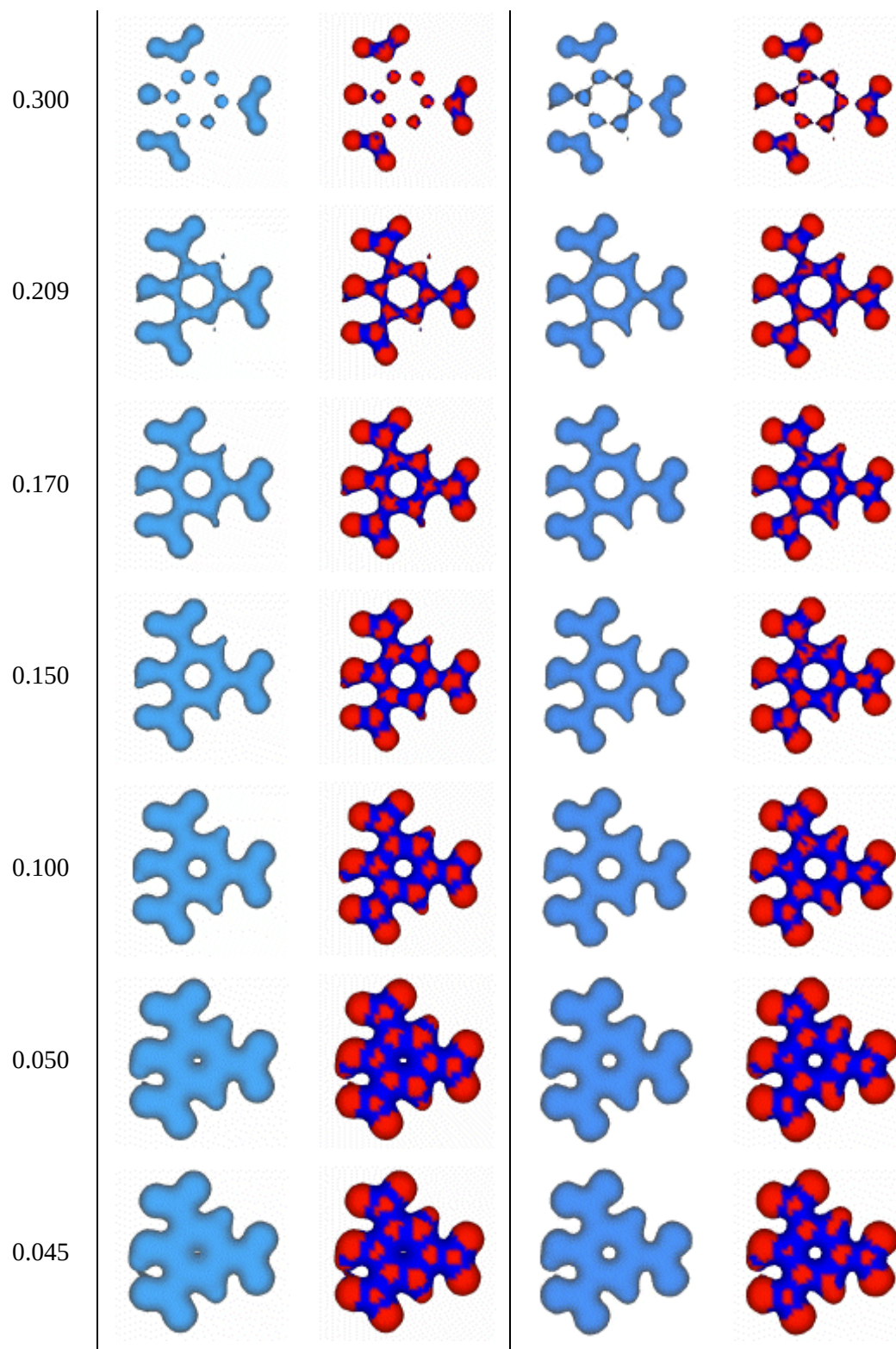
Nivell	ASA			ab initio (HF)		
	Pm. Simple	Pm. Límit	Mol. 3-21G	STO-3G	3-21G	6-31 ⁺⁺ G ^{**}
0.01						
0.15						
0.22						
0.35						
0.45						

Figura 14.- Comparativa de MIDCOs per a la molècula de glicina.

Tal i com es pot veure de la figura 8 a la 14, tant els valors elevats de densitat (0.45) com els baixos (0.01), les superfícies obtingudes amb qualsevol de les metodologies són gairebé idèntiques, donant una representació dels àtoms pesants i el *contorn molecular* exterior, respectivament. A valors intermitjos s'observen les formacions dels enllaços i l'aparició dels àtoms d'hidrogen. Cal remarcar que aquestes dues últimes característiques no estan tant ben descrites en les metodologies ASA com ho estan en les ab initio, degut fonamentalment a que la densitat ASA es troba col·lapsada als nuclis en forma de funcions 1S. En el cas dels MIDCOs per la PASA simple, els àtoms d'hidrogen no arriben a aparèixer com a entitats independents, sinó que tan sols es poden intuir, a nivells intermitjos, a partir de deformacions en els contorns dels àtoms pesants. Així i tot, els MIDCOs ASA permeten clarament diferenciar entre els diferents tipus d'enllaços presents a les molècules. Tanmateix, tenint en compte les semblances a nivell gràfic i els temps requerits en el seu càlcul, el seu ús esdevé justificable.

Seguidament es presenta una comparativa similar, on addicionalment s'ha realitzat un anàlisi de curvatura sobre els diferents contorns. L'anàlisi de curvatura permet classificar zones dels isocontorns en base a la curvatura que presenta la superfície: convexa o còncava. En aquest exemple l'estructura molecular emprada és el 2,4,6-trinitrofenol, la densitat de la qual s'ha calculat sobre la base 6-311G. En l'aproximació promolecular, s'ha utilitzat un ajust amb funcions 1S sobre aquesta base, i en la representació ab initio, la densitat s'ha calculat mitjançant el mètode de Hartree-Fock [57]. Els fitxers de grid de densitat van ser processats amb el paquet gràfic AVS [70], i els resultats es mostren a la següent figura 15.

Isovalor	Promolecular ASA		Ab initio	
	Densitat	Curvatura	Densitat	Curvatura
0.500				
0.400				



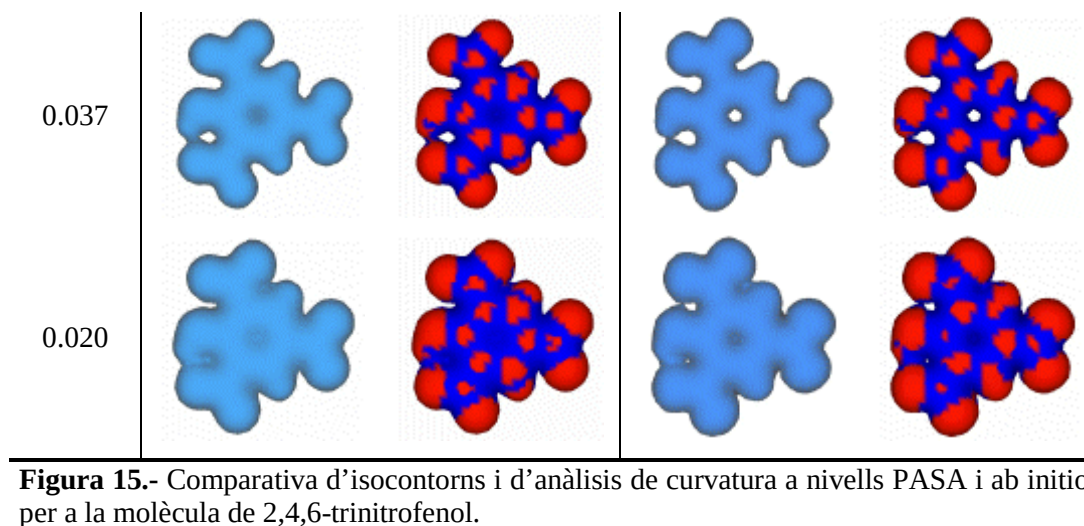


Figura 15.- Comparativa d'isocontorns i d'anàlisi de curvatura a nivells PASA i ab initio per a la molècula de 2,4,6-trinitrofenol.

La figura 15 mostra les mateixes tendències que els exemples anteriors. A nivells elevats de densitat (superiors a 0.300), els isocontorns mostren bàsicament les localitzacions dels àtoms pesants. A valors de densitat intermitjos (0.150-0.300) es representen els patrons d'enllaços i els àtoms d'hidrogen. I finalment a valors petits de densitat (inferiors a 0.150) es remarquen les característiques estèriques i d'ocupació de l'espai. També d'igual manera es poden veure alguns dels defectes que incorpora la PASA, com és el cas de l'aparició tardana dels àtoms d'hidrogen. Al nivell 0.300, els hidrògens comencen a aparèixer en els MIDCOs ab initio, mentre que aquests no apareixen en els MIDCOs PASA fins al nivell 0.209; nivell al qual els àtoms d'hidrogen ja es troben enllaçats a l'anell benzènic dins de la representació ab initio. Tanmateix, cal remarcar el temps necessari per a calcular les representacions, ja que mentre el grid PASA fou calculat en segons, el grid ab initio va requerir 20 minuts en la mateixa màquina. A més a més, els MIDCOs PASA obtinguts per a valors elevats són suficientment semblants per a poder realitzar comparacions qualitatives, i els de valors baixos esdevenen gairebé indistingibles en ambdues metodologies.

Apart de la simple comparació dels isocontorns, l'anàlisi de curvatura confirma l'aplicabilitat de les densitats promoleculars en la representacions de la forma molecular. Les regions de la superfícies properes als nuclis atòmics presenten una forma convexa, mentre que les rodalies donen una aparença local de punt de sella (dominis D_1 en la terminologia de Mezey [43]). A mesura que el nivell de densitat de referència es redueix, les zones convexes s'expandeixen, fins a obtenir-se, teòricament, una superfície totalment convexa per a nivells de densitat molt baixos. Tal i com es pot observar a la figura, les tendències en les curvatures són equivalents, donant crèdit a la validesa de les densitats PASA per a calcular MIDCOs realístics i computacionalment econòmics.

3.4.2. Isocontorns PASA per a molècules grans

En aquest subapartat es mostren MIDCOs generats per a dues molècules de mida més gran que les anteriors per a poder comprovar l'eficiència de la metodologia PASA per a construir superfícies. Les molècules presentades són el 2a-methyl-9a-fluoro-cortisol i un derivat de la benzamidina, la densitat PASA de les quals ha sigut construïda sobre ajustos a la base 3-21G, calculada a una resolució de 0.5 unitats atòmiques i representada amb el paquet gràfic AVS [70].

Els MIDCOs resultants, per a la molècula de 2a-methyl-9a-fluoro-cortisol, es mostren a la següent figura 16:

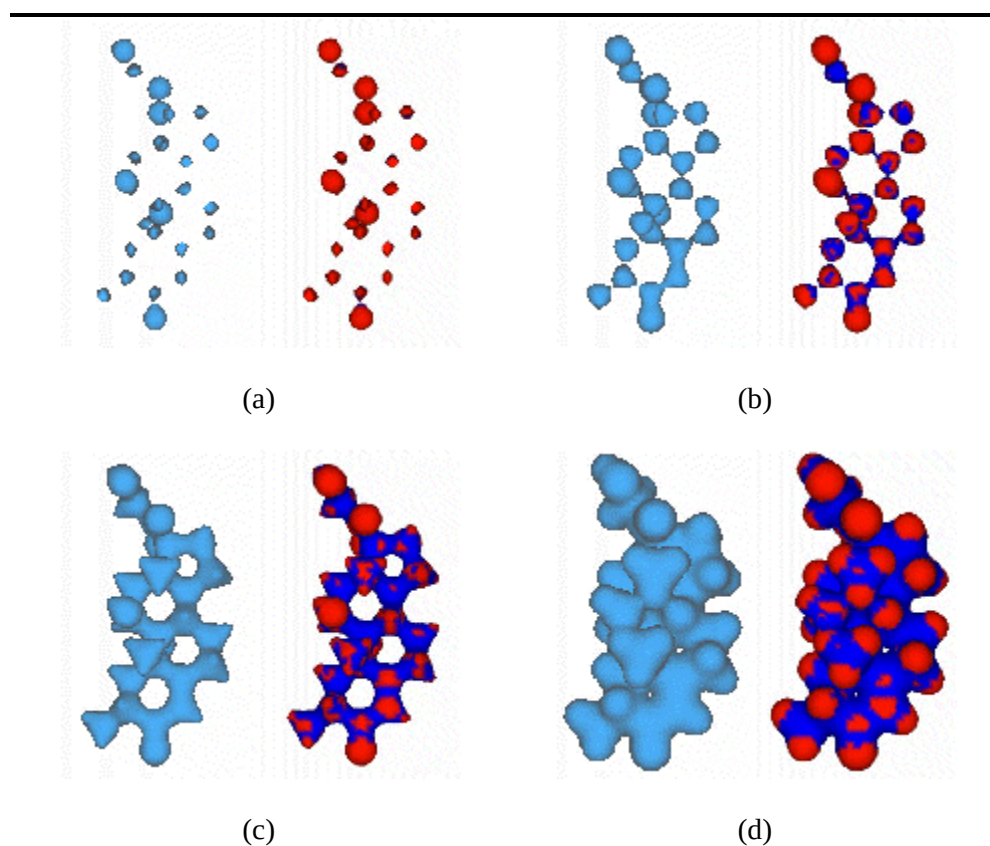


Figura 16.- Isocontorns per a la molècula de 2a-methyl-9a-fluoro-cortisol. Els nivells representats són (a) 0.371, (b) 0.212, (c) 0.132 i (d) 0.026.

I els MIDCOs corresponents al derivat de la benzamidina a la següent figura 17:

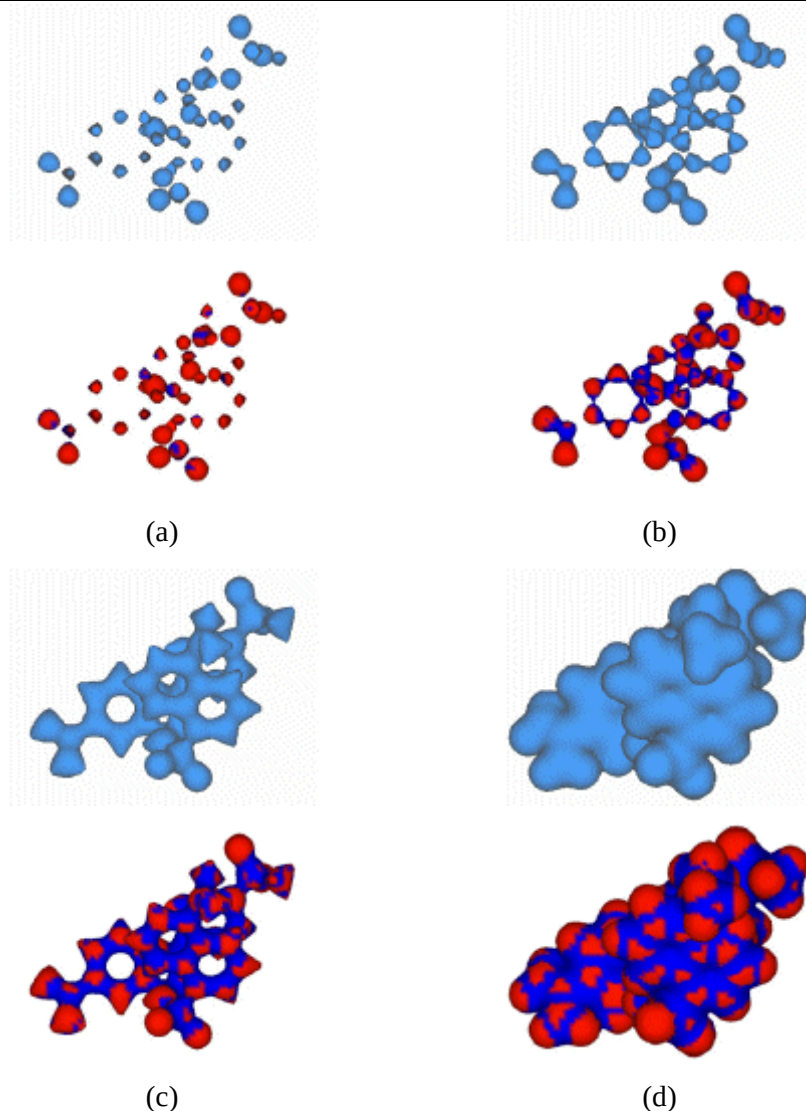


Figura 17.- Isocontorns per un derivat de la benzamidina. Els nivells representats són (a) 0.350, (b) 0.250, (c) 0.150 i (d) 0.010.

En ambdues figures s'observen, al igual que els casos anteriors, les mateixes tendències tant en les representacions de densitat com en els MIDCOs on es realitza l'anàlisi de curvatura; és a dir, a valors alts de densitat es localitzen les posicions atòmiques, a valors mitjans tota l'estructura d'enllaços, i finalment a valors baixos es mostra l'ocupació de l'espai. Tanmateix, els MIDCOs de les molècules presentades, formades per 59 i 75 àtoms respectivament, han estat calculats en poc menys d'un minut, oferint una qualitat apreciable per a la realització de d'anàlisis semiquantitatius.

La funció densitat PASA per les dues molècules contava de 115 i 157 funcions 1S respectivament, i els grids on s'han avaluat les densitats tenien unes dimensions de 19×33×15 i

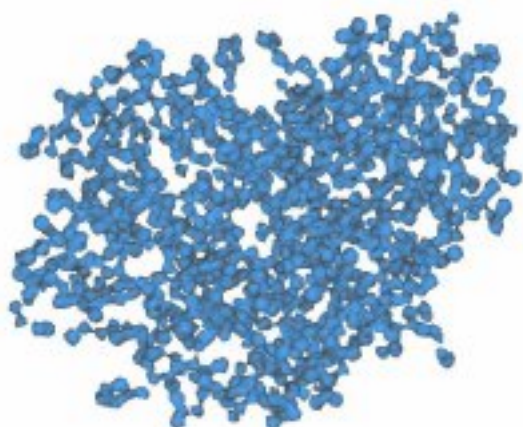
$34 \times 25 \times 29$ unitats atòmiques, que equival a 75240 i 197200 avaluacions de la funció densitat respectivament. Si es té en compte la quantitat de punts a calcular conjuntament amb el temps necessari per a avaluar-los, la utilitat del mètode esdevé clara.

3.4.3. Isocontorns PASA per a macromolècules

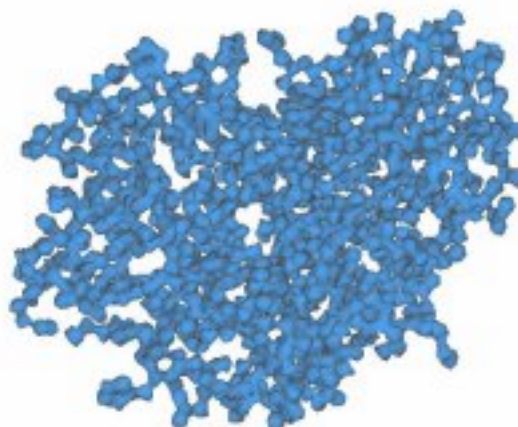
Un pas més endavant en l'aplicació de densitats ajustades en la creació de MIDCOs és intentar el seu càlcul en sistemes difícilment tractables per metodologies més complexes. Estructures moleculars composades per a més de 1000 àtoms no són tractables per mètodes *ab initio*, ni tan sols en el seu nivell més bàsic en les màquines actuals més potents, en temps raonables, deixant de banda els sistemes periòdics, pels que existeixen diferents aproximacions per al seu tractament.

Els MIDCOs que es presentaran corresponen a dues proteïnes: la mioglobina i el gen V, les coordenades de les quals han estat extretes del Protein Data Bank [63], amb referències 1MBN i 1VQB. Aquestes estructures estan composades per 2541 i 1362 àtoms, després de extreure les aigües de solvatació presents a l'estructura cristal·lina. Per a construir les funcions densitat PASA han calgut 5065 i 2709 funcions 1S respectivament, les quals han estat avaluades en grids de $91 \times 83 \times 85$ i $95 \times 107 \times 83$ unitats atòmiques, fixant la resolució a una unitat atòmica.

Els MIDCOs obtinguts es mostren a continuació:



(a)



(b)

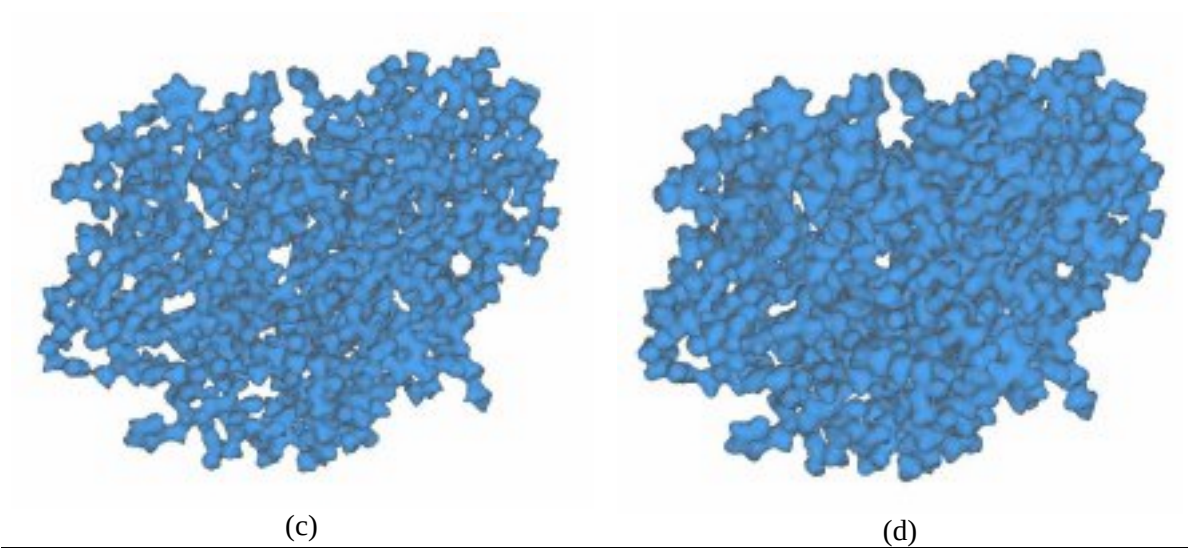


Figura 18.- MIDCOs per a la proteïna mioglobina a valors d'isodensitat (a) 0.20, (b) 0.15, (c) 0.10 i (d) 0.05.

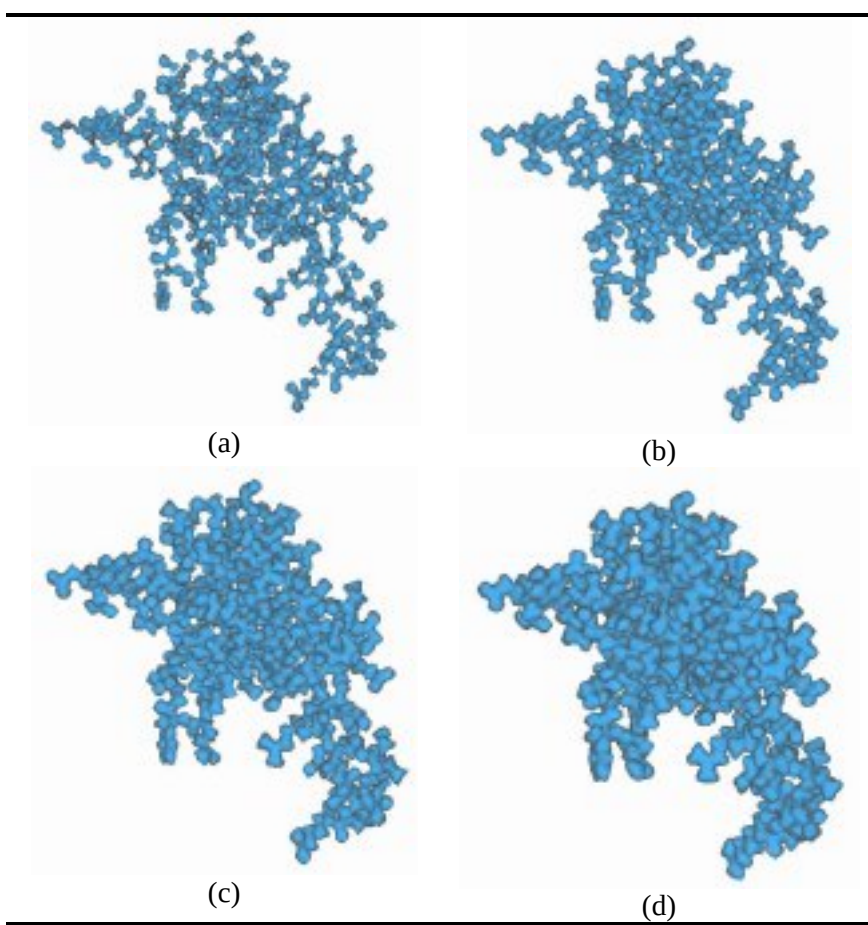


Figura 19.- MIDCOs per a la proteïna gen V a valors d'isodensitat (a) 0.20, (b) 0.15, (c) 0.10 i (d) 0.05.

Tal i com es pot veure a les figures 18 i 19, les tendències seguides pels diferents MIDCOs s'equiparen amb les prèviament exposades. El que destaca en aquestes representacions és la qualitat obtinguda en els gràfics a partir d'un protocol tan senzill com l'exposat, però sobretot, el fet més remarcable és el temps necessari per a la seva construcció: aproximadament 9 i 5 hores (en el DEC Alpha prèviament citat) per la mioglobina i el gen V, respectivament. Tot i que pot semblar una xifra elevada, la quantitat d'àtoms presents, així com de funcions 1S necessàries per a descriure el sistema, minimitza la grandesa del càlcul, possibilitant-lo en equips relativament modestos.

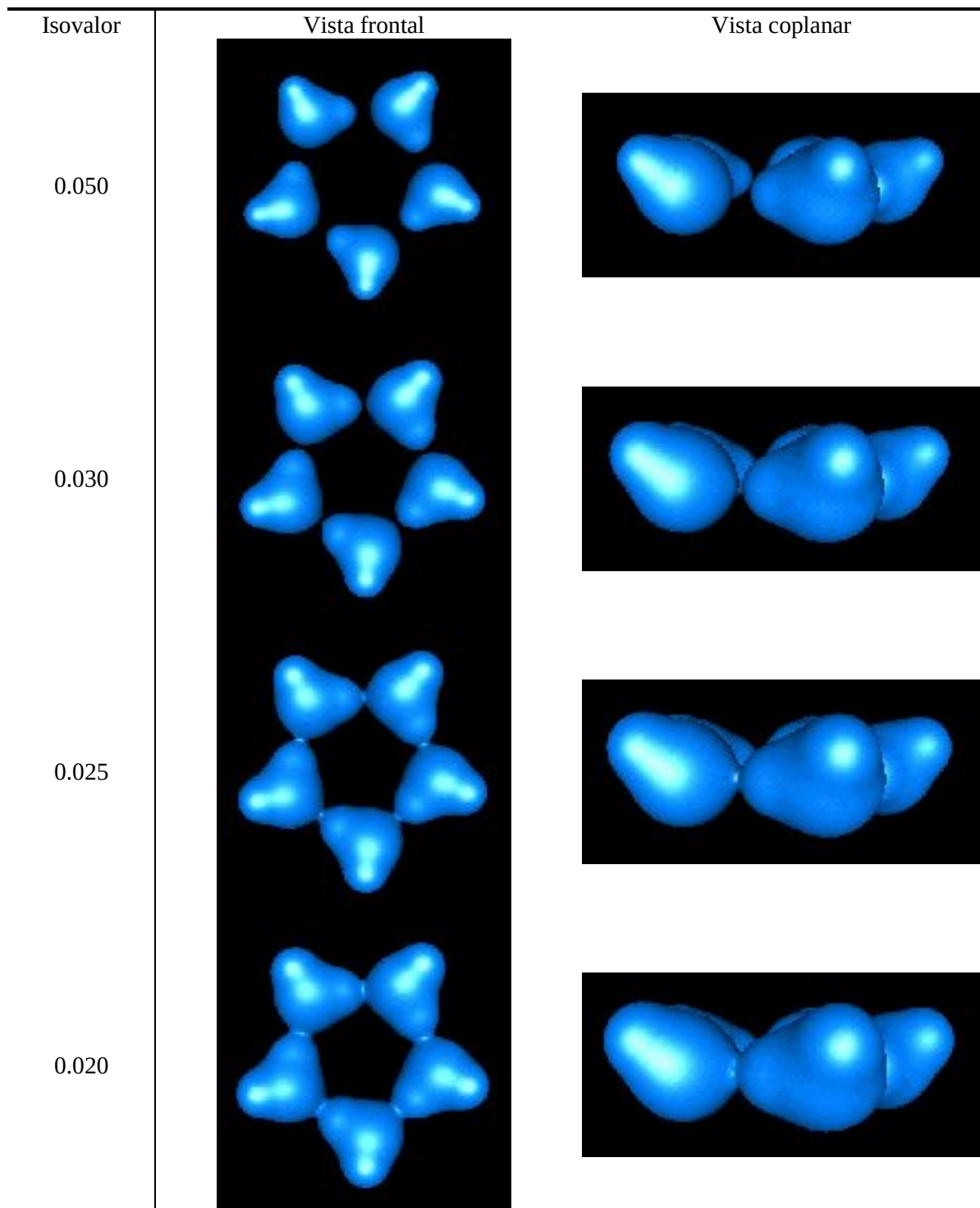
Les representacions del gen V, prèviament presentades per Mezey dins de l'entorn del mètode MEDLA [80,81], permeten una font de comparació de resultats. El MEDLA (de l'anglès *Molecular Electron Density Lego Approach*, o Aproximació a la densitat electrònica molecular per 'peces') consisteix en un procediment que permet la construcció de MIDCOs de qualitat ab initio en temps lineal, fins i tot per a macromolècules, a partir de fragments moleculars difosos prèviament calculats, els quals, quan són adequadament rotats i traslladats, reconstrueixen els contorns de densitat electrònica de la molècula original. Els fragments necessaris es calculen en un entorn similar al que tindrien dins de la molècula a representar, assumint que la densitat electrònica és gairebé *transferible*. Els fragments calculats comporten força càlculs ab initio per a sistemes petits, els quals són ràpidament duts a terme, podent-se guardar en una base de dades per a possibles usos futurs. Degut justament a aquests càlculs previs de fragments amb entorns similars, el MEDLA permet reflexar en la superfície interaccions febles de manera més detallada que no la PASA, la qual només està fonamentada sobre densitats atòmiques discretes. Tanmateix, un càlcul MEDLA elaborat, tot i que un cop els fragments han estat calculats la superfície serà ràpidament construïda, requerirà la creació o ús de múltiples porcions moleculars, encarint-se la globalitat del càlcul.

Més detalls sobre el càlcul, així sobre la funció biològica de les proteïnes presentades es troba a l'article adjunt Gironés, Carbó-Dorca, Mezey. *J. Mol. Graph. Model.* 19, 343-348.

3.4.4. Isocontorns PASA per a sistemes interaccionants

Tot i que les densitats ajustades no modelen de manera comparable als ab initio les interaccions moleculars degut al col·lapse de la densitat a les posicions atòmiques, els isocontorns PASA de sistemes moleculars que els presentin poden ser d'interès si més no qualitatiu. La següent figura presenta com a exemple una de les múltiples configuracions del pentàmer d'aigua ((H₂O)₅) [82], on en aquest cas s'adopta una disposició cíclica amb els hidrògens en *cis*. Aquesta estructura s'ha

elucidat experimentalment, i tot i no ser un mínim global, constitueix un mínim local. Els MIDCOs que es presenten per aquest sistema corresponen als nivells on s'observa la interacció per pont d'hidrogen intermolecular.



0.015

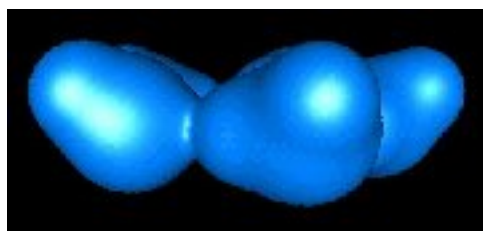
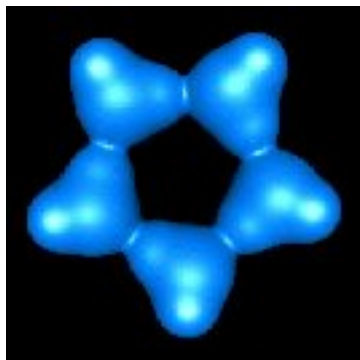


Figura 20.- MIDCOs PASA pel pentàmer d'aigua.

És interessant veure com es va formant amb el valor decreixent de la densitat l'enllaç entre els oxígens i els hidrògens de les molècules adjacents, podent-se fer interpretacions qualitatives dins del marc de les densitats ajustades. El càlcul dels MIDCOs, amb els ajustos corresponents a la base 6-311G i amb una resolució de 0.25 unitats atòmiques, fou completat en 15 segons escassos en un P200MHz.

3.5. Refinaments del mètode

Tal i com s'ha comentat en el sub-apartat de representació de MIDCOs de macromolècules, els temps de càlcul per a la representació d'aquestes estructures, tot i que acceptable en termes relatius de costos computacionals, pot arribar a ser excessiu si cal representar múltiples estructures o el temps de càlcul resulta un factor limitant. A fi de retallar al màxim el temps de càlcul i convertir-lo en un procés lineal del nombre d'àtoms, es van plantejar diferents refinaments, tenint en compte el comportament de la funció densitat i tècniques acurades de programació, tot aprofitant les idees del mètode MEDLA i la rapidesa de càlcul amb funcions PASA.

3.5.1. Particionament de l'espai

És més que conegut que la funció densitat electrònica es concentra bàsicament en les posicions atòmiques, i que aquesta presenta una caiguda exponencial amb la distància del seu centre. Aquesta tendència matemàtica es pot veure clarament a les figures 21 i 22, on es presenten un gràfic monodimensional de la densitat per l'àtom d'oxigen i un de bidimensional per la molècula de benzè. En ambdós casos, els paràmetres PASA emprats corresponen a ajustos a la base 6-311G o càlculs Hartree-Fock en la mateixa base [57].

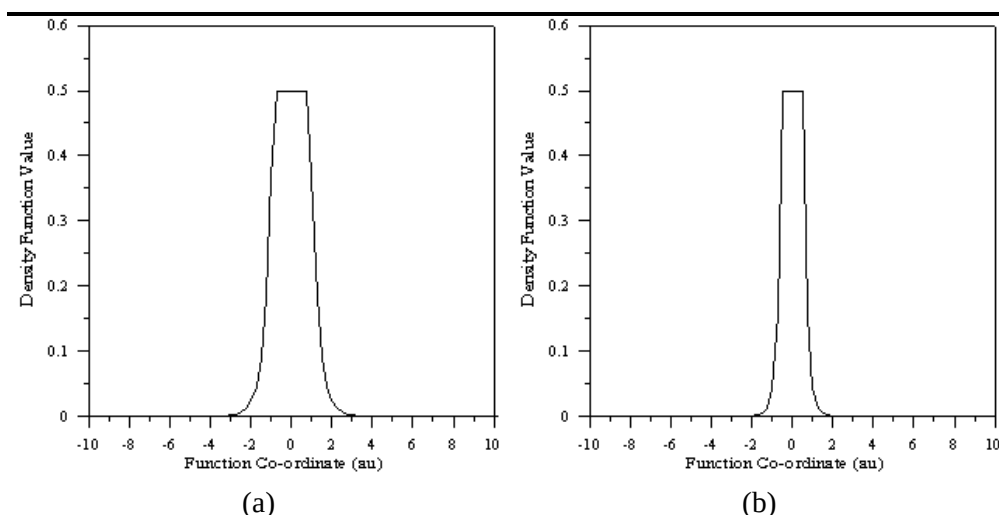


Figura 21.- Representació de densitat l'àtom d'oxigen, (a) a nivell PASA i (b) ab initio. Els valors superiors a 0.5 han estat eliminats per claredat.

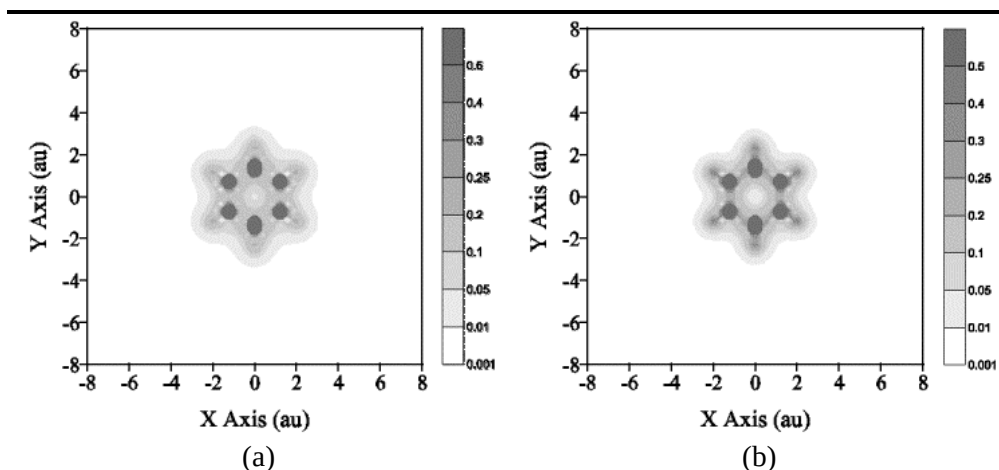


Figura 22.- Representació de densitat per la molècula de benzè, (a) a nivell PASA i (b) ab initio.

Les figures anteriors, apart de les evidents semblances entre ambdues metodologies, corroboren l'anterior afirmació sobre el comportament de la funció densitat al allunyar-se dels nuclis on està centrada.

Tanmateix, la figura 21 podria donar una visió distorsionada de la realitat, al considerar tan sols un àtom. Es podria pensar que a distàncies curtes del centre, la densitat assoleix valors tan baixos que ja no val la pena considerar-los. Per això, la següent figura 22 demostra com a valors baixos de densitat, aquesta continua sent important, i més des d'un punt de vista químic, ja que els enllaços químics es manifesten a valors força baixos comparats amb els assolits sobre els nuclis. Al mateix temps, no només els enllaços són químicament rellevants, sinó també la distribució exterior de la densitat, la qual ha de ser correctament modelada a fi de tenir un model vàlid i realístic.

Tal i com s'ha comentat en subapartats anteriors, la generació de MIDCOs comporta la creació d'un grid i l'avaluació de la densitat a cada punt d'aquesta xarxa. Aquest procés comporta un increment cúbic dels recursos, excepte per sistemes simètrics, on tan sols caldria ampliar una o dues direccions de l'espai. A fi de convertir aquest procés en lineal, es proposa calcular només la densitat de cada àtom que constitueix la molècula en un entorn proper a ell, i no al llarg de tot el grid, on la contribució en punts llunyans és totalment menyspreable, o fins i tot no avaluable per raons de precisió i arrodoniment numèric. Tenint compte el comportament de la funció densitat, ni tan sols cal construir entorns monoatòmics molt grans, i l'experiència acumulada ha demostrat que subgrids de $10 \times 10 \times 10$ unitats atòmiques són més que suficients, ja que la densitat a aquesta distància baixa per sota de 0.002 per a la majoria dels àtoms més comuns de les molècules orgàniques. A més a més, aquest valor troba també justificació en la literatura, ja que 0.002 significa abarcar més del 98% de la càrrega electrònica, reproduir en gran mesura els volums CPK [1] i també es va trobar que reproduïa correctament els volums de molècules diatòmiques [83]. Així, dins d'aquests subgrids de 1000 au^3 queda continguda tota la informació rellevant, tal com el pic de densitat centrat a cada àtom, el patró d'enllaços (si s'escau) i la forma general de la distribució de càrrega exterior.

En la següent figura 23 es mostra un esquema de particionament en subgrids per a la molècula de 2-amino,2-hidroximetil,1,3-propandiol.

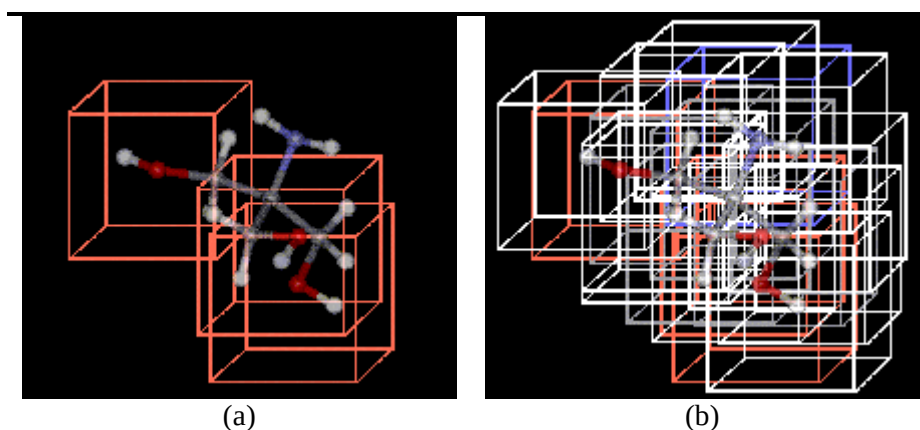


Figura 23.- Esquema de particionament de l'espai per al càlcul de MIDCOs, (a) espais de càlcul dels oxígens (b) totalitat dels subgrids pels diferents àtoms: oxigen (vermell), nitrogen (blau), carboni (gris) i hidrogen (blanc) per a la molècula de 2-amino,2-hidroximetil,1,3-propandiol.

La molècula de 2-amino,2-hidroximetil,1,3-propandiol està composta per 19 àtoms, pel que l'espai es divideix, tal i com es veu a la figura 23b en 19 subregions que envolten cadascun dels àtoms.

També, i tal com s'havia comentat anteriorment, es pot observar com el volum que s'assigna a cada àtom és suficient com per abarcar la majoria d'interaccions rellevants amb altres àtoms de la molècula. D'aquesta manera, l'estalvi en temps de càlcul és considerable, ja que les funcions que pertanyen a un àtom seran avaluades només en aquesta regió, enlloc de en tot l'espai definit pel grid molecular. D'aquí la semblança conceptual amb la metodologia MEDLA, ja que l'espai es troba definit en base a fragments molecular, en el nostre cas àtoms, només que la densitat encara no està avaluada.

3.5.2. Definició de variables-àtom

Un altre estalvi de recursos que es pot dur a terme, el qual contribueix notablement en l'eficiència del càlcul és la minimització en el nombre de funcions presents en memòria en la realització de l'avaluació de la funció densitat.

En la metodologia PASA i donat un conjunt de base de funcions ajustades, els àtoms del mateix tipus venen definits per el mateix tipus i nombre de gaussianes 1S. Llavors, poc sentit té definir per a cada àtom el seu conjunt de funcions, si aquestes ja han estat definides per a un àtom previ del mateix tipus. El que es proposa és la definició de *variables-àtom*: conjunt de paràmetres (coeficients i exponents de funcions 1S) definits per a cada tipus distint d'àtom present en la molècula a estudiar, les quals poden ser referenciades dinàmicament en temps d'execució. D'aquesta manera, i independentment del nombre d'àtoms que tingui realment la molècula, el conjunt de funcions ajustades presents en memòria serà sempre relativament petit, estalviant-se més memòria en funció de la mida creixent de la molècula.

En el següent tall de codi 3 es mostra el codi Fortran 90 per a la definició de les variables-àtom:

```
! nf: nombre de funcions per un àtom en concret
! c: vector que conté els coeficients de les funcions
! ex: exponents de les funcions 1S
TYPE :: atom
  INTEGER*1 nf
  REAL, DIMENSION(:), POINTER :: c, ex
END TYPE atom
```

Tall de Codi 3.- Definició de les *variables-àtom*.

El següent pas és la determinació del nombre d'àtoms diferents dins la molècula, i la seva identificació. El mètode emprat es mostra al següent tall de codi 4:

```

! Se suposa la lectura de les coordenades moleculars a
! partir d'un fitxer estàndard XYZ
! xyz: matriu de les coordenades moleculars
! sym: variable alfanumèrica on s'emmagatzema el símbol
!      atòmic llegit
! simbols: vector alfanumèric que conté tots els símbols
!          atòmics
! presents: vector lògic que enregistra l'incidència
!           d'un àtom. Iniciat a .false.
!
open (UNIT=10, FILE=nom, STATUS='old')
! lectura del nombre d'àtoms
read (10,*) na
ALLOCATE (xyz(na,3), nz(na))
read (10,*) ! dummy line
! el nombre d'àtoms diferents
nad = 0
do i=1,na
  read (10,*) sym,xyz(i,:)
  do j=1,105
    if (sym.eq.simbols(j)) then
      if (.not.presents(j)) then
        nad=nad+1
        presents(j)=.true.
      end if
      ! determinació dels nombres atòmics
      nz(i)=j
    end if
  end do
end do
close (10)

```

Tall de Codi 4.- Determinació i identificació dels diferents àtoms que formen una molècula.

Sabent quants i quins àtoms diferents formen la molècula, es pot procedir a la lectura de la base per a cada àtom, la qual serà mantinguda i referenciada en memòria en temps d'execució. El següent tall de codi 5 en mostra el procediment seguit, suposant que el fitxer base té el següent format:

- Símbol atòmic (2 caràcters)
- Nombre de funcions de base (nombre enter)
- coeficient₁, exponent₁ (nombres reals)
- coeficient₂, exponent₂
- ...
- següent definició

```

do i=1,105 ! es busca per els 105 primers àtoms existents
  if (presents(i)) then
    open (UNIT=10, FILE=base, STATUS='old')
    1 read (10,'(a)') sim
    if (sim.eq.simbols(i)) then
      read (10,*) atoms(i)%nf
      ALLOCATE (atoms(i)%c(atoms(i)%nf),atoms(i)%ex(atoms(i)%nf))
      do j=1,atoms(i)%nf
        read (10,*) atoms(i)%c(j),atoms(i)%ex(j)
      end do
      go to 2
    end if
    go to 1
    2 continue
  close (10)
end if
end do

```

Tall de Codi 5.- Assignació de la base als àtoms.

3.5.3. Càlcul de la densitat

Definit el grid molecular, els subgrids per a cada àtom i les variables-àtom, només cal calcular la densitat en cada una de les caixes que envolten a cada àtom. El càlcul de la densitat en aquest cas, només comporta l'avaluació de les funcions que defineixen l'àtom de la caixa i no la resta.

Resumint, el nou protocol seguit es simplifica a:

- a. entrada de coordenades moleculars
- b. definició dels paràmetres de la densitat PASA per a cada àtom distint
 - i. reserva d'espai en memòria
 - j. definició de variables-àtom per als àtoms distints
- c. definició del grid molecular: límits (automàtics) i resolució
- d. posada a zero de cada punt del grid molecular
- e. per a cada àtom:
 - i. definició d'un subgrid $10 \times 10 \times 10 \text{ au}^3$ dins del grid molecular, envoltant l'àtom.
 - j. avaluació de les funcions de l'àtom a cada punt del subgrid emprant les funcions prèviament definides. Cas que el punt fos prèviament calculat (pertanyent a un altre subgrid) afegir el valor calculat al previ.
- f. eliminació de tota informació irrellevant de la memòria
- g. descàrrega del grid molecular a disc en format adequat al paquet gràfic a utilitzar.

Tal i com es pot veure en el pas *e-j* del protocol, es considera clarament que el valor de la funció densitat en un punt està format pràcticament per la superposició de funcions discretes localitzades sobre els àtoms més propers, assumint que la contribució d'un àtom distant és zero, i per tant descartada. Així mateix, aquest pas proporciona les bases de la dependència lineal del temps de càlcul del grid molecular amb el nombre d'àtoms presents a la molècula.

3.5.4. Exemple gràfic

Per a il·lustrar aquest nou protocol presentat, es presenten una sèrie de MIDCOs per a la molècula de 2-amino,2-hidroximetil,1,3propandiol. Les representacions obtingudes es mostren a la següent figura 24

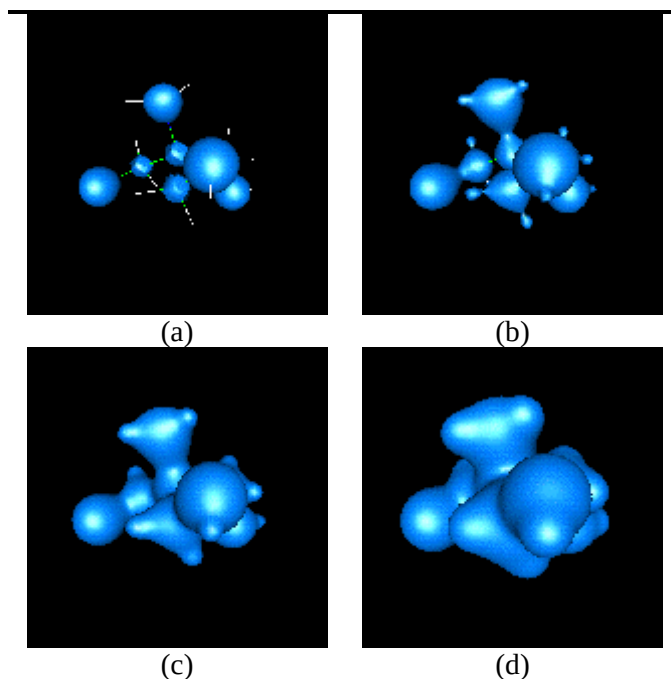


Figura 24.- MIDCOs per a la molècula de 2-amino,2-hidroximetil,1,3propandiol a valors d'isodensitat (a) 0.35, (b) 0.22, (c) 0.15 i (d) 0.05.

Com es pot observar a la figura 24, no hi ha diferències apreciables en la forma general dels MIDCOs després d'aplicar les modificacions del protocol. Tanmateix, el temps de càlcul ha estat de tan sols 3 segons per a calcular el grid de densitat en un Pentium II/400MHz/64Mb sota Linux RedHat 6.0. La molècula presentada, formada per a 19 àtoms, tan sols en presentava 4 tipus distints: hidrogen, carboni, nitrogen i oxigen, requerint només 4 variables-àtom.

Aquest exemple tan sols demostra que els MIDCOs obtinguts són totalment equivalents als resultants de l'anterior protocol, calent exemples més grans per a poder avaluar completament l'eficiència del càlcul, tal i com s'exposa en el següent subapartat.

3.5.5. Exemples macromoleculars

En aquests subapartat es mostraran diferents MIDCOs per a estructures moleculars compostes per un nombre elevat d'àtoms. D'aquesta manera, es pot verificar l'eficiència del protocol exposat en funció del temps de càlcul associat, així com de la qualitat de les representacions obtingudes. Les estructures que es presentaran com a exemples són la Proto-Oncogen Tirosina Quinasa i la Endonucleasa EC3.1.21.4 emprant densitats PASA sobre ajustos a la base 6-311G.

La Proto-Oncogen Tirosina Quinasa (referència 1ABL en el PDB[63]) és una proteïna existent en el cos humà, la funció de la qual és la transferència de grups fosfat. La seva estructura fou determinada teòricament emprant tècniques de modelatge homòleg i simulacions de dinàmica molecular [84], i es mostra a la figura 25.



Figura 25.- Estructura tridimensional de la proteïna 1ABL.

La proteïna consta de 57 residus i un total de 446 àtoms, calent un grid de $65 \times 71 \times 63 \text{ ua}^3$ i utilitzant una resolució de 0.5 ua. Tanmateix, tot i l'elevat nombre d'àtoms presents, tan sols 5 tipus diferents estaven presents: hidrogen, carboni, nitrogen, oxigen i sofre, pel que es van definir en temps d'execució només 5 variables-àtom. Els resultats obtinguts amb l'1ABL es mostren a la següent figura 26.

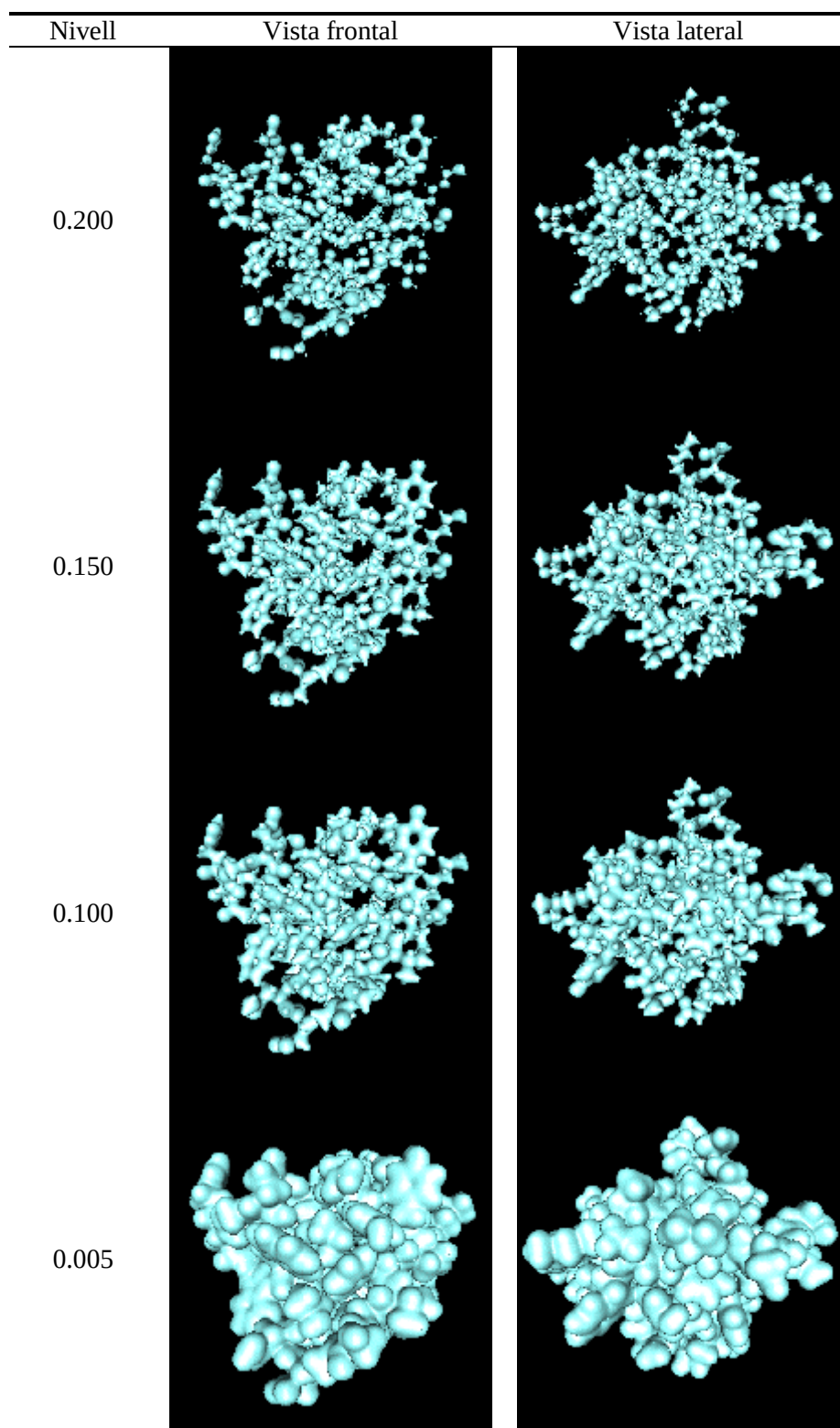


Figura 26.- MIDCOs per a la proteïna 1ABL.

Els MIDCOs obtinguts per la proteïna 1ABL presenten les mateixes tendències que els casos anteriors, donant bona idea de l'espai efectivament ocupat a nivells baixos. Més destacat que els gràfics en sí, és el temps de càlcul requerit, ja que tan sols aproximadament 10 segons foren necessaris per a computar els 446 subgrids en que es subdividia l'estructura. Val a dir que gairebé la meitat d'aquest temps de càlcul fou consumit per a desar el fitxer (traspàs de memòria a disc), i diferents proves en diferent maquinari ha revelat que aquest pot arribar a ser el pas més lent del procés¹⁰.

El següent exemple que es mostra consisteix en la proteïna Endonucleasa (codi 1RVA al PDB), estreta de la *E. Coli* i elucidada per Kostrewa i col·laboradors [85], la qual presenta en la seva estructura un fragment de ADN, tal i com es mostra a la figura 27.

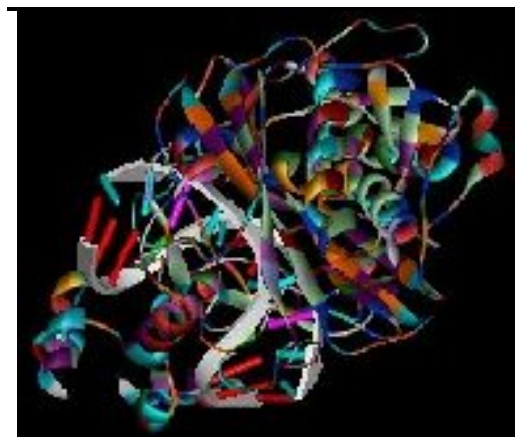


Figura 27.- Estructura tridimensional de la proteïna 1RVA.

Aquests tipus de proteïnes tenen una funció protectora pel material genètic bacterià. Degut a que les bactèries estan contínuament sota l'atac de bacteriòfags, les bactèries han desenvolupats mecanismes per a erradicar qualsevol fragment d'ADN forani, com el provinent d'un fag hostil. Aquestes bactèries que han desenvolupat aquest tipus de mecanisme sintetitzen endonucleases, o enzims capaços de 'tallar' seqüències d'ADN, les quals suren pel citoplasma bacterià en espera d'ADN de fag. D'aquí el nom d'enzim de restricció per les endonucleases, ja que restringeixen la infecció per bacteriòfags.

L'estructura original en el PDB constava 510 residus, que sumen un total de 4492 àtoms, els quals al afegir posteriorment els hidrògens, van arribar a 8686. Tanmateix, tan sols 6 tipus d'àtoms són presents a l'estructura: hidrogen, carboni, nitrogen, oxigen, fòsfor i sofre, calguent poques funcions definides en memòria. El grid molecular tenia unes dimensions de 129×121×147 unitats atòmiques, calculat a una resolució d'una unitat atòmica.

Els resultats gràfics es mostren a continuació en la següent figura 28:

¹⁰ En el cas d'un Pentim 200MHz/64Mb el grid fou calculat en uns 25 segons, però uns 4 minuts foren necessaris per a desar el fitxer a disc.

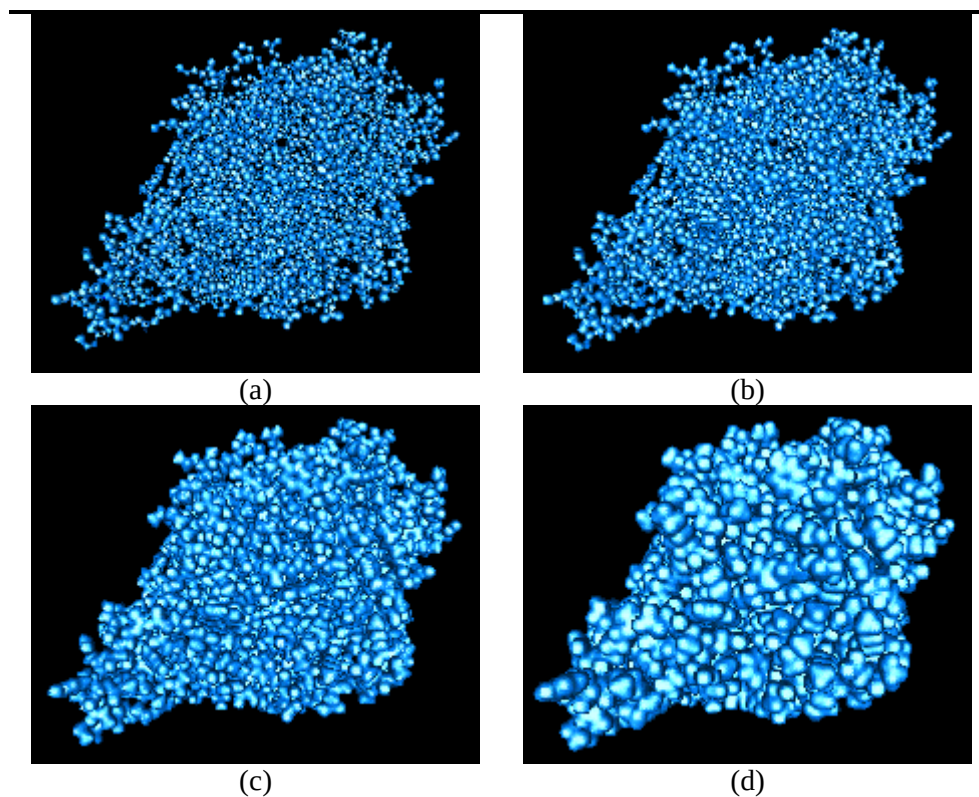


Figura 28.- MIDCOs per a la proteïna 1RVA a valors d'iso-densitat (a) 0.200, (b) 0.150, (c) 0.050 i (d) 0.005.

A l'igual que l'exemple anterior, apareixen les mateixes tendències de localització atòmica, patró d'enllaços i estericitat a mesura que el llindar de densitat decreix. Sorpren en aquest cas el temps necessari per a completar el grid molecular, ja que solament mig minut fou necessari per al seu càlcul, i anàlogament a l'exemple anterior, bona part del temps fou dedicat a desar el fitxer resultant a disc.

Aquesta proteïna, al ser elucidada amb un fragment d'ADN, permet un anàlisi d'aquest fragment, el qual es representa a la següent figura 29. Aquest fragment de material genètic constava de 700 àtoms a l'estructura original, augmentant a 770 després d'afegir els hidrògens. 5 tipus d'àtoms formen la molècula: hidrogen, carboni, nitrogen, oxigen i fòsfor. El grid molecular tenia unes dimensions de $81 \times 79 \times 55$ unitats atòmiques, i fou calculat a una resolució de 0.5 en tan sols 15 segons. Els resultats es mostren a la següent figura 30.

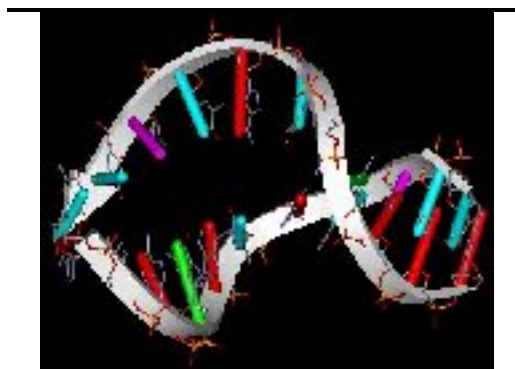


Figura 29.- Estructura tridimensional d'un fragment d'ADN.

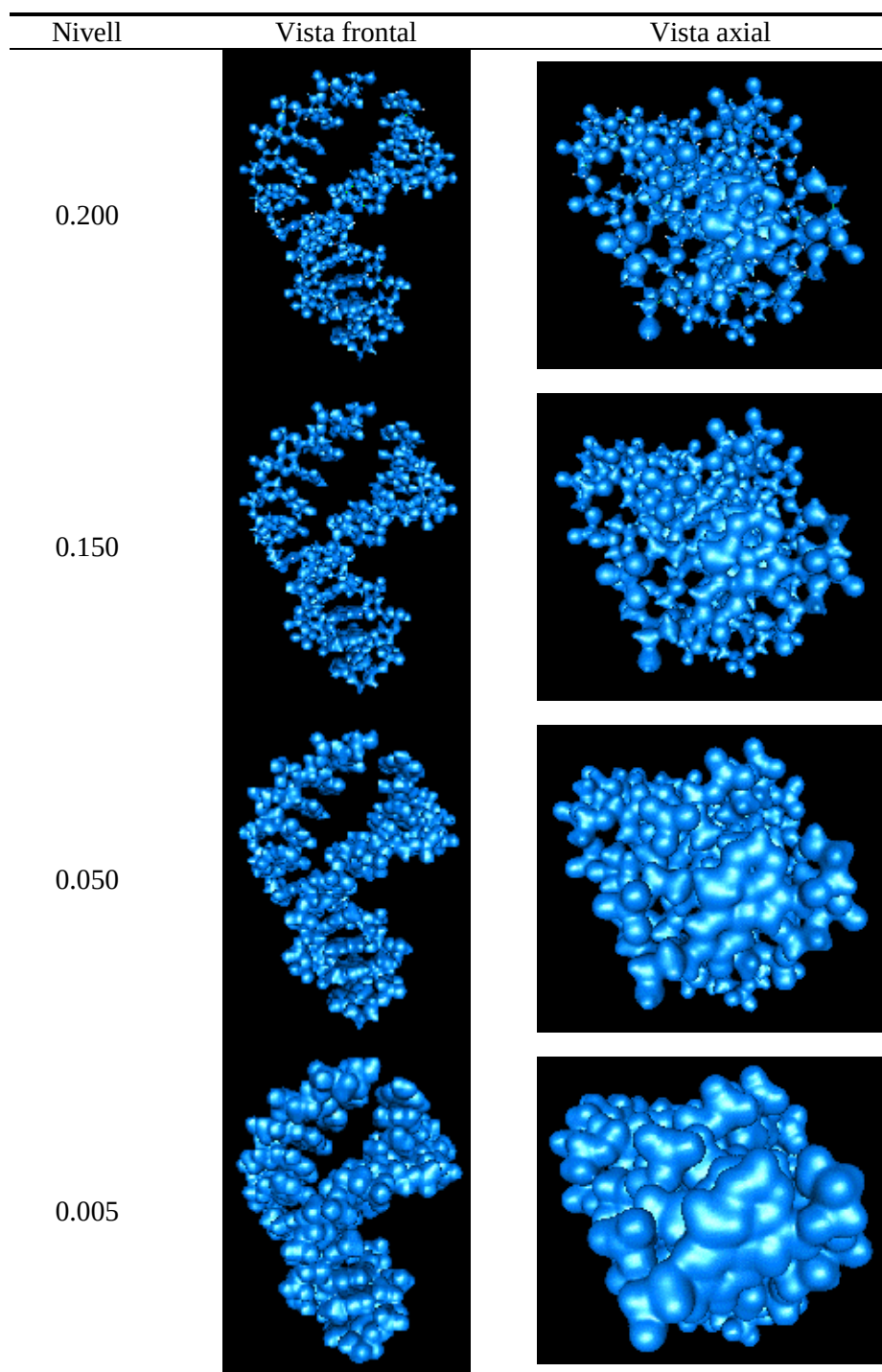


Figura 30.- MIDCOs per a un fragment d'ADN.

Tal i com es veu a la figura 30, a la part inferior de les representacions frontals es visualitza l'estructura en doble hèlix característica de l'ADN, mentre que la part superior denota la separació

de les cadenes, com a pas previ a les incisions que durà a terme la proteïna endonucleasa. Les representacions axials denoten més l'empaquetament de la molècula.

3.6. Conclusions

S'ha desenvolupat un mètode general per a la generació de contorns d'isodensitat electrònica molecular (MIDCOs), el qual permet representar gràficament, i amb uns recursos computacionals mínims, isocontorns per a una gran varietat de tipus de molècules dins de marges de temps molt reduïts. El mètode proposat s'ha treballat en tots els aspectes, des de la construcció de les funcions densitat ajustades, càlcul de la densitat en una xarxa de punts, reconstrucció de les superfícies en el valor llindar d'interès i representació gràfica.

S'ha realitzat una comparació visual entre aquests isocontorns, obtinguts a partir de densitats ajustades, i els obtinguts a partir de metodologies més acurades, així com més costoses, donant resultats molt satisfactoris, oferint una alternativa a l'ús d'aquestes últimes en les representacions de densitats electròniques. Aquesta comparació es pot afegir a treballs previs, en matèria de funcions ajustades, en semblança molecular quàntica, on els càlculs de tal mesura resulten quantitativament equivalents, bé emprant densitats ajustades o *densitats ab initio*. La comparació gràfica d'aquestes funcions ajustades permet reforçar el seu ús en treballs de semblança.

El camp d'aplicació de funcions densitats ajustades en la representació de MIDCOs ha estat ampliat mitjançant l'ús de variables-àtom i al particionament de l'espai, permetent reduir més el temps necessari per a la creació de la xarxa de punts de densitat. Aquesta etapa del procés, que aproximadament escala amb el nombre d'àtoms al cub, s'ha pogut convertir amb un procés lineal amb aquest nombre d'àtoms, podent-se abastar molècules de la mida de proteïnes amb temps assequibles en maquinari modest.

3.7. Contribucions i articles adjunts

En aquest sub-apartat final d'aquest capítol es llisten les diferents contribucions originades, totalment o parcial, per el treball exposat :

- *Construcció de superfícies d'isodensitat utilitzant funcions ASA*. Contribució oral al Taller conjunt de sobre topologia i semblança molecular, Deltebre, 3-5 juny, 1998.

- *Aplicació de funcions ASA a la construcció de superfícies d'isodensitat*. Contribució oral a la XIV Reunió del Grup de Química Teòrica de Catalunya, Girona, 29-30 juny, 1998.
- Gironés, X., Amat, L. i Carbó-Dorca, R. *A comparative study of isodensity surfaces using ab initio and ASA density functions*. J. Mol. Graph. & Mod., 16, 1998, pp 190-196. Inclós en aquest volum.
- Robert, D., Gironés, X. and Carbó-Dorca, R. *Facet diagrams for quantum similarity data*. J. Comp-Aided Mol. Des., 13, 1999, pp 597-610.
- Carbó-Dorca, R., Robert, D., Amat, L., Gironés, X. and Besalú, E. *Molecular quantum similarity in QSAR and drug design*. Springer Lecture Notes in Chemistry, 73, 2000. Springer Verlag.
- Carbó-Dorca, R., Amat, L., Besalú, E., Gironés, X. and Robert, R. *Quantum mechanical origin of QSAR: theory and applications*. J. Mol. Struct. (Theochem), 504, 2000, pp 181-228.
- Carbó-Dorca, R., Besalú, E. and Gironés, X. *Extended density functions and quantum chemistry*. Adv. Quantum Chem., 35, 2001, pp 3-63.
- Gironés, X., Carbó-Dorca, R. and Mezey, P.G. *Application of promolecular ASA densities to graphical representation of density functions of macromolecular systems*. J. Mol. Graph. & Mod., 3-4, 2001, pp 343-348. Inclós en aquest volum.
- Carbó-Dorca, R., Amat, L., Besalú, E., Gironés, X. and Robert, D. *Quantum molecular similarity: theory and applications to the evaluation of molecular properties, biological activities and toxicity*. In: *The Fundamentals of Molecular Similarity*. Carbó-Dorca, R., Gironés, X., and Mezey, P.G. (eds.). Kluwer Academic / Plenum Publishers, New York, 2001
- Gironés, X., Amat, L. and Carbó-Dorca, R. *Modelling Large Macromolecular Structures using Promolecular Densities*. J. Chem. Inf. Comput. Sci. Acceptat. Inclós en aquest volum.

3.8. Referències

1. Corey, R. B. and Pauling, L. *Molecular models of amino acids, peptides and proteins*. Rev. Sci. Instr. 24, 1953, pp 621-627.
2. Yankeelov, J. A., Jr. and Coggins, J.R. *Construction of space-filling models of proteins using dihedral angles*. Cold Spring Harbor Symposia on Quantitative Biology. 36, 1971, pp 585-587.
3. Kendrew, J. C., Dickerson, R.E., Strandberg, B.E., Hart, R.G., Davies, D.R., Phillips, D.C. and Shore, V.C. *Structure of myoglobin: A three-dimensional Fourier synthesis at 2 Å resolution*. Nature 185, 1960, pp 422-427.
4. Rupley, J. A. and Bruice, T.C. *Conversion of skeletal models to space-filling*. J. Mol. Biol. 37, 1968, pp 521-523.
5. Clarke, F. H. *New skeletal-space-filling models*. J. Chem. Education 54, 1977, pp 230-235.
6. Barrett, E. J. *Biopolymer models of nucleic acids*. J. Chem. Educ. 56, 1979, pp 168-169.
7. Wells, A. F. (1970). *Models in structural inorganic chemistry*. Oxford University Press, Oxford.
8. Walton, A. (1978). *Molecular crystal structure models*. Ellis Horwood Limited, Chichester, England.
9. Bassow, H. (1968). *Construction and Use of Atomic and Molecular Models*. Pergamon Press, Oxford.
10. Petersen, Q. R. *Some reflections on the use and abuse of molecular models*. J. Chem. Education 47, 1970, pp 24-29.
11. Gordon, A. J. *A survey of atomic and molecular models*. J. Chem. Education 47, 1970, pp 30-32.
12. Rubin, B. *Macromolecule backbone models*. Methods in Enzymology 115, 1985, pp 391-397.
13. Fletterick, R. J. and Matela, R. *Color-coded alpha-carbon models of proteins*. Biopolymers 21, 1982, pp 999-1003.
14. Fletterick, R. J., Schroer, T. and Matela, R.J. (1985). *Molecular Structure*. Blackwell Scientific Publications, Oxford.
15. Kendrew, J. C., Dickerson, R.E., Strandberg, B.E., Hart, R.G., Davies, D.R., Phillips, D.C. and Shore, V.C. *Structure of myoglobin: A three-dimensional Fourier synthesis at 2 Å resolution*. Nature 185, 1960, pp 422-427.
16. Richards, F. M. *The matching of physical models to three-dimensional electron-density maps: A simple optical device*. J. Mol. Biol. 37, 1968, pp 225-230.
17. Langridge, R. and MacEwan, A.W. In *IBM Scientific Computing Symposium on Computer Aided Experimentation*, 1965.
18. Levinthal, C. *Molecular model building by computer*, Scientific American 214, 1966, pp 42-52.
19. Barry, C.D., Levinthal, C., Ward, S.A. and Zwick, M. In *Computer graphics in macromolecular chemistry*. Emerging Concepts in Computer Graphics. Serest, D. and Nievergelt, J. (eds.), Benjamin, New York, 1968, pp 251-283.
20. Meyer, E.F., Jr. *Interactive computer display for the three dimensional study of macromolecular structures*, Nature, 232, 1971, pp 255-257.
21. Barry, C.D. and North, A.C.T. *The use of a computer-controlled display system in the study of molecular conformations*. Cold S. Harb. 36, 1971, pp 577-584.
22. Katz, L. and Levinthal, G. *Interactive computer graphics and representation of complex biological structures*. Ann. Rev. Biophys. Bioeng. 1, 1972, pp 465-504.
23. Levinthal, C., Wodak, S.J. Kahn, P. and Dadvanian, A.K. *Hemoglobin interaction in sickle cell fibers I: Theoretical approaches to the molecular contacts*. Proc. Nat. Acad. Sci., U.S.A. 72, 1975, pp 1330-1334.
24. Langridge, R. *Interactive three-dimensional computer graphics in molecular biology*. Fed. Proc. 33, 1974, pp 2332-2335.

25. Collins, D. M., Cotton, F.A., Hazen, E.E. Jr., Meyer, E.F. Jr. and Morimoto, C.N. *Protein crystal structures: Quicker, cheaper approaches*. Science. 190, 1975, pp 1047-1052.
26. Porter, T.K. *Spherical shading*. Comp. Graphics. 12, 1978, pp 282-285.
27. Feldmann, R.J., Bing, D.H., Furie, B.C. and Furie, B. *Interactive computer surface graphics approach to study of the active site of bovine trypsin*. Proc. Natl. Acad. Sci. USA, 75, 1978, pp 5409-5412.
28. Max, N.L. *ATOMLLL, ATOMS with shading and highlights*. Computer Graphics. 13, 1979, pp 165-173.
29. Palmer, T.C., Hausheer, F.H. and Saxe, J.D. *Applications of Ray Tracing in Molecular Graphics*. J. Mol. Graphics. 7, 1989, pp 160-164.
30. Evans, S.V. *SETOR: hardware lighted three-dimensional solid model representation of macromolecules*. J. Mol. Graphics, 11, 1993, pp 134-138.
31. Sayle, R. and Bissell, A. (1992). *RasMol: A Program for Fast Realistic Rendering of Molecular Structures With Shadows*. Proceedings of the 10th Eurographics UK '92 Conference, University of Edinburgh, Scotland.
32. Pique, M.E., Macke, T.J. and Arvai, A.S. *Flex: A light-weight molecular display program*. J. Mol. Graphics. 9, 1991, pp 40-41.
33. Surles, M.C., Richardson, J.S., Richardson, D.C., Brooks, F.P. Jr. *Sculpting proteins interactively: Continual energy minimization embedded in a graphical modeling system*. Protein Sci. 3, 1994, pp 198-210.
34. Bell, G., Parisi, A. and Pesce, M. (1995). *The Virtual Reality Modeling Language (VRML) 1.0 Specification*. Silicon Graphics, Mountain View, CA.
35. El llenguatge Java fou creat per Sun Technologies. Més informació, així com descàrrega de programari, a la seva adreça web: <http://java.sun.com>.
36. Més informació a la adreça web de Silicon Graphics: <http://www.sgi.com>.
37. La pàgina oficial de les llibreries MESA es troba a <http://www.mesa3d.org>
38. Mulliken, R.S. *Electronic population analysis on LCAO molecular wavefunctions. I*. J. Chem. Phys. 23, 1955, pp 1833-1840.
39. Mulliken, R.S. *Electronic population analysis on LCAO molecular wavefunctions. II. Overlap populations, bond orders and covalent bond energies*. J. Chem. Phys. 23, 1955, pp 1841-1846.
40. Mulliken, R.S. *Electronic population analysis on LCAO molecular wavefunctions. III. Effects of hybridization on overlap and gross AO populations*. J. Chem. Phys. 23, 1955, pp 2338-2342.
41. Mulliken, R.S. *Electronic population analysis on LCAO molecular wavefunctions. IV. Bonding and antibonding in LCAO and valence-bond theories*. J. Chem. Phys. 23, 1955, pp 2343-2346.
42. Mulliken, R.S. *Criteria for the construction of good self-consistent-field molecular orbital wavefunctions, and the significance of LCAO-MO population analysis*. J. Chem. Phys. 32, 1962, pp 3428-3439.
43. Mezey, P.G. (1991). *Shape in Chemistry: an introduction to molecular shape and topology*. VCH Publishers, New York.
44. Lee, B. and Richards, F.M. *The interpretation of protein structures: estimation of static accessibility*. J. Mol. Biol., V55, 1971, pp 379-400.
45. Agishtein, M.E. *Fuzzy molecular surfaces*. J. Biomolec. Struct. Dynamics. 9(4), 1992, pp 759-768.
46. Foresman, J.B. In *What every chemist should know about computing*. Ab initio techniques in chemistry: interpretation and visualization. Swift, M.L. and Zillinski, T.J. (eds.). ACS Books, Washington DC, 1996, Chpt 14.
47. Szabo, A. and Ostlund, N.S. *Modern quantum chemistry*. McGraw-Hill, New York, 1986.
48. Hehre, W.J., Radom, L., Schleyer, P.v.P. and Pople, J.A. *Ab initio molecular orbital theory*. Wiley, New York, 1986.
49. Frisch, M.J., Frisch, A., Foresman, J.B. *Gaussian 94 user's reference*. Gaussian Inc., Pittsburgh, 1995.

50. Frisch, M.J. and Frisch, A.E. *Exploring chemistry with electronic structure*. Gaussian Inc., Pittsburgh, 1995-6.
51. Steward, J.P. *Optimization of Parameters for Semiempirical Methods I. Method*. J. Comp. Chem. 10, 1989, pp 209-220.
52. Steward, J.P. *Optimization of Parameters for Semiempirical Methods II. Applications*. J. Comp. Chem. 10, 1989, pp 221-264.
53. Atkins, P.W. and Friedman, R.S. *Molecular quantum mechanics*, 3rd edition. Oxford, New York, 1997.
54. Zerner, M.C. *Semiempirical Molecular Orbital Methods*. Rev. Comp. Chem. Vol. 2, 1991, pp 313-365.
55. Veure per exemple:
 - a. http://www.uni-muenster.de/Chemie/OC/research/grimme/tc_links.html
 - b. <http://zeus.polsl.gliwice.pl/~nikodem/linux4chemistry.html>
 - c. <http://www.inform.umd.edu/UMS+State/UMD-Projects/MCTP/Technology/Chemistry.html#Chemistry Software>
 - d. http://www.chemistry-software.com/software_guide/modelling_index.htm
56. Constants, P. and Carbó, R. *Atomic Shell Approximation. Electron Density Fitting Program Algorithm Restricting Coefficients to Positive Values*. J. Chem. Inf. Comput. Sci., 35, 1995, pp 1046-1053.
57. Frisch, M.J., Trucks, G. W., Schlegel, H.B., Scuseria, G.E., Robb, M.A., Cheeseman, J.R., Zakrzewski, V.G., Montgomery, Jr., J.A., Stratmann, R.E., Burant, J.C., Dapprich, S., Millam, J.M., Daniels, D., Kudin, K.N., Strain, M.C., Farkas, O., Tomasi, J., Barone, V., Cossi, M., Cammi, R., Mennucci, B., Pomelli, C., Adamo, C., Clifford, S., Ochterski, J., Petersson, G.A., Ayala, P.Y., Cui, Q., Morokuma, K., Malick, D.K., Rabuck, A.D., Raghavachari, K., Foresman, J.B., Cioslowski, J., Ortiz, J.V., Stefanov, B.B., Liu, G., Liashenko, A., Piskorz, P., Komaromi, I., Gomperts, R., Martin, R.L., Fox, D.J., Keith, M. T., Al-Laham, A., Peng, C.Y., Nanayakkara, A., Gonzalez, C., Challacombe, M., Gill, P.M.W., Johnson, B., Chen, W., Wong, M.W., Andres, J.L., Gonzalez, C., Head-Gordon, M., Replogle, E.S., and Pople, J.A. Gaussian 98, Revision A.6, Gaussian, Inc., Pittsburgh PA, 1998.
58. Amat, L. and Carbó-Dorca, R. *Quantum Similarity Measures under Atomic Shell Approximation: First Order Density Fitting Using Elementary Jacobi Rotations*. J. Comput. Chem., 18, 1997, pp 2023-2039.
59. Amat, L., and Carbó-Dorca, R. *Fitted Electronic Density Functions from H to Rn for Use in Quantum Similarity Measures: cis-Diammine-Dichloroplatinium(II) Complex as an Application Example*. J. Comput. Chem., 20, 1999, pp 911-920.
60. Amat, L., and Carbó-Dorca, R. *Molecular Electronic Density Fitting Using Elementary Jacobi Rotations under Atomic Shell Approximation*. J. Chem. Inf. Comput. Sci., 40, 2000, pp 1188-1198.
61. Coeficients i exponents ASA per a diferents conjunts de funcions de base i sobre diferents grups atòmics es poden descarregar de: <http://iqc.udg.es/cat/similarity/ASA>.
62. Carbó, R., Leyda, L. and Arnau, M. *How similar is a molecule to another? An electron density measure of the similarity between two molecular structures*. Int. J. Quant. Chem., 17, 1980, pp 1185-1189.
63. The Protein Data Bank (PDB) es troba a la següent adreça Web: <http://www.rcsb.org/pdb>. Referència principal: Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. *The Protein Data Bank*. Nucleic Acids Research, 28, 2000, pp 235-242.
64. Allen, F.H., and Kennard, O. *3D research using the Cambridge Structural Database*. Chemical Design Automation News, 8, 1993, pp 31-37.

65. Bruno, I.J., Cole, J.C., Lommerse, J.P.M., Rowland, R.S., Taylor, R., and Verdonk, M. *IsoStar: a library of information about nonbonded interactions*. J. Chem. Inf. Comput. Sci., 11, 1997, pp 525-537.
66. Amat, L. and Carbó-Dorca, R. Programa Buildfdd, Institut de Química Computacional, Universitat de Girona, Girona, 1999.
67. GaussView, Gaussian, Inc., Pittsburgh, PA, 2000.
68. OrbDraw, Serena Software, Bloomington, IN, 2000.
69. Laaksonen, L. gOpenMol, Centre of Scientific Computing, Espoo, Finland, 1998..
70. AVS, rev H, 1994, Advanced Visual Systems Inc., 300 Fifth Avenue, Waltham, MA 02154, USA.
71. Althaus, E., Mehlhorn, K. Näher, A., and Schirra, S. *Experiments in Curve Reconstruction*. ALENEX, 2000, pp 103-114.
72. Althaus, E., and Mehlhorn K. *TSP-Based Curve Reconstruction in Polynomial Time*. Symposium on Discrete Algorithms (SODA), 2000, pp 686-695.
73. Amenta, N., and Choi, S. *One-Pass Delaunay filtering for homeomorphic 3D surface reconstruction*. UT Technical report number TR99-08, 1999.
74. Dey, T.K., Mehlhorn, K., and Ramos, E. *Curve reconstruction: connecting dots with good reason*. Proc. 15th. Sympos. Computational Geometry, 1999, pp 197-206.
75. Dey, T.K., and Kumar, P. *A simple provable algorithm for curve reconstruction*. Proc. 10th. ACM-SIAM Symposium on Discrete Algorithms (SODA '99), 1999, pp 893-894.
76. Cheng, S.W., and Dey, T.K. *Improved construction of Delaunay based contour surfaces*. Proc. ACM Sympos. Solid Modeling and Applications, 99, 1999, pp 322-323.
77. Amenta, N., Bern, M., and Kamvysselis, M. *A new Voronoi-based surface reconstruction algorithm*. Siggraph '98, 1998, pp 415-421.
78. Amenta, N., Bern, M., and Eppstein, D. *The crust and the beta-skeleton: combinatorial curve reconstruction*. Graphical Models and Image Processing, 60/2:2, 1998, pp 125-135.
79. Lorensen, W.E., and Cline H.E. *Marching Cubes: a high resolution 3D surface construction algorithm*. Computer Graphics (Proceedings of SigGraph'87), 21, 1987, pp 163-169.
80. Walker, P.D. and Mezey, P.G. *Molecular electron density lego approach to molecule building*. J. Am. Chem. Soc., 115, 1993, pp 12423-12430.
81. Walker, P.D., and Mezey, P.G. *Ab initio quality electron densities for proteins: A MEDLA approach*. J. Am. Chem. Soc., 116, 1994, pp 12022-12032.
82. Més informació a: <http://www.sbu.ac.uk/water/h10o5orb.html>
83. Bader, R.F.W., Henneker, W.H., and Cade, P.E.J. *Molecular charge distributions and chemical binding*. J. Chem. Phys., 46, 1967, pp 3341-3363.
84. Pisabarro, M.T., Ortiz, A.R., Serrano, L. and Wade, R.C. *Homology modeling of the Abl-SH3 domain*. Proteins, 20, 1994, pp 203-215.
85. Kostrewa, D. and Winkler, F.K. *Mg²⁺ binding to the active site of EcoRV endonuclease: a crystallographic study of complexes with substrate and product DNA at 2 Å resolution*. Biochemistry, 34, 1995, pp. 683-696.

3.9. Llistat d'Articles Adjunts

En aquesta secció s'inclouen a mode d'annex les contribucions llistades a continuació, les quals han estat acceptades per a publicació en diferents revistes especialitzades.

- Gironés, X., Amat, L. i Carbó-Dorca, R. *A comparative study of isodensity surfaces using ab initio and ASA density functions*. J. Mol. Graph. & Mod., 16, 1998, pp 190-196.
- Gironés, X., Carbó-Dorca, R. and Mezey, P.G. *Application of promolecular ASA densities to graphical representation of density functions of macromolecular systems*. J. Mol. Graph. & Mod., 3-4, 2001, pp 343-348.
- Gironés, X., Amat, L. and Carbó-Dorca, R. *Modelling Large Macromolecular Structures using Promolecular Densities*. J. Chem. Inf. Comput. Sci. Acceptat.

A comparative study of isodensity surfaces using *ab initio* and ASA density functions

Xavier Gironés, Lluís Amat, and Ramon Carbó-Dorca

Institute of Computational Chemistry, University of Girona, Girona, Catalonia, Spain

In this article, we report a visual comparison between several of the available methods for constructing electronic density functions. The density forms studied include *ab initio*, atomic shell approximation, and promolecular densities. A graphical comparison is made for six different molecules at different levels of density function values. The differences between the various density functions are analysed by considering a molecular quantum self-similarity measure and the required computational time for all molecules at all computation levels is considered. © 1999 by Elsevier Science Inc.

Keywords: *ab initio* densities, atomic shell approximation densities, marching cubes algorithm, promolecular densities, molecular quantum similarity measures.

INTRODUCTION

A significant amount of the research in our laboratory is related to molecular quantum similarity measures (MQSM),^{1–10} which can be obtained using the integral shown in Eq. (1):

$$Z_{AB} = \iint \rho_A(\mathbf{r}_1) \Omega(\mathbf{r}_1, \mathbf{r}_2) \rho_B(\mathbf{r}_2) d\mathbf{r}_1 d\mathbf{r}_2 \quad (1)$$

Where ρ_A and ρ_B are the corresponding density functions for two arbitrary molecules A and B and Ω is a positive definite operator. In this way, an MQSM, Z_{AB} , is constructed from a scalar product of the pair of density functions weighted by Ω . As the measure Z_{AB} depends on the relative positioning of both molecules in space, the optimal value of Eq. (1) will be derived from the optimal alignment of the molecules with each other.¹¹ This optimization procedure, which requires multiple evalua-

tions of the MQSM, and, consequently, the computation of many center integrals, can be very time consuming, depending on the method used to compute the density functions. In addition, it is important to note that the quality of the MQSM depends on the quality of the density functions used in the calculation.

DENSITY FUNCTIONS FORMS

Ab initio calculations

Ab initio calculations using the LCAO-MO approximation are widely used to evaluate density functions.¹² These density functions are formulated as shown in Eq. (2):

$$\rho_A(\mathbf{r}) = \sum_{\mu, \nu \in A} D_{\mu\nu} \chi_\mu^*(\mathbf{r}) \chi_\nu(\mathbf{r}) \quad (2)$$

where $D_{\mu\nu}$ is the charge and bond order matrix and $\{\chi_\mu\}$ is the atomic orbitals basis set.

The MQSMs of systems calculated by this procedure are as accurate as the level of theory used. However, this methodology has some disadvantages. Computational requirements are large owing to the amount of calculations needed to construct the density function. Furthermore, when an MQSM must be evaluated, four center integrals must be calculated, because Eq. (1), using expression (2) for ρ_A and ρ_B , adopts the form shown in Eq. (3):

$$Z_{AB} = \sum_{\mu \in A} \sum_{\nu \in A} \sum_{\lambda \in B} \sum_{\sigma \in B} D_{\mu\nu} D_{\lambda\sigma} \iint \chi_\mu^*(\mathbf{r}_1) \chi_\nu(\mathbf{r}_1) \Omega(\mathbf{r}_1, \mathbf{r}_2) \chi_\lambda^*(\mathbf{r}_2) \chi_\sigma(\mathbf{r}_2) d\mathbf{r}_1 d\mathbf{r}_2 \quad (3)$$

As a result, Eq. (3) is computationally time consuming. Molecular systems composed of a large number of atoms are difficult to treat at this level owing to constraints of computer resources. As an MQSM needs to be optimized, the amount of time required to obtain a maximal value for an MQSM can be large, as Z_{AB} integrals must be repeatedly calculated.

Address reprint requests to: Ramon Carbó-Dorca, Institute of Computational Chemistry, University of Girona, Girona, Catalonia, 17071 Spain. Fax: 34-972-418356. E-mail: director@iqc.udg.es.

Received 24 October 1998; revised 21 January 1999; accepted 13 April 1999.

Molecular atomic shell approximation

The molecular atomic shell approximation (ASA) approach^{13,14} expresses molecular electron densities as linear combinations of 1S Gaussian functions. The resulting density function, constructed in this way, is as shown in Eq. (4):

$$\rho_A^{\text{ASA}}(\mathbf{r}) = \sum_{i \in A} w_i |s_i(\mathbf{r})|^2 \quad (4)$$

where $s_i(\mathbf{r})$ are normalized 1S spherical functions, that is: $\int |s_i(\mathbf{r})|^2 d\mathbf{r} = 1$, and the coefficients w_i must fulfill the convex conditions shown in Eqs. (5) and (6):

$$w_i \in \mathbf{R} \wedge w_i > 0, \quad \forall i \in A \quad (5)$$

in order to guarantee that the density function is positive definite, and

$$\sum_{i \in A} w_i = 1 \quad (6)$$

to ensure the normalization of ASA density to unity. In this way the statistical probability distribution of expressions such as Eq. (4) is preserved.

When density functions are obtained for molecular systems in this way, then the calculation of MQSM in expression (3) transforms to the simplified form shown in Eq. (7):

$$Z_{AB} = \sum_{i \in A} \sum_{j \in B} w_i w_j \int \int |s_i(\mathbf{r}_1)|^2 \Omega(\mathbf{r}_1, \mathbf{r}_2) |s_j(\mathbf{r}_2)|^2 d\mathbf{r}_1 d\mathbf{r}_2 \quad (7)$$

However, two disadvantages arise from this procedure. The first is that the ASA density function needs to be fitted, which requires the *ab initio* calculation to be computed beforehand, providing density functions as in Eq. (2). The other problem is that for each molecule, there are coefficients w_i that must be fitted to the molecular density (2), and this involves extra computer requirements and time spent. Table 1 lists the times taken to fit some ASA functions.

Promolecular atomic shell approximation

The promolecular atomic shell approximation (PASA)¹⁴ is based on the same ideas as the ASA methodology, and involves the construction of the density function as a sum of atomic

density contributions, which are expressed in turn as linear combinations of 1S spherical functions as in Eq. (4). The resulting density function is written as in Eq. (8):

$$\rho_A^{\text{ASA}}(\mathbf{r}) = \frac{1}{P_A} \sum_{a \in A} P_a \rho_a^{\text{ASA}}(\mathbf{r}) \quad (8)$$

where P_a is the atomic number of each atom and P_A is the total number of electrons in the studied molecule.

In this way, molecular density is a simple sum of atomic densities, ρ_a^{ASA} centered at the atomic positions $\mathbf{R}_a$, which are defined as in Eq. (9):

$$\rho_a^{\text{ASA}}(\mathbf{r} - \mathbf{R}_a) = \sum_{i \in a} w_i |s_i(\mathbf{r} - \mathbf{R}_a)|^2 \quad (9)$$

with coefficients $\{w_i\}$ fulfilling convex conditions (5) and (6). In this way, PASA functions automatically include the conditions expressed in Eqs. (5) and (6).

The PASA needs an ASA density function as in Eq. (9). Therefore, for each atom, *ab initio* calculations and fitting processes are used to parameterize 1S function exponents and w_i coefficients, which are stored for later calculations.¹⁵

The main advantage of this method is that calculations take little time to be completed. There is no need to calculate the molecular density function at any level because atomic density calculations are to be carried out only once. Within the PASA the molecular density function is constructed simply by means of atomic coordinates and ASA densities, and such coordinates can be extracted from any source, including crystallographic data or previous theoretical calculations.

GRAPHICAL VISUAL COMPARISON BETWEEN DENSITIES

Computational details

In Tables 2–7, several comparisons are presented between six different molecules calculated at six different computation levels. The choice of molecular systems studied is similar to that used by Mezey,¹⁶ allowing comparison with the results of this earlier work. The molecules used are water, methanol, ethanol, allyl alcohol, benzene, and glycine. The calculation levels are (1) promolecular ASA at two different levels (simple and limit), (2) molecular ASA, and (3) HF/STO-3G, HF/3-21G, and HF/6-31++G**. In all cases, gas-phase geometries are used arising from a full optimization at the HF/6-31++G** level.¹⁷

A simple ASA calculation is a PASA calculation¹⁴ that involves one Gaussian function for each H, and three Gaussian functions for C, O and N. A limit ASA calculation is again a PASA calculation; however, in this case the number of functions per atom has increased to seven Gaussian functions for H, eight for C, and nine for O and N. It has been found that further increase in the number of functions does not improve the description of the atomic density. In molecular ASA the density function is fitted with as many Gaussian functions as needed, to the LCAO form, as presented in Eq. (3), calculated from a previous computation performed using an *ab initio* procedure.¹⁷ In this case this was chosen to be the HF/3-21G computation level.

Table 1. Time needed (in seconds) to fit ASA to *ab initio* functions obtained from a Gaussian HF/3-21G calculation^a

Molecule	Time ^b
Water	13
Allyl alcohol	1 478
Methanol	300
Ethanol	1 675
Benzene	4 815
Glycine	1 750

^a Frisch *et al.*¹⁷

^b Times calculated with a PC Intel Pentium 200 MHz, running under Windows 95. They should be taken as approximate, because Windows 95 runs self-programs in the background that can slow the calculation down.

Table 2. Comparison of water molecule density function shapes

Density Value	ASA			ab initio		
	Simple	Limit	Molecular	STO-3G	3-21G	6-31++G**
0.01						
0.15						
0.22						
0.35						
0.45						

Image construction

In all approximations, the steps involved in the graphical representation have been chosen in the same way. First, a grid, large enough to envelop the molecule, is defined in

three-dimensional space. At each grid point, a density function value is computed and stored in a file. When all grid points have been calculated, the wireframe model is created, choosing first the isodensity value, which is then computed according to a marching cubes algorithm (see Reference 18

Table 3. Comparison of methanol molecule density function shapes

Density Value	ASA			ab initio		
	Simple	Limit	Molecular	STO-3G	3-21G	6-31++G**
0.01						
0.15						
0.22						
0.35						
0.45						

Table 4. Comparison of ethanol molecule density function shapes

Density Value	ASA			ab initio		
	Simple	Limit	Molecular	STO-3G	3-21G	6-31++G**
0.01						
0.15						
0.22						
0.35						
0.45						

and Appendix 1). Finally, the input file is rendered by the program GiD.¹⁹

RESULTS AND DISCUSSION

As can be seen in Tables 2–7, the lowest and highest density function values correctly describe the surfaces in all cases.

For the middle density function values, bond formation and hydrogen position are not so accurately described with the ASA approaches, as the density function has collapsed over the atoms because only 1S Gaussian functions are used, and the hydrogen density is not well described. However, when working with quantum similarity measures, the major part of

Table 5. Comparison of allyl alcohol molecule density function shapes

Density Value	ASA			ab initio		
	Simple	Limit	Molecular	STO-3G	3-21G	6-31++G**
0.01						
0.15						
0.22						
0.35						
0.45						

Table 6. Comparison of benzene molecule density function shapes

Density Value	ASA			ab initio		
	Simple	Limit	Molecular	STO-3G	3-21G	6-31++G**
0.01						
0.15						
0.22						
0.35						
0.45						

the density function surrounds the heavy atoms. These inconveniences are not so important, and do not affect the final results in any significant way. This was demonstrated by performing a comparison employing overlap-like molec-

ular quantum self-similarity measures (MQS-SM), Z_{AA} . In Eq. (1) the operator $\Omega(r_1, r_2)$ is chosen to be equal to a Dirac's delta function, $\delta(r_1 - r_2)$, for the six molecules considered. The results are presented in Table 8, and the

Table 7. Comparison of glycine molecule density function shapes

Density Value	ASA			ab initio		
	Simple	Limit	Molecular	STO-3G	3-21G	6-31++G**
0.01						
0.15						
0.22						
0.35						
0.45						

Table 8. Comparison between overlap-like MQS-SM for the six considered molecules

Molecule	ASA			<i>Ab initio</i>		
	Simple	Limit	Molecular	STO-3G	3-21G	6-31 ⁺⁺ G**
Water	80.86	80.52	80.25	79.41	80.24	81.08
Methanol	112.56	112.25	107.89	107.02	108.03	107.74
Ethanol	144.32	144.00	141.05	139.66	141.18	142.24
Allyl Alcohol	175.94	175.60	172.47	170.10	171.61	176.28
Benzene	190.09	190.02	179.59	179.43	180.12	189.81
Glycine	277.16	276.53	273.33	266.53	272.53	274.30

Table 9. Comparison between overlap-like MQS-SM calculation time^a for the six considered molecules (in seconds)

Computational level		Water	Methanol	Ethanol	Allyl alcohol	Benzene	Glycine
ASA	Simple	≈0 ^b	≈0	≈0	≈0	≈0	≈0
	Limit	≈0	≈0	0.17	0.22	0.22	0.44
	Molecular	≈0	0.06	0.33	≈0	0.33	0.38
<i>Ab initio</i>	STO-3G	1	1	3	7	16	12
	3-21G	1	2	4	8	17	13
	6-31 ⁺⁺ G**	1	21	125	286	576	484

^a Times calculated with a PC Intel Pentium 200 MHz, running under Windows 95. Times should be taken as approximate, because Windows 95 runs self-programs in the background that can slow the calculation down.

^b A zero value means that the computer processor has not already updated the clock; a PC, such as the one used, takes 0.056 s to update the system clock.

related times needed to compute them are reported in Table 9.

From Tables 8 and 9, it can be seen that no really significant differences appear between the values of computed overlap-like MQS-SM. However, the times required to compute them differ greatly and become large as the computational level achieves more accuracy.

CONCLUSIONS

The use of ASA density functions in quantum mechanical computations has opened a new door to fast and precise results, compared with the results obtained using *ab initio* computations. We have shown here that the ASA density function describes molecular shapes in almost the same way as *ab initio* computations, and the MQS-SM example verifies the validity of the use of the ASA density function.

ACKNOWLEDGMENTS

This research has been partially supported by a CICYT grant (SAF 96-0158), by the Fundació Maria Francisca de Roviralta, and by a EU contract (ENV4-CT97-0508). The authors acknowledge Dr. M. Pastor (Laboratory of Chemometrics, Department of Chemistry, University of Perugia) for his kindness in providing us with the information about the marching cubes algorithm. X. Gironés wishes to acknowledge University of Girona for a predoctoral fellowship.

APPENDIX-MARCHING CUBES ALGORITHM

Marching Cubes Algorithm (MCA) is used to render isosurfaces in volumetric data.^{20,21} The basic notion in MCA consists in that one can define a voxel as the sequence of the pixel values at the eight corners of a cube. If one or more pixels of a cube have values less than the user-specified isovalue, and one or more have values greater than this value, one knows the voxel must contribute to some component of the isosurface. By determining which edges of the cube are intersected by the isosurface, one can create triangular patches that divide the cube between regions within the isosurface and regions outside. By connecting the patches from all cubes on the isosurface boundary, a surface representation is obtained.

Algorithm details

There are two major components of MCA. The first consists into deciding how to define the section or sections of surface which chop up an individual cube. If each corner is classified as either being below or above the isovalue, there are 256 possible configurations of corner classifications. However if one accounts for symmetry, only 15 configurations remain (see Reference 22a for a figure). Two of these are trivial: all points are inside or outside the cube and do not contribute to the isosurface. For all configurations it is needed to determine where, along each cube edge, the isosurface crosses, and these edge intersection points can be used to create one or more triangular patches for the isosurface.

When surface patches can be created for a single voxel, this process can be applied to the entire volume. One can process the volume in slabs, where each slab is comprised between 2 slices of pixels. Each cube can be treated either independently, or edge intersections between cubes which share the edges can be propagated.

More information about the MCA can be found on the WWW (see Reference 22).

REFERENCES

- 1 Carbó, R., Leyda L., and Arnau, M. How similar is a molecule to another? An electron density measure of similarity between two molecular structures. *Int. J. Quantum Chem.* 1980, **17**, 1185–1189
- 2 Besalú, E., Carbó, R., Mestres, J., and Solà, M. Foundations and recent advances on molecular quantum similarity. *Topics Curr. Chem.* 1995, **173**, 31–62
- 3 Carbó, R., Besalú, E., Calabuig, B., and Vera, L. Molecular quantum similarity: Theoretical framework, ordering principles, and visualization techniques. *Adv. Quantum Chem.* 1994, **25**, 253–313
- 4 Carbó R., Ed. *Molecular Similarity and Reactivity: From Quantum Chemical to Phenomenological Approaches*. Kluwer, Amsterdam, 1995
- 5 Carbó-Dorca, R., and Mezey, P.G., Eds. *Advances in Molecular Similarity*, Vol. 1. JAI Press, Greenwich, Connecticut, 1996
- 6 Cioslowski, J., and Fleischmann, E.D. Assessing molecular similarity from results of ab initio electronic structure calculations. *J. Am. Chem. Soc.* 1991, **113**, 64–67
- 7 Burt, C., Richards, W.G., and Huxley, P. The application of molecular similarity calculations. *J. Comput. Chem.* 1990, **11**, 1139–1146
- 8 Mezey, P.G. Density domain bonding topology and molecular similarity measures. *Topics Curr. Chem.* 1995, **173**, 63–82
- 9 Allan, N.L., and Cooper, D.L. Momentum-space electron densities and quantum molecular similarity. *Topics Curr. Chem.* 1995, **173**, 85–111
- 10 Ponc, R. Similarity models in the theory of pericyclic reactions. *Topics Curr. Chem.* 1995, **174**, 1–26
- 11 Constants, P., Amat, L., and Carbó-Dorca, R. Toward a global maximization of the molecular similarity function: Superposition of two molecules. *J. Comput. Chem.* 1997, **18**, 826–846
- 12 Parr, R.G. *The Quantum Theory of Molecular Electronic Structure*. W.A. Benjamin, New York, 1963
- 13 Constants, P., and Carbó, R. Atomic shell approximation. Electron density fitting program algorithm restricting coefficients to positive values. *J. Chem. Inf. Comput. Sci.* 1995, **35**, 1046–1053
- 14 Amat, L., and Carbó-Dorca, R. Quantum similarity measures under atomic shell approximation: First order density fitting using elementary Jacobi rotations. *J. Comput. Chem.* 1997, **18**, 2023–2039
- 15 ASA coefficients and exponents for an assorted sample of atoms can be seen and downloaded from the WWW site: <http://iqc.udg.es/cat/similarity/ASA/funcset.html>
- 16 Mezey, P.G. *Shape in Chemistry*. VCH Publishers, New York, 1993
- 17 Frisch, M.J., Trucks, G.W., Schlegel, H.B., Gill, P.M.W., Johnson, B.G., Robb, M.A., Cheeseman, J.R., Keith, T., Petersson, G.A., Montgomery, J.A., Raghavachari, K., Al-Laham, M.A., Zakrzewski, V.G., Ortiz, J.V., Foresman, J.B., Cioslowski, J., Stefanov, B.B., Nanayakkara, A., Challacombe, M., Peng, C.Y., Ayala, P.Y., Chen, W., Wong, M.W., Andres, J.L., Replogle, E.S., Gomperts, R., Martin, R.L., Fox, D.J., Binkley, J.S., Defrees, D.J., Baker, J., Stewart, J.P., Head-Gordon, M., Gonzalez, C., and Pople, J.A. *Gaussian-94*, revision E.2. Gaussian, Inc., Pittsburgh, Pennsylvania, 1995
- 18 A Fortran 90 procedure for the MCA can be seen and downloaded from the WWW site: <http://iqc.udg.es/cat/similarity/ASA/mca.html>
- 19 GiD, Geometry and Data, a pre/postprocessor graphical interface. An academic version can be downloaded at the WWW site: <http://gatxan.upc.es>. International Center for Numerical Methods in Engineering (CIMNE), 1998, Barcelona, e-mail: cimne@etseccpb.upc.es
- 20 Lorensen, W.E., and Cline, H.E. A high resolution 3D surface construction algorithm. *Computer Graphics* 1987, **21**, 163
- 21 Watt, A., and Watt, M. *Advanced Animation and Rendering Techniques*. Addison-Wesley, Reading, Massachusetts, 1992
- 22 WWW sites with more information about the marching cubes algorithm: (a) <http://exaflop.org/docs/marchcubes>; (b) http://panda.uchc.edu/htbit/indiv/software_docs/marching_cubes.html

Application of promolecular ASA densities to graphical representation of density functions of macromolecular systems

X. Gironés,* R. Carbó-Dorca,* and P.G. Mezey†

*Institute of Computational Chemistry, University of Girona, Campus Montilivi, 17071 Girona, Catalonia, Spain

†Department of Chemistry and Department of Mathematics and Statistics, University of Saskatchewan, 110 Science Place, Saskatoon, S7N 5C9 Canada

In this article we report the application of the Promolecular Atomic Shell Approximation (Promolecular ASA) to the graphical representation of the density function (DF) of large macromolecular systems. Promolecular ASA DF, constructed from previously computed and fitted atomic densities, provides a fast and practical representation of Molecular IsoDensity Contours (MIDCOs). These representations can be extended to macromolecular systems composed by >1000 atoms easily and with low computational costs, allowing the visualization of protein DF. The method is at first presented with a small molecule (2,4,6-trinitrophenol), comparing the resulting ASA MIDCOs with direct ab initio contours. For macromolecular tests the Promolecular ASA densities are also applied to the generation of macromolecular density surfaces of two proteins: myoglobin (2541 atoms) and gene V protein (1362 atoms). © 2001 by Elsevier Science Inc.

Keywords: ASA, promolecular densities, density function, MIDCOs, graphical representation, macromolecular systems

INTRODUCTION

The representation and visualization of molecular shape constitute a powerful tool in understanding a molecular system and its chemical behavior or biological response. As molecular

shape may be a useful concept to many scientists, several models have been introduced to describe molecular shape.

Concerning local shapes, the earliest quantum chemical approaches are those of the orbital representations of Mulliken.^{1–5} Quantum chemistry provides, in general, the most reliable shape representations,⁶ strongly connected to molecular similarity concepts based on quantum mechanics.^{7,8} Novel macromolecular Quantum chemistry approaches provide new linear time-dependent computational techniques and *ab initio* quality results for truly large systems, including proteins.^{9–13} The special aspects of local shape vs. global shape are reflected in The Holographic Electron-Density Theorem.¹⁴

Shape analysis on a simple level also has a rich history. A first crude approximation to molecular shape is given by the wire-frame model, where the molecule is simply represented using vectors corresponding to the bonds. This model can be improved using the ball-and-stick model, which constructs the molecular structure by means of small balls, representing the existing atoms in the molecule, linked by cylinders, accounting for the bonds (simple, double, triple) between the constituent atoms. In this way ball-and-stick models provide a discrete vision of molecular shape using atoms and bonds as primary building blocks. An alternative to this last model is ribbon representation of the molecular structure, which is mainly used in biochemistry to visualize proteins.¹⁵ These models, however, become inappropriate when the molecular surface is to be explored, since they are unable to create a valid representation of the effective occupied space in the molecular environment.

A common and simple procedure to obtain a molecular surface is the space-filling model, which assigns to each atom a sphere of its Van der Waals radius. This empirical method provides a better understanding of the molecular shape, as it accounts for some individual properties of the corresponding atoms. Within this procedure the molecular shape is repre-

Corresponding author: R. Carbó-Dorca, Institute of Computational Chemistry, University of Girona, Campus Montilivi, 17071 Girona, Catalonia, Spain. Tel.: +34 972418367; fax: +34 972418356.

E-mail address: director@iqc.udg.es

sented by a set of fused spheres. Other empirical approaches to molecular surface are the solvent accessible¹⁶ and Connolly surfaces.¹⁷ The solvent accessible surface corresponds to the molecular envelope of the surface of each atom that is accessible to the center of a probe sphere of a given radius. The Connolly surface is composed of two kinds of surface patches: the part of the Van der Waals surface of each atom that is accessible to a probe sphere of a given radius, generally 1.4 Å (sphere including most of a water molecule), and the inward-facing part of the probe sphere when it is simultaneously in contact with more than one atom (the reentrant surface).

However, these methods lack realism when representing the molecular shape as they truncate the existing interactions of the molecule. A practical way to introduce better molecular shape representations is the use of the system's electronic density function (DF). The DF describes the actual molecular electronic distribution and reflects its real, fuzzy characteristics, hence DF provides a suitable molecular shape representation.⁶ Among the several strategies to obtain DF, the most used are the *ab initio* procedures, which are implemented in available quantum chemistry software packages, such as Gaussian 98¹⁸ or Gamess,¹⁹ with an assorted collection of basis sets. Even though this methodology provides precise DF, insofar as the basis used is accurate enough, the procedure requires a large amount of computer time, which implies that large systems may not be accurately described within reasonable time limits. So far, direct *ab initio* calculations are presently restricted to molecules containing no more than 150 atoms, with the exception of periodical systems.

To overcome this problem, some procedures have been described to obtain accurate DF in computationally inexpensive ways. Successful linear, time-dependent *ab initio* quality approaches include the MEDLA^{20,21} and ADMA methods.¹³ Other techniques use data from experimental sources or from previous theoretical calculations to simplify the computation. Alternative semi-empirical methodologies, even though they produce valuable results, fail when they deal with molecular features that were not previously parameterized. A related procedure, developed in our laboratory, is the Promolecular Atomic Shell Approximation (Promolecular ASA),²² which will be explained below. The major use of Promolecular ASA DF has been devoted to Molecular Quantum Similarity Measures,²³ however, the successful application of Promolecular ASA densities to molecular shape representations was demonstrated in a previous work.²⁴ It was shown that Promolecular ASA DF were capable of generating isodensity contours comparable to those obtained by direct *ab initio* methods, while only using a fraction of the time and computational resources. This method opened a door to the possible use of Promolecular ASA densities to the simple generation of Molecular IsoDensity COntours (MIDCOs). The intention of this article, as a continuation of the previous study, is to prove the applicability of Promolecular densities as an easy, inexpensive method for representing ASA MIDCOs of large macromolecular systems.

The article is organized as follows. First, the main characteristics of the Promolecular ASA approach are detailed and illustrated with a graphical example. Next, MIDCOs for two proteins, myoglobin and gene V, obtained with the exposed methodology, are presented and commented upon. Finally, the conclusions of this work, as well as proposals for future research are given.

PROMOLECULAR ATOMIC SHELL APPROXIMATION (PROMOLECULAR ASA)

The theoretical basis of Promolecular densities²² overlays the ideas of the ASA approach.²⁵ ASA is a model in which the molecular electron DF is fitted with spherical 1S-type Gaussian functions to the original *ab initio* DF. This methodology allows the simplification of the mathematical expression of the DF, thus reducing the computational cost of the subsequent computations, with a reasonable error margin.

As the fitting procedure to obtain the ASA DF was computationally tedious and, in addition, a previous *ab initio* calculation was needed, this methodology was hard to extend to large molecular systems. To expand the applicability of ASA, another simplification was applied, resulting in the Promolecular ASA approach. The Promolecular approach considers the whole molecular DF as a linear combination of discrete atomic contributions, which in turn are expressed as linear combinations of spherical 1S-like Gaussian functions. The Promolecular ASA DF can be mathematically formulated as:

$$\rho_A^{PASA}(\mathbf{r}) = \sum_{a \in A} P_a \rho_a^{ASA}(\mathbf{r} - \mathbf{R}_a), \quad (1)$$

where P_a is the atomic number of the a -th atom. In this way, the molecular density is a simple addition of atomic densities, ρ_a^{ASA} , centered at the atomic positions $\mathbf{R}_a$. These atomic densities can be expressed as:

$$\rho_a^{ASA}(\mathbf{r} - \mathbf{R}_a) = \sum_{i \in a} w_i |S_i(\mathbf{r} - \mathbf{R}_a)|^2, \quad (2)$$

where S_i are the normalized spherical 1S Gaussian functions and w_i are the coefficients of the linear expansion, which must fulfill the following conditions:

$$w_i \in \mathbb{R} \wedge w_i > 0, \forall i \in a, \quad (3)$$

to guarantee that the DF is positive definite in its whole range, and

$$\sum_{i \in a} w_i = P_a, \quad (4)$$

to assure that the integral of ρ_a^{PASA} over all space is the number of electrons of molecule A , thus ensuring the proper definition of charge distribution function.

As seen in Equation 1, Promolecular ASA densities require atomic ASA DF. For this reason and for each atom, *ab initio* calculations and fitting computations have been carried out to parameterize the optimal exponents of the S_i functions and coefficients w_i . These results are stored in a database for later calculations.²⁶ So, the construction of Promolecular ASA DF is done simply by means of these previously computed ASA densities and atomic coordinates. Molecular coordinates can be extracted from many sources, such as crystallographic data or theoretical calculations.

A Graphical Example

To illustrate the validity of Promolecular densities when representing MIDCOs, a graphical example is provided. The chosen molecule is 2,4,6-trinitrophenol, which presents interesting

features such as aromaticity, due to the presence of the benzene ring and an internal hydrogen bond. The DF of the selected molecule has been constructed in two different ways. The first involves an *ab initio* method computed with Gaussian 98,¹⁸ in this case the calculation chosen was a restricted Hartree-Fock calculation using the 6-31G basis set. The second one uses the Promolecular ASA approach, employing previously computed functions present in the exponent and coefficient database.²⁶ These parameters were obtained from fitting atomic calculations using the 6-31G basis set to spherical 1S Gaussian functions. In both cases, the employed molecular geometry is the same and has been optimized using the AM1 Hamiltonian with the Ampac 6.01 software package.²⁷

From the computed DF, the corresponding MIDCOs that will be used to analyze the differences between the two methodologies are created. MIDCOs are created by evaluating the DF in a grid. The molecule is placed on a grid of points large enough to adequately envelop the molecule. At each point of the grid, the corresponding DF is evaluated and stored. Then a curvature analysis is performed over the previously computed points. The curvature analysis is used to determine features of the molecular shape, like convexity and concavity in the molecular surfaces, and to observe how these features change by modifying the isodensity level. The different representations have been graphically plotted with AVS software²⁸ and are presented in Table 1.

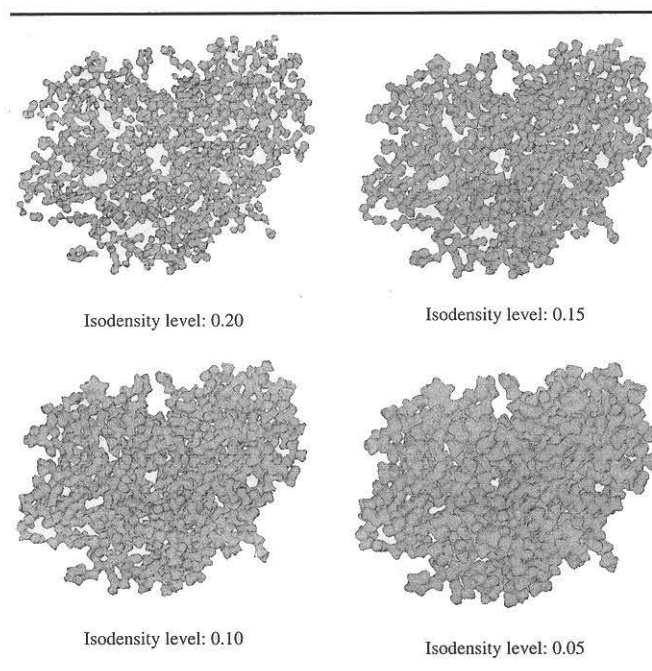
As seen in the graphical representations in Table 1, there are no significant differences between the Promolecular ASA MIDCOs and the respective *ab initio* ones. At first sight, high-density levels (>0.300) reflect the atomic positions in space, intermediate density levels (0.150–0.300) define the bonding pattern, and the low-density levels outline the volume and steric features of the molecular shape for both methods. However, a closer inspection of the figures reveals that the carbon–carbon bonds are manifested earlier in the *ab initio* representation at a density value of about 0.300, whereas there is no sign of them in the Promolecular approach at the same density level. In the Promolecular density, hydrogens appear at a lower density value, 0.209, whereas they are already defined in the *ab initio* density. These facts are due to the nature of the Promolecular ASA method. As DF is collapsed at the nuclear positions with 1S Gaussian functions, the bond pattern appears at a later stage, compared with the *ab initio* procedure, which includes higher-quantum number functions and consequently offers a better description. At the same time, the description of hydrogen atoms within the Promolecular framework is not accurate enough, when compared with the *ab initio* description.

However, these differences may not seem important if the reduced time needed to obtain the representations is taken into account. While the Promolecular density grid was calculated in seconds, the *ab initio* procedure lasted about 20 minutes using the same computer (a DEC Alpha 3000 with 64 MB RAM). In addition, the Promolecular ASA DF was constructed nearly instantly with the coefficients and exponents of the database,²⁶ whereas the *ab initio* calculation needed about 35 minutes to be completed. Furthermore, MIDCOs obtained for high-density levels are similar enough to perform qualitative comparisons. For low-density levels, figures from both methodologies become almost identical.

Besides the MIDCOs comparison, curvature analysis confirms the applicability of Promolecular densities to represent molecular shape. The regions of the surfaces close to the

Table 1. Graphical comparison and curvature analysis of 2,4,6-trinitrophenol, computed with Promolecular ASA and *ab initio* levels

Density Value	Promolecular ASA		<i>Ab initio</i>	
	Density function	Curvature analysis	Density function	Curvature analysis
0.500				
0.400				
0.300				
0.209				
0.170				
0.150				
0.100				
0.050				
0.045				
0.037				
0.020				

Table 2. Representation of myoglobin MIDCOs

atomic locations present a convex shape, whereas the intermediate vicinities evidence a locally saddle-type appearance (D_1 domains in the terminology of Mezey⁶). As the density level decreases, the convex regions tend to grow until a completely convex surface would be obtained for very low-density levels. The curvature tendencies for both methods are entirely equivalent, evidencing the validity of Promolecular ASA DF to provide realistic and inexpensive MIDCOs.

Construction of Macromolecular MIDCOs

In this section, we intend to show the applicability of Promolecular ASA DF as a powerful tool for creating MIDCOs of macromolecular systems. As examples, two proteins will be presented: myoglobin and gene V protein. The calculation involves the same steps as the previous example:

- obtaining the atomic coordinates
- constructing of the Promolecular ASA DF
- creating a grid large enough to completely surround the selected molecule
- calculating the DF value at each grid point
- representing the obtained MIDCOs with the AVS program.²⁸

The atomic coordinates have been extracted from the Protein Data Bank (PDB),²⁹ the PDB Id of the proteins are 1MBN for myoglobin and 1VQB for gene V protein. The original files contained several solvating water molecules, which were removed before the calculations were carried out. The construction of the Promolecular DF has been done using the parameters present in our database,²⁶ but in this case the fitting parameters used correspond to the 3-21G basis set, even though this basis set offers a poorer description. It was also demonstrated in a previous work²⁴ that acceptable ASA MIDCOs

could be derived from such fitting. In this way, the number of functions used is small, which translates to a reduction of computational time while still preserving a good visualization of the main, important features of the molecular shape, as will be seen in the examples.

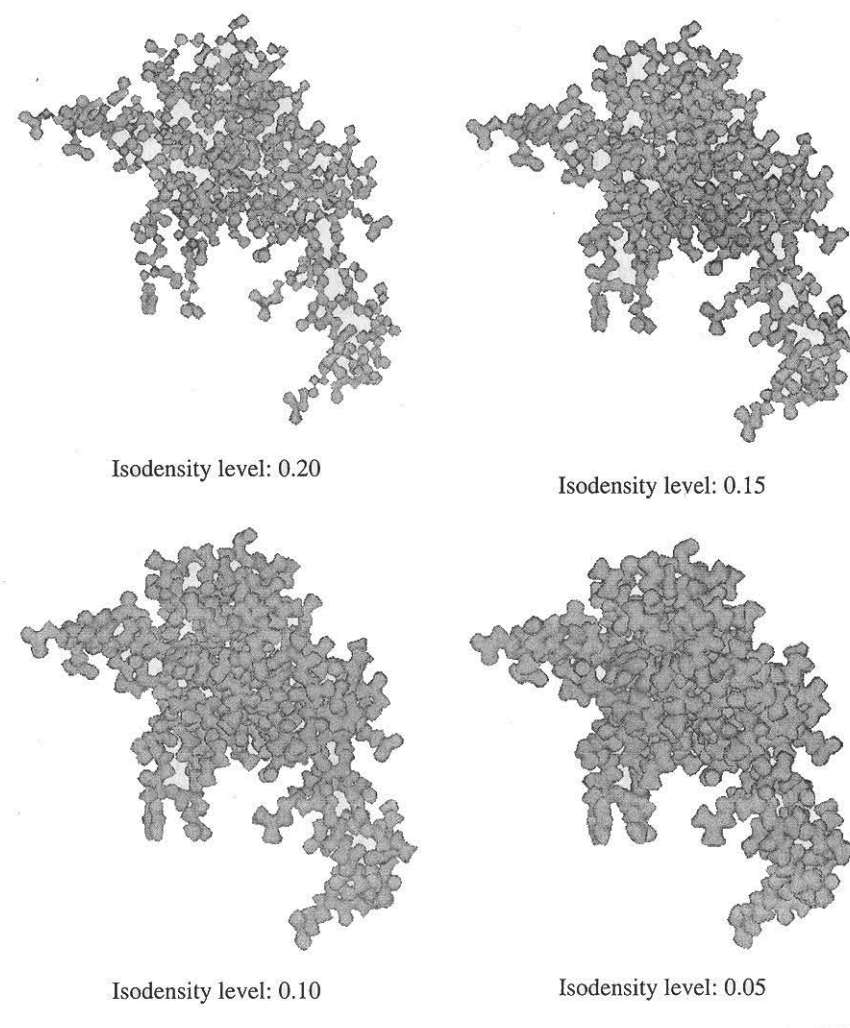
The first example of this is myoglobin, a protein that stores molecular oxygen. Kendrew and Watson,³⁰ first described this protein by extracting it from whale sperm, as myoglobin is present in larger quantities in marine mammals due to their extra oxygen requirements when doing deep dives under sea. The group responsible for this capacity is an iron atom encased in a heme group; this protein is also called ferric iron metmyoglobin. This molecule was chosen because it was the first protein whose atomic structure could be determined, and thus it laid the foundation for an era of biological understanding. This protein is formed by 2,541 atoms and 5,065 1S Gaussian functions are needed to construct de Promolecular ASA DF. The dimensions of the grid of points were $91 \times 83 \times 85$ atomic units (au) and the spacing between the points (resolution) was chosen to be 1 au. The MIDCOs obtained for myoglobin are presented in Table 2.

As seen from the MIDCOs presented in Table 2, the representation of the molecular shape shows the same tendencies as the previous 2,4,6-trinitrophenol. At high-density values the atomic positions are visualized, as are high-order bonds. Intermediate values provide the bonding pattern and low-density levels reflect the bulkiness of the molecular system. These graphical representations can provide valuable information about reaction centers or pathways due to the volume of information they provide, accounting for the realistic sterical features of MIDCOs. In addition, these molecular representations are easily obtained with low computational costs. In this case, the evaluation of the Promolecular DF at all grid points required about 9 hours to complete in the test machine, whereas a direct *ab initio* calculation and DF evaluation for this system would last years in the most modern computers.

The second example presented consists of the gene V, a DNA binding protein. The structure of this protein was determined by Terwilliger et al.,³¹ and was extracted from Bacteriophage F1. This molecular structure was chosen because its MIDCOs were previously studied by Mezey et al. within the MEDLA approach,^{20,21} hence these methodologies can be compared. Briefly, MEDLA can construct in linear time *ab initio* quality MIDCOs of large molecules from previously computed fuzzy molecular fragments, which when adequately translated and rotated reproduce the entire molecular DF. The molecular fragments are chosen to possess environment similar to that of the final molecule, assuming that the fuzzy DF is almost transferable. This procedure requires calculating many small molecular systems, regarded as building blocks, at *ab initio* levels, which requires low computational costs. In addition, as the fragments resulting from these calculations are stored in a database, they can be used in the construction of other MIDCOs. In contrast, in the ASA method, the gene V protein is formed by 1,362 atoms and 2,709 1S Gaussian functions are needed to fit the DF at the 3-21G basis set. In this case, the dimensions of the grid box were $95 \times 107 \times 83$ au and the resolution, as in the previous example, has been set to 1 au. The obtained ASA MIDCOs for gene V protein are presented in Table 3.

As seen in Table 3, the ASA MIDCOs obtained follow the same model as the previous examples. As the DF level decreases, atomic locations, bonding patterns, and volumetric information are described. When the obtained ASA MIDCOs are compared

Table 3. Representation of gene V MIDCOs



with the previously studied ones,²¹ no relevant differences arise at first. However, because the MEDLA approach^{20,21} is based on fuzzy density fragments, it is better able to reflect the interatomic interactions. Promolecular ASA, as found in discrete atomic densities, results in a poorer description of these features. However, as only these precomputed atomic contributions are required in Promolecular ASA, there is no need to recompute them and thus the fitted DF obtained can be used throughout regardless of the molecular environment.

CONCLUSIONS

We have demonstrated the practical application of Promolecular ASA DF to the graphical representation of MIDCOs. Using this simplified model to obtain DF has been proved to achieve valuable representations of the molecular shape and opens a new door for the potential use of fitted densities in the scientific areas where accurate descriptions of molecular surfaces are needed. Future applications of Promolecular ASA densities, besides Quantum Molecular Similarity calculations, are intended to be developed. These new capabilities will be focused on obtaining various molecular properties such as electrostatic

potentials and interaction energy surfaces. Considering the low computational requirements of the Promolecular Approach, these new features will also be easily applied to macromolecular systems.

ACKNOWLEDGMENTS

This research was completed while X. Gironés stayed at the University of Saskatchewan, and has been supported by a European Commission Project No. ENV4-CT97-0508. X.G. benefited from a predoctoral fellowship from the University of Girona. We also acknowledge financial support from the Fundació Maria Francisca de Roviralta and from the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

- 1 Mulliken, R.S. Electronic population analysis on LCAO molecular wave functions. I. *J. Chem. Phys.* 1955, **23**, 1833–1840
- 2 Mulliken, R.S. Electronic population analysis on LCAO molecular wave functions. II. Overlap populations, bond

- orders, and covalent bond energies. *J. Chem. Phys.* 1955, **23**, 1841–1846
- 3 Mulliken, R.S. Electronic population analysis on LCAO molecular wave functions. III. Effects of hybridization on overlap and gross AO populations. *J. Chem. Phys.* 1955, **23**, 2338–2342
- 4 Mulliken, R.S. Electronic population analysis on LCAO molecular wave functions. IV. Bonding and antibonding in LCAO and valence-bond theories. *J. Chem. Phys.* 1955, **23**, 2343–2346
- 5 Mulliken, R.S. Criteria for the construction of good self-consistent-field molecular orbital wave functions, and the significance of LCAO-MO population analysis. *J. Chem. Phys.* 1962, **36**, 3428–3439
- 6 Mezey, P.G. *Shape in chemistry: an introduction to molecular shape and topology*. VCH Publishers, New York, 1991
- 7 Carbo, R. (Ed.) *Molecular similarity and reactivity: from quantum chemical to phenomenological approaches*. Kluwer Acad. Publ., Dordrecht, The Netherlands, 1995
- 8 Johnson, J.A., and Maggiora, G.M. (Eds.) *Concepts and applications of molecular similarity*. Wiley, New York, 1990
- 9 Mezey, P.G. Quantum chemistry of macromolecular shape. *Int. Rev. Phys. Chem.* 1997, **16**, 361–388
- 10 Mezey, P.G. Shape analysis. In: *Encyclopedia of computational chemistry*. Schleyer, P.v.R., Allinger, N.L., Clark, T., Gasteiger, J., Kollman, P.A., Schaefer III, H.F., and Schreiner P.R., Eds., John Wiley & Sons, Chichester, UK, 1998, Vol. 4, 2582–2589
- 11 Mezey, P.G. Computational microscopy: pictures of proteins. *Pharmaceutical News* 1997, **4**, 29–34
- 12 Mezey, P.G. Functional groups in chemistry. *Adv. Quant. Chem.* 1996, **27**, 163–222
- 13 Mezey, P.G. Quantum similarity measures and Löwdin's transform for approximate density matrices and macromolecular forces. *Int. J. Quantum Chem.* 1997, **63**, 39–48
- 14 Mezey, P.G. The holographic density theorem and quantum similarity measures. *Mol. Phys.* 1999, **96**, 169–178
- 15 Richardson, J.S. The anatomy and taxonomy of protein structure. *Adv. Protein Chem.* 1981, **34**, 167–339
- 16 Lee, B., and Richards, F.M. The interpretation of protein structures: Estimation of static accessibility. *J. Mol. Biol.* 1971, **V55**, 379–400
- 17 Agishtein, M. E. Fuzzy molecular surfaces. *J. Biomolec. Struct. Dynamics* 1992, **9(4)**, 759–768
- 18 Frisch, M.J., Trucks, G.W., Schlegel, H.B., Scuseria, G.E., Robb, M.A., Cheeseman, J.R., Zakrzewski, V.G., Montgomery, Jr., J.A., Stratmann, R.E., Burant, J.C., Dapprich, S., Millam, J.M., Daniels, D., Kudin, K.N., Strain, M.C., Farkas, O., Tomasi, J., Barone, V., Cossi, M., Cammi, R., Mennucci, B., Pomelli, C., Adamo, C., Clifford, S., Ochterski, J., Petersson, G.A., Ayala, P.Y., Cui, Q., Morokuma, K., Malick, D.K., Rabuck, A.D., Raghavachari, K., Foresman, J.B., Cioslowski, J., Ortiz, J.V., Stefanov, B.B., Liu, G., Liashenko, A., Piskorz, P., Komaromi, I., Gomperts, R., Martin, R.L., Fox, D.J., Keith, M. T., Al-Laham, A., Peng, C.Y., Nanayakkara, A., Gonzalez, C., Challacombe, M., Gill, P.M.W., Johnson, B., Chen, W., Wong, M.W., Andres, J.L., Gonzalez, C., Head-Gordon, M., Replogle, E.S., and Pople, J.A. Gaussian 98, Revision A.6, Gaussian, Inc., Pittsburgh PA, 1998
- 19 Schmidt, M.W., Baldridge, K.K., Boatz, J.A., Elbert, S.T., Gordon, M.S., Jensen, J.H., Koseki, S., Matsunaga, N., Nguyen, K.A., Su, S., Windus, T.L., Dupuis, M., and Montgomery, J.A. General atomic and molecular electronic structure system. *J. Comput. Chem.* 1993, **14**, 1347–1363
- 20 Walker, P.D., and Mezey, P.G. Molecular electron density lego approach to molecule building. *J. Am. Chem. Soc.* 1993, **115**, 12423–12430
- 21 Walker, P.D., and Mezey, P.G. Ab initio quality electron densities for proteins: A MEDLA approach. *J. Am. Chem. Soc.* 1994, **116**, 12022–12032
- 22 Amat, L., and Carbó-Dorca, R. Quantum similarity measures under atomic shell approximation: First order density fitting program using elementary Jacobi rotations. *J. Comput. Chem.* 1997, **18**, 2023–2039
- 23 Carbó, R., Leyda, L., and Arnau, M. How similar is a molecule to another? An electron density measure of similarity between two molecular structures. *Int. J. Quantum Chem.* 1980, **17**, 1185–1189
- 24 Gironés, X., Amat, L., and Carbó-Dorca, R. A comparative study of isodensity surfaces using ab initio and ASA density functions. *J. Mol. Graph. Model.* 1998, **16**, 190–196
- 25 Constants, P., and Carbó, R. Atomic shell approximation. Electron density program algorithm restricting coefficients to positive values. *J. Chem. Inf. Comput. Sci.* 1995, **35**, 1046–1053
- 26 Promolecular ASA coefficients and exponents for an assorted set of atoms and for different basis sets can be consulted and downloaded in our web site, <http://iqc.udg.es/cat/similarity/ASA/funcset.html>
- 27 Ampac, version 6.0, 1994, Semichem, 7128 Summit, Shawnee, KS 66216DA
- 28 AVS, rev. H, 1994, Advanced Visual Systems Inc., 300 Fifth Avenue, Waltham, MA 02154, USA
- 29 The PDB can be consulted through its website: <http://www.rcsb.org/pdb/index.html>
- 30 Watson, H.C., The stereochemistry of the protein myoglobin. *Prog. Stereochem.* 1969, **4**, 299–303
- 31 Skinner, M. M., Zhang, H., Leschnitzer, D.H., Guan, Y., Bellamy, H., Sweet, R.M., Gray, C.W., Konings, R.N., Wang, A.H., and Terwilliger, T.C. Structure of the gene V protein of Bacteriophage F1 determined by multiwavelength X-ray diffraction on the selenomethionyl protein. *Proc. Nat. Acad. Sci. USA.* 1994, **91**, 2071–2075.

Modelling Large Macromolecular Structures using Promolecular Densities

Xavier Gironés, Lluís Amat and Ramon Carbó-Dorca^{*}

Institute of Computational Chemistry, University of Girona,
Campus Montilivi, 17071, Girona, Catalonia, Spain.

Reference: CI010348T

ABSTRACT

A procedure to easily construct fitted density functions is presented. This methodology, based on the promolecule approach, is able to handle large macromolecular systems, like proteins. The usual procedure dealing with fitted densities has been improved by adding some restrictions, which allow faster calculations. As a main application example, molecular isodensity contours (MIDCOs) are constructed for two proteins, one of them composed by more than 50000 atoms. MIDCOs, as a visual representation of the molecular density function, and thus, an important descriptor of the molecular charge distribution, constitute a powerful tool in the understanding of molecular systems. MIDCOs are presented for both proteins, allowing to explore their surfaces, as well as to analyse their shape. Also, as quantum mechanical calculation example, Molecular Quantum Self-Similarity Measures are calculated for several proteins.

KEYWORDS

Molecular Isodensity Contours, MIDCOs, density function, DF, promolecular density, Promolecular Atomic Shell Approximation, PASA, Molecular Quantum Similarity Measures, MQSM

1. INTRODUCTION

An important tool for the description of three-dimensional (3D) electronic density functions (DF) is the representation of molecular isodensity contours (MIDCO's), which characterise the molecular shape in a detailed and realistic way. First defined by Mezey [1-6], MIDCO's generation and analysis, which inherently include molecular surface areas and occupied volumes, provide valuable information that eases the understanding of a molecular system, as well as they assist when performing molecular recognition, drug-receptor interactions or bulk physical properties prediction studies. Due to their importance, several research has been carried out on MIDCO's and shape analysis [7-16], leading to the Holographic Electron-Density Theorem [17].

^{*} Corresponding author. Phone: +34 972418367. Fax: +34 972418356. E-mail: quantum@iqc.udg.es.

Given a formal nuclear configuration, and assuming a 3D molecular co-ordinate system, the electronic DF, $\rho(\mathbf{r})$, is a function of the 3D position variable $\mathbf{r}$. Within this co-ordinate system, $\rho(\mathbf{r})$ assigns a density value at each point $\mathbf{r}$ of the 3D space. This charge DF can be experimentally determined [18,19] or theoretically calculated using an appropriate quantum chemical method provided by, for example, available quantum chemistry software [20,21].

In order to faithfully visualise detailed shape features, theoretical models are the most affordable available methodology, due to experimental data from X-ray diffraction [22], complemented with aspherical atom libraries [23], is still not accurate enough to provide a reasonable description of the outer molecular region. Since molecular properties and reactivity, as well as biological responses, are strongly connected to the external electronic density, rather than to the core one, it seems natural that an accurate representation should correctly model it. Besides $\rho(\mathbf{r})$ -based MIDCO's, other usual methodologies have been described, such as fused sphere models, solvent accessible surfaces or motif patterns, however, they only yield to rough and incomplete representations of the molecular shape features [24-28]. MIDCO's, as founded in quantum mechanical ideas, reflect the actual fuzzy characteristics of the molecular shape given by $\rho(\mathbf{r})$; however, one of the major problems in *ab initio* calculations is the ability to quickly generate these $\rho(\mathbf{r})$ for large macromolecules, so easy for small molecules.

Ordinary *ab initio* techniques, with assorted procedures and a wide selection of basis sets [29], are able to provide accurate $\rho(\mathbf{r})$. However, when the studied systems consist of about 150 atoms or more, the computational requirements become unaffordable, with the exception of periodic systems. Many workarounds have been formulated within *ab initio* techniques to both increase the range of application and reduce computational costs. Between these methodologies, the most relevant are semiempirical methods [29], which avoid many intermediate calculations approximating with experimental or theoretical data, thus providing good quality $\rho(\mathbf{r})$ within appealing time limits for those systems where accurate parameterisation exists; and linear scaling DFT methods [29,30], which sacrifice part of the integral calculation and uses approximate algorithms in order to speed up the whole process. However,

when dealing with very large macromolecular systems these procedures are extremely time consuming, even in present day multiprocessor hardware.

Several procedures have been developed in order to circumvent the large computational requirements of treating these systems. The *Molecular Electron Density Approach* (MEDLA) [7-10], developed by Walker and Mezey, was the first method that allowed constructing inexpensive and accurate electron densities for large molecules. MEDLA, based on simple fragment additivity principle, consists of constructing fragment densities, stored in a databank, which can be used as building blocks for building DF for larger molecules by selecting and arranging the interpenetrating fuzzy charge clouds of previously computed density fragments. Other approaches by the same group were also reported [13,31].

Another possibility is the use of *promolecular densities* [32-35] as primary building block instead of genuine *ab initio* densities. Promolecular densities are simplified models designed to emulate an *ab initio* DF, considering that the whole molecular DF is made from discrete atomic, non-interacting, contributions. This procedure, along with the use of lower angular momentum functions, such as spherical 1S gaussian functions that fit the original atomic *ab initio* DF, allows to speed considerably up any related calculation, such as the necessary grid of $\rho(\mathbf{r})$ around the molecule, needed to represent MIDCOs. Within this scope, detailed MIDCOs have been obtained, indeed for large molecules [14-16], and it is general enough to permit the inclusion of any element of the periodic table. A related approach, developed to replace hard sphere models, is the *Gaussian Method for Molecular Shape* [36,37] by Pickup and co-workers, where gaussian functions are fitted to emulate molecular arrays of van der Waals hard spheres, but not electronic density functions.

Even if promolecular approaches save much computational time, they are still unable to treat large macromolecular structures within reasonable time limits. The intention of this work is to provide a new insight by adding shortcuts when evaluating $\rho(\mathbf{r})$, so the calculation becomes a linear process of the number of atoms present in the molecule, instead of a cubic one. In addition, these cut-offs can also be applied when performing some related quantum chemical computations, such as Molecular Quantum

Similarity Measures (MQSM) [38-44], thus extending the range of application of these methods.

The present work is structured as follows: first, the methodological section exposes the basis of promolecular densities, MIDCO generation and a brief explanation of MQSM. An application example, dealing with two proteins: *Proto-Oncogene Tyrosine Kinase* [45], a small protein composed by 872 atoms, and *Bovine Heart Cytochrome C Oxydase* [46], a larger enzyme composed by 56863, will be presented, in which MIDCOs and similarity measures are performed. Finally, the concluding remarks are given.

2. METHODOLOGY

Promolecular Atomic Shell Approximation. In order to avoid expensive theoretical calculations to obtain electronic DF, the Promolecular Atomic Shell Approximation (PASA) [47-49] can be used to simplify such procedure. This approach considers the molecular density as a sum of discrete atomic density contributions, which are taken as 1S Gaussian functions and fitted to previously computed atomic *ab initio* ones. Promolecular ASA densities are mathematically expressed as:

$$\rho_A^{ASA}(\mathbf{r}) = \sum_{a \in A} P_a \rho_a^{ASA}(\mathbf{r} - \mathbf{R}_a), \quad (1)$$

where P_a is the atomic number of the a -th atom. In this way, the molecular density is a simple addition of atomic densities, ρ_a^{ASA} , centered at the atomic positions $\mathbf{R}_a$. These atomic densities can be expressed in turn as:

$$\rho_a^{ASA}(\mathbf{r} - \mathbf{R}_a) = \sum_{i \in a} w_i |S_i(\mathbf{r} - \mathbf{R}_a)|^2, \quad (2)$$

where S_i are the normalized spherical 1S Gaussian functions and w_i are the coefficients of the linear expansion, which must fulfil the following convexity conditions:

$$w_i \in (0,1], \quad \forall i \in a \quad \text{and} \quad \sum_{i \in a} w_i = 1 \quad (3)$$

in order to guarantee that the DF is positive definite in its whole range, and to normalize the atomic DF in order to assure that the integral of ρ_A^{ASA} over all space yields the number of electrons of molecule A, thus ensuring the proper definition of the charge distribution function. Once the overall atomic densities are built, each molecular density is constructed by adding, as appropriate, these atomic density elementary pieces.

An improvement can be introduced considering the promolecular nature. In earlier works [14,15] a set of 1S gaussian functions was assigned to each atom forming a molecule. Even if it is of minor importance for small molecules, as the molecular system grows larger, an incredible amount of 1S functions may be needed. As illustrative example, a protein like gene V (PDB Id: 1VQB), composed by 1362 atoms, would require about 2700 functions [15]. However, a simple analysis of this protein reveals that only 5 different atom types are present, namely hydrogen, carbon, nitrogen, oxygen and sulphur. Taking into account that in the promolecule each atomic type is described by a unique set of 1S gaussian functions, there is no need to place in memory a repetitive number of copies of the same set for the same type of atom. In this way, gene V protein would have been described only by 4 sets of 1S gaussian functions (1 for H, 3 for C, N and O, and 4 for S), with a total number of 14 functions.

At this point, it needs to be stated that promolecular DF constructed in this way, even fitted, constitute themselves a DF, which may be used in different quantum chemical applications, such as MQSM [38-44], which are explained later in this section. Due to its simple formulation, PASA DF may not be suitable, or accurate enough, for all kind of calculations; however in those applicable, the computational performance increase is outstanding compared to the usage of *ab initio* densities.

Molecular Isodensity Contours (MIDCOs) using promolecular densities.

Present day DF representations of MIDCOs [2], based on the early orbital representations of Mulliken [50-54], are constructed defining a grid enveloping the studied molecule, whose DF has been previously constructed, and evaluating this DF in

each grid point. Needless to say that the grid must be large and fine enough to provide complete and smooth representations of the different isodensity levels.

Even if $\rho(\mathbf{r})$ is defined in the entire 3D space and only becomes zero at infinite distance from the nuclei, this function converges rapidly to zero already at short distances, becoming imperceptible, and thus computationally negligible, at about 5 atomic units (au.) from the nearest nucleus [2]. Assuming this fact, and considering the structure of promolecular densities, it seems reasonable that the contribution of an atomic DF does not need to be computed further than 5 au. from the respective atomic location.

So, the proposed simplified flowchart for the generation of PASA MIDCOs can be written as follows:

- Input of atomic coordinates.
- Retrieve how many different atomic types are present and allocate the necessary sets of PASA functions.
- Define the molecular grid according to molecular size and desired resolution. All grid points are set initially to zero.
- For each atom: evaluate the addressed functions in a $(10 \text{ au.})^3$ box surrounding the atom.
- Download the grid box to disk using a suitable file format to be imported by any graphics application.

As seen in step d), it has been considered that the contribution of an atom to a point located further than 5 au. is practically zero; however, all relevant interactions are contained within the defined box, such as the bonding pattern and outer electronic distribution. This approximation sets the basis of the linear dependence with the number of atoms of the proposed algorithm. This procedure has been coded in Fortran 90 [55] and used in the forthcoming examples.

Molecular Quantum Similarity Measures. Given a pair of molecular structures, a possible way to determine their resemblance, accounting by their native quantum character, can be done using Molecular Quantum Similarity Measures (MQSM) [38-44]. This methodology states the degree of similarity between two objects

by means of their electronic distributions. The simplest MQSM, also known as Overlap-like MQSM, is mathematically expressed as:

$$Z_{AB} = \int \rho_A(\mathbf{r}) \rho_B(\mathbf{r}) d\mathbf{r} \quad (4)$$

where $\rho_A(\mathbf{r})$ and $\rho_B(\mathbf{r})$ are the respective DF of molecules A and B, and Z_{AB} is the resulting value of the MQSM between A and B, which takes into account the shared amount of density between both structures. In the particular case when A is equal to B, a Molecular Quantum Self-Similarity Measure (MQS-SM) is defined, accounting for the concentration of the charge in the studied molecular system.

The major use of Promolecular ASA DF has been devoted to compute MQSM because its simplicity yields to high computational performance and considerable time reductions compared to MQSM calculation based on *ab initio* densities [14]. Since it was proved that these measures differed by less than 2% from *ab initio* densities [14,47], their use was clearly justified in this way.

The computational cost of computing MQSM scales quadratically with the number of functions used, which is roughly proportional to the number of atoms. Even if this quadratic behavior cannot be linearized, the calculation of MQSM can also benefit from a cut-off in order to improve the computational performance when calculating the similarity values. The proposed approximation consists of including only in the calculation the contribution of those functions that are close in space. For this purpose, the same distance threshold of 5 au. is also applied.

3. RESULTS AND DISCUSSION

As commented in the introduction section, MIDCO's and MQS-SM will be calculated for two macromolecules using PASA DF. The featuring proteins are *Proto-Oncogene Tyrosine Kinase* (PDB Id: 1ABL) [45] and *Bovine Heart Cytochrome C Oxidase* (PDB Id: 1OCO) [46]. 1ABL, which is found in humans, is involved in the transference of phosphate groups, and its structure was elucidated by theoretical methods. This protein is formed by 872 atoms, but only 5 different kinds (H, C, N, O,

S). On the other hand, 1OCO controls the last step of nutrient molecules oxidation by transferring their electrons to oxygen molecules, thus catalysing the reaction of oxygen to water and generating energy to later build ATP. 56863 atoms, but only 10 different types (H, C, N, O, Na, Mg, S, Fe, Cu, Zn), compose this enzyme, which contains metals in a prosthetic group located inside a cavity.

Molecular Modelling. The molecular geometries were obtained from the Protein Data Bank (PDB) [56]. 1ABL, as theoretically derived, was ready for the forthcoming calculations. 1OCO contained solvating waters, which were previously removed, as well as it lacked from hydrogens in its structure, and those were later added using Babel software [57]. PASA DF for both molecules were constructed using previously computed fitting parameters to atomic Hartree-Fock calculations using 6-311G basis set [49]. All these steps, as well as the forthcoming calculations, have been carried out on a Intel Pentium III 600MHz with 256Mb RAM running Linux Red Hat 6.2 operative system.

MIDCOs. As a graphical example, some isodensity contours plots are presented in Figures 1 and 2, for 1ABL and 1OCO respectively, at different density threshold values. 1ABL has been plotted at 0.5 au. resolution, whereas 1OCO at 1 au., due to its larger size, but still providing an acceptable representation. As seen from the plots, as the isodensity level decreases, the following shape characteristics [5] are outlined:

- At high values (0.250 e/au³ and over) only atomic locations and pre-bonding patterns are present. These plots are in the *localized density range*, which mostly exhibits those density domains collapsed around nuclei or delimiting a well-defined functional group, like OH or COOH.
- At intermediate values (0.150 and 0.050 e/au³), the prominent density domains are those defining the whole molecular bonding range, as well as outlining a skinny representation of the molecule, similar to other kinds of molecular surface representations, like solvent-accessible surfaces [25]. At these levels, the volumetric features of density function, as well as initial indications of the occupied space, are evinced.
- At low isodensity levels (0.005 e/au³ and below) effective occupied space and the volumetric characteristics are fully exhibited. At these levels, the unique density domain corresponds only to the full molecule, being only the

outer regions visible. A further decrease in the isovalue would yield to a quasi-spherical domain, whose edge would be already far from the nuclei.

Figures 1 and 2, about here

It is also interesting to consider, regardless of the large number of atoms treated, the time needed to complete the calculation, as well as the dimensions of the grid used. Table I lists this information for the previous proteins, as well as for two more, whose size is intermediate between the graphically represented.

Table I.- about here

From Table I, it can be observed a sound relationship between the number of atoms present and the necessary time to compute the grid. The listed times are not only affordable in present day computers, but also very small if the large number of atoms present in the structure are taken into account. As it was commented before, no other methodology has been still able to treat, or even construct, DF of such magnitude. Even if the PASA DF consist of functions fitted to *ab initio* ones, they constitute a suitable starting point for simple calculations involving these vast molecular systems, as it is demonstrated in the next section, where MQS-SM are computed for these structures.

MQS-SM. As a quantum chemical application using PASA DF calculations of MQS-SM will be presented. As previously commented in the methodological section, a distance threshold of 5 au. is also be applied, so two functions further than this threshold will not contribute to the measure. However, and for comparative purposes, unrestricted calculations are also performed, so that the time savings can be valued. The results are given in Table II.

Table II.- about here

Perusing Table II manifests a rough quadratic dependence between the number of atoms and the values of MQS-SM in both approaches. Even if unrestricted MQS-SM values are still affordable in present day hardware, the application of the distance restriction considerably reduces the calculation time, practically without any loss of

precision. As seen in Table II, differences between both kind of values are below 0.005% in all cases, pointing out the usefulness of the proposed approximation.

According to the definition of MQS-SM, if a quantum object is to be compared to itself, there is no need for structural superposition, however, if the more general case is to be handled, the value of the similarity measure strongly depends on the relative position of both analysed objects in space [47,58]. Hence, a superposition procedure, as well as repeated samples of the measures, needs to be computed. As quantum similarity measures require intensive calculation, it is interesting to have the possibility to compute them faster within an acceptable precision.

4. CONCLUSIONS

The practical use of Promolecular Densities derived from Atomic Shell Approximation applied to the construction of Density Functions of large macromolecular systems has been demonstrated. These fitted densities have been employed to construct Molecular Isodensity Contours of these large systems with both affordable computational and time requirements. As previously demonstrated, molecular surfaces derived from promolecular densities do not significantly differ from those derived from *ab initio* ones, but decreasing the amount of time and computational effort required to both works to build the density, instantly done within PASA DF, and to construct the MIDCOs. Also, the proposed distance restriction allows turning density evaluation in a linear process dependent only of the number of atoms composing the molecular system.

Even if a fitted representation is used, PASA DF are themselves a kind of DF that truly emulate a higher order DF. In this way, they can be included in some quantum mechanical calculations. As demonstrated, they can provide quantum similarity measures, even for large macromolecular systems, composed by up to 50,000 atoms or more, within reasonable time limits. Furthermore, if the proposed distance restriction is applied, the similarity calculation is still more appealing. Similarity measures computed with distance restrictions have been demonstrated to differ less than 0.005% from those calculated unrestricted.

ACKNOWLEDGEMENTS

This research has been supported by the Investigation Project No. SAF2000-0223-C03-01 from the Spanish *Ministerio de Ciencia y Tecnología*. One of us (X.G.) benefited from a predoctoral fellowship from the University of Girona. Financial support from the *Fundació Maria Francisca de Roviraltà* is also acknowledged.

REFERENCES

1. Mezey, P.G. Molecular Surfaces. In *Reviews in Computational Chemistry*; Lipkowitz, K.B., Boyd, D.B., Eds.; VCH Publishers: New York, 1990; Vol. 1, pp 265-294.
2. Mezey, P.G. The Topology of Molecular Surfaces and Shape Graphs. In *Computational Chemical Graph Theory*; Rouvray, D.H., Ed.; Nova: New York, 1990; pp 175-197.
3. Mezey, P.G. Non-Visual Molecular Shape Analysis: Shape Changes in Electronic Excitations and Chemical Reactions. In *Computational Advances in Organic Chemistry (Molecular Structure and Reactivity)*; Ogretir, C., Csizmadia, I., Eds.; Kluwer Academic Publishers: Dordrecht, 1991; pp 261-288.
4. Arteca, G.A.; Mezey, P.G. Algebraic Approaches to the Shape Analysis of Biological Macromolecules. In *Computational Chemistry: Structure, Interactions and Reactivity*, Part A; Fraga, S., Ed.; Elsevier: Amsterdam, 1992; pp 463-487.
5. Mezey, P.G. *Shape in Chemistry: An Introduction to Molecular Shape and Topology*; VCH Publishers: New York, 1993.
6. Mezey, P.G. Shape analysis. In *Encyclopedia of computational chemistry*; Schleyer, P.v.R., Allinger, N.L., Clark, T., Gasteiger, J., Kollman, P.A., Schaefer III, H.F., Schreiner P.R., Eds.; John Wiley & Sons: Chichester, UK, 1998; Vol. 4, pp 2582-2589.
7. Walker, P.D.; Mezey, P.G. Molecular electron density lego approach to molecule building. *J. Am. Chem. Soc.* **1993**, *115*, 12423-12430.
8. Walker, P.D.; Mezey, P.G. Ab initio quality electron densities for proteins: A MEDLA approach. *J. Am. Chem. Soc.* **1994**, *116*, 12022-12032.
9. Walker, P.D.; Maggiora, G.M.; Johnson, M.A.; Petke, J.D.; Mezey, P.G. Shape Group Analysis of Molecular Similarity: Shape Similarity of Six-Membered Aromatic Rings Systems. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 568-578.
10. Walker, P.D.; Mezey, P.G. Realistic, detailed images of proteins and tertiary structure elements: ab initio quality electron density calculations for bovine insulin. *Can. J. Chem.* **1994**, *72*, 2531-2536.
11. Mezey, P.G. Macromolecular density matrices and electron densities with adjustable nuclear geometries. *J. Math. Chem.* **1995**, *18*, 141-168.

12. Walker, P.G.; Mezey, P.G. A new computational microscope for molecules: High-resolution MEDLA images of taxol and HIV-1 protease, using additive electron density fragmentation principles. *J. Math. Chem.* **1995**, *17*, 203-234.
13. Mezey, P.G. Quantum chemistry of macromolecular shape. *Int. Rev. Phys. Chem.* **1997**, *16*, 361-388.
14. Gironés, X.; Amat, L.; Carbó-Dorca, R. A comparative study of isodensity surfaces using ab initio and ASA density functions. *J. Mol. Graph. Model.* **1998**, *16*, 190-196.
15. Gironés, X.; Carbó-Dorca, R.; Mezey, P.G. Applications of promolecular ASA densities to graphical representation of density functions of macromolecular systems. *J. Mol. Graph. Model.* **2001**, *19*, 343-348.
16. Mitchell, A.S.; Spackman, M.A. Molecular Surfaces from the Promolecule: A Comparison with Hartree-Fock Ab Initio Electron Density Surfaces. *J. Comput. Chem.* **2000**, *21*, 933-942.
17. Mezey, P.G. The holographic density theorem and quantum similarity measures. *Mol. Phys.* **1999**, *96*, 169-178.
18. Coppens, P.; Hall, M.B. (Eds.) *Electron Distribution and the Chemical Bond*; Plenum: New York and London, 1982.
19. Fliszár, S. *Charge Distributions and Chemical Effects*; Springer: New York, 1983.
20. Frisch, M.J.; Trucks, G. W.; Schlegel, H.B.; Scuseria, G.E.; Robb, M.A.; Cheeseman, J.R.; Zakrzewski, V.G.; Montgomery, Jr., J.A.; Stratmann, R.E.; Burant, J.C.; Dapprich, S.; Millam, J.M.; Daniels, D.; Kudin, K.N.; Strain, M.C.; Farkas, O.; Tomasi, J.; Barone, V.; Cossi, M.; Cammi, R.; Mennucci, B.; Pomelli, C.; Adamo, C.; Clifford, S.; Ochterski, J.; Petersson, G.A.; Ayala, P.Y.; Cui, Q.; Morokuma, K.; Malick, D.K.; Rabuck, A.D.; Raghavachari, K.; Foresman, J.B.; Cioslowski, J.; Ortiz, J.V.; Stefanov, B.B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Gomperts, R.; Martin, R.L.; Fox, D.J.; Keith, M. T.; Al-Laham, A.; Peng, C.Y.; Nanayakkara, A.; Gonzalez, C.; Challacombe, M.; Gill, P.M.W.; Johnson, B.; Chen, W.; Wong, M.W.; Andres, J.L.; Gonzalez, C.; Head-Gordon, M.; Replogle, E.S.; Pople, J.A. Gaussian 98, Revision A.6; Gaussian, Inc.: Pittsburgh PA, 1998.
21. Schmidt, M.W.; Baldridge, K.K.; Boatz, J.A.; Elbert, S.T.; Gordon, M.S.; Jensen, J.H.; Koseki, S.; Matsunaga, N.; Nguyen, K.A.; Su, S.; Windus, T.L.; Dupuis, M.;

- Montgomery, J.A. General Atomic and Molecular Electronic Structure System. *J. Comput. Chem.* **1993**, *14*, 1347-1363.
22. Coppens, P. *X-ray Charge Densities and Chemical Bonding*; Oxford: NewYork, 1997.
23. Rossman, M.G.; Arnold, E. (Eds.) *International Tables for Crystallography*, Vol. F; Kluwer Academic Publishers: Dordrecht, 2001.
24. Corey R.B.; Pauling L. Molecular models of amino acids, peptides and proteins; *Rev. Sci. Instr.* **1953**, *24*, 621-627.
25. Lee, B.; Richards, F.M. The interpretation of protein structures: Estimation of static accessibility; *J. Mol. Biol.* **1971**, *V55*, 379-400
26. Richardson, J. S. The anatomy and taxonomy of protein structure; *Adv. Protein Chem.* **1981**, *34*, 167-339.
27. Agishtein, M. E. Fuzzy molecular surfaces. *J. Biomolec. Struct. & Dynamics* **1992**, *9(4)*, 759-768.
28. Connolly, M.L. In *Network Science*, 1996 (<http://www.netsci.org/Science/Compchem/feature14.html>).
29. Foresman, J.B.; Frisch, A.E. *Exploring Chemistry with Electronic Structure Methods*; Gaussian, Inc.: Pittsburgh, 1996.
30. Scuseria, G.E. Linear Scaling Density Functional Calculations with Gaussian Orbitals. *J. Phys. Chem. A* **1999**, *103*, 4782-4790.
31. Mezey, P.G. Quantum similarity measures and Löwdin's transform for approximate density matrices and macromolecular forces. *Int. J. Quantum Chem.* **1997**, *63*, 39-48.
32. Hirshfeld, F.L.; Rzotkiewicz, S. Electrostatic Binding in First-Row Ah and A2 Diatomic-Molecules. *Mol. Phys.* **1974**, *27*, 1319-1343.
33. Spackman, M.A.; Maslen, E.N. Chemical properties from the promolecule. *J. Phys. Chem.* **1986**, *90*, 2020-2027.
34. Ruedenberg, K.; Schwarz, W.H.E. Nonspherical Atom Densities and Chemical Deformation Densities from X-Ray Scattering. *J. Chem. Phys.* **1990**, *92*, 4956-4969.
35. Coppens, P. The Structure Factor. In *International Tables for Crystallography*; Shmueli, U., Ed.; Kluwer Academic Publishers: Dordrecht, The Netherlands, 1996, Volume B., pp. 10-22.

36. Grant, J.A.; Gallardo, M.A.; Pickup, B.T. A Fast Method of Molecular Shape Comparison A simple application of a gaussian description of molecular shape. *J. Comput. Chem.* **1996**, *17*, 1653-1666.
37. Grant, J.A.; Pickup, B.T. A Gaussian description of molecular shape. *J. Phys. Chem.* **1996**, *100*, 2456-2456.
38. Carbó, R.; Arnau, J.; Leyda, L. How similar is a molecule to another? An electron density measure of similarity between two molecular structures. *Int. J. Quantum Chem.*, **1980**, *17*, 1185-1189.
39. Carbó-Dorca, R.; Besalú, E. A general survey of molecular quantum similarity. *TEOCHEM*, **1998**, *451*, 11-23.
40. Carbó-Dorca, R.; Amat, L.; Besalú, E.; Lobato, M. Quantum Similarity. In *Advances in Molecular Similarity*; Carbó-Dorca, R.; Mezey, P.G., Eds.; JAI Press: Greenwich, CT, 1998, Vol. 2, pp 1-42.
41. Carbó-Dorca, R. Fuzzy sets and Boolean tagged sets; vector semispaces and convex sets; quantum similarity measures and ASA density functions; diagonal vector spaces and quantum chemistry. In *Advances in molecular similarity*; Carbó-Dorca, R.; Mezey, P.G., Eds.; JAI Press: Greenwich, CT, 1998, Vol. 2, pp 43-72.
42. Besalú, E.; Carbó, R.; Mestres, J.; Solà, M. Foundations and recent developments on molecular quantum similarity, *Top. Curr. Chem.*, **1995**, *173*, 31-62.
43. Carbó, R.; Besalú, E.; Theoretical foundations of quantum molecular similarity. In *Molecular similarity and reactivity: from quantum chemical to phenomenological approaches*; Carbo, R., Ed.; Kluwer: Amsterdam, 1995, pp 3-30.
44. Carbó-Dorca, R.; Besalú, E.; Amat, L.; Fradera, X. Quantum molecular similarity measures: concepts, definitions and applications to quantitative structure-properties relationships. In *Advances in molecular similarity*; Carbó-Dorca, R.; Mezey, P.G., Eds.; JAI Press, Greenwich, CT, 1996, Vol. 1, pp 1-42.
45. Pisabarro, M. T.; Ortiz, A. R.; Serrano, L.; Wade, R. C. Homology modeling of the Abl-SH3 domain. *Proteins* **1994**, *20*, 203-215.
46. Yoshikawa, S.; Shinzawa-Itoh, K.; Nakashima, R.; Yaono, R.; Yamashita, E.; Inoue, N.; Yao, M.; Fei, M.J.; Libeu, C.P.; Mizushima, T.; Yamaguchi, H.; Tomizaki, T.; Tsukihara, T.; Redox-Coupled Crystal Structural Changes in Bovine Heart Cytochrome C Oxydase. *Science* **1998**, *280*, 1723-1729.

47. Amat, L.; Carbó-Dorca R. Quantum Similarity Measures under Atomic Shell Approximation: First-Order Density Fitting using Elementary Jacobi Rotations. *J. Comput. Chem.* **1997**, *18*, 2023-2039.
48. Amat, L.; Carbó-Dorca, R. Fitted Electronic Density Functions from H to Rn for Use in Quantum Similarity Measures: cis-Diammine-Dichloroplatinium(II) Complex as an Application Example. *J. Comput. Chem.* **1999**, *20*, 911-920.
49. Amat, L.; Carbó-Dorca, R. Molecular Electronic Density Fitting Using Elementary Jacobi Rotations under Atomic Shell Approximation. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1188-1198.
50. Mulliken, R.S. Electronic Population Analysis on LCAO Molecular Wave Functions. I. *J. Chem. Phys.* **1955**, *23*, 1833-1840.
51. Mulliken, R.S. Electronic Population Analysis on LCAO Molecular Wave Functions. II. Overlap Populations, Bond Orders, and Covalent Bond Energies. *J. Chem. Phys.* **1955**, *23*, 1841-1846.
52. Mulliken, R.S. Electronic Population Analysis on LCAO Molecular Wave Functions. III. Effects of Hybridization on Overlap and Gross AO Populations. *J. Chem. Phys.* **1955**, *23*, 2338-2342.
53. Mulliken, R.S. Electronic Population Analysis on LCAO Molecular Wave Functions. IV. Bonding and Antibonding in LCAO and Valence-Bond Theories. *J. Chem. Phys.* **1955**, *23*, 2343-2346.
54. Mulliken, R.S. Criteria for the Construction of Good Self-Consistent-Field Molecular Orbital Wave Functions, and the Significance of LCAO-MO Population Analysis. *J. Chem. Phys.* **1962**, *36*, 3428-3439.
55. Gironés, X.; Amat, L.; Carbó-Dorca, R. ASACUBE v1.0, Institute of Computational Chemistry: University of Girona, Girona, 2000.
56. Protein Data Bank, Research Collaboratory for Structural Bioinformatics (<http://www.rcsb.org/pdb>). Main Citation: Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E. The Protein Data Bank. *Nucleic Acids Research* **2000**, *28*, 235-242.
57. Walters, P.; Stahl, M. Babel, version 1.6; University of Arizona: Arizona, 1996.
58. Constants, P.; Amat, L.; Carbó-Dorca, R. Toward a Global Maximization of the Molecular Similarity Function: Superposition of Two Molecules. *J. Comput. Chem.* **1997**, *18*, 826-846.

Table I.- Dimension of the molecular boxes, resolution used and time required to compute MIDCOs for a series of macromolecular structures.

Structure (PDB Id)	# atoms	Dimensions of the molecular box (a.u.)	Resolution (a.u.)	Time (seconds)
1ABL	872	65×71×63	0.5	18
1RVA	8686	129×121×147	1.0	24
1QLE	21358	275×237×177	1.0	83
1OCO	56863	299×245×245	1.0	177

Note: resolution stands for spacing between points in the grid. 1RVA corresponds to *Eco Rv Endonuclease* and 1QLE to *Paracoccus Denitrificans Cytochrome C Oxydase*. Both structures have been identically treated as 1OCO.

Table II.- Results of the MQS-SM (a.u.) calculations for the proteins. The unrestricted values correspond to the full calculation, whereas the restricted ones correspond to those where the distance threshold has been applied.

Structure (PDB Id)	Number of atoms	MQS-SM (unrestricted)	Time (seconds)	MQS-SM (restricted)	Time (seconds)
1ABL	872	20994.86	45.76	20993.00	4.88
1RVA	8686	218541.39	4701.3	218531.17	475.30
1QLE	21358	551956.91	40200	551929.00	3044.8
1OCO	56863	1461628.55	282480	1461626.52	3283.9

Figure 1.- MIDCOs of protein 1ABL at four isodensity levels: (a) 0.250, (b) 0.150, (c) 0.050 and (d) 0.005. (In e/au^3)

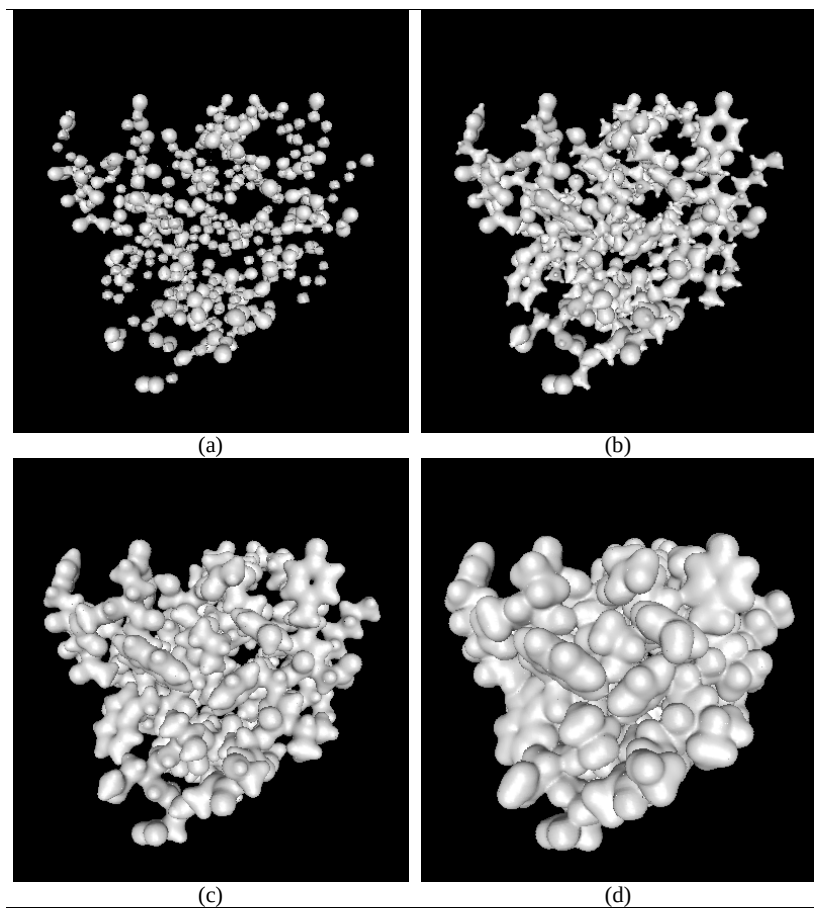
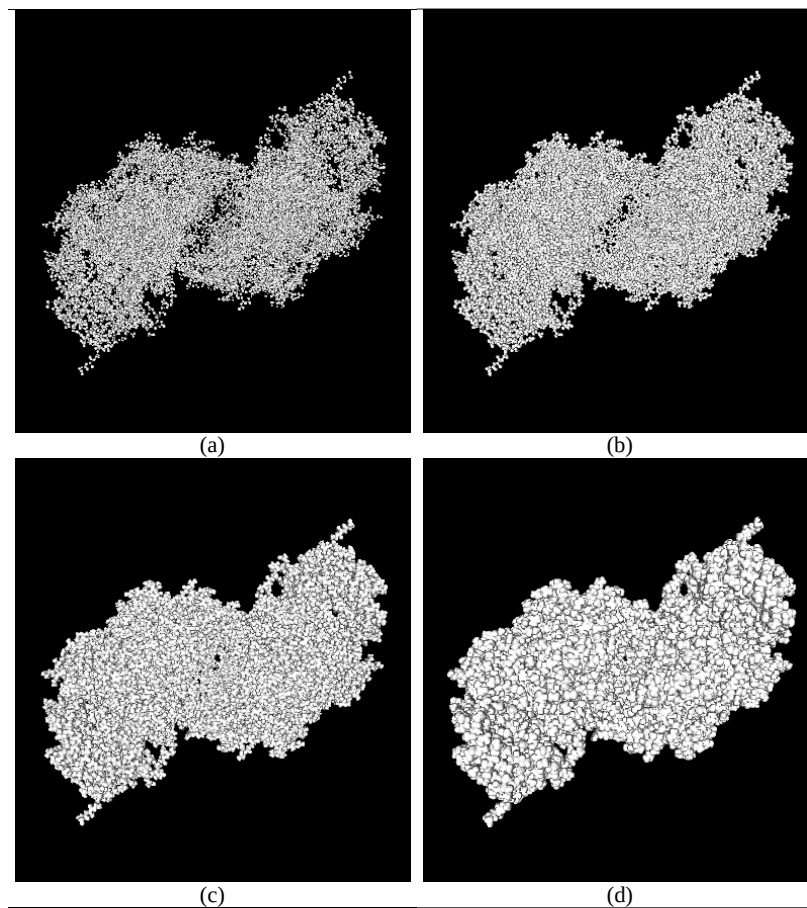


Figure 2.- MIDCOs of protein 1OCO at four isodensity levels: (a) 0.250, (b) 0.150, (c) 0.050 and (d) 0.005. (In e/au^3)



4. Relacions Quantitatives Estructura-Activitat (QSAR)

- **Resum:** En aquest capítol es donarà un repàs temporal de les tècniques QSAR, des dels seus primitius orígens fins a les avançades tècniques de modelatge molecular assistit per ordinador de l'actualitat. Així mateix també es farà un incís especial en base a la necessitat de sobreposició molecular en aquestes tècniques. Finalment es farà una pinzellada sobre les tècniques estadístiques més emprades.
- **Paraules Clau:** QSAR (Relacions Quantitatives Estructura-Activitat), QSPR (Relacions Quantitatives Estructura-Propietat), MQS (Semblança Molecular Quàntica), MQSM (MQS Moleculars)

4.1. Importància del QSAR

El gran desenvolupament que ha sofert la química en els últims anys en totes les seves vessants experimentals ha portat a la síntesi d'una xifra espectacular de compostos. En el cas concret de la química combinatòria, on s'apliquen complexos sistemes d'enginyeria per a la producció de múltiples compostos amb característiques estructurals diverses, es provoca la nova aparició de centenars de milers de compostos nous cada any. En xifres aproximades, es calcula que només pel que fa a compostos orgànics, la quantitat de nous compostos sintetitzats volta els 3 milions de compostos per any, amb els costos econòmics que aquesta xifra significa.

Moltes d'aquestes noves molècules sintetitzades són tan sols avaluades en base a una propietat molt concreta, deixant altres possibles característiques potencials de banda. El cas més típic es dona en recerca farmacèutica, on es seleccionen aquelles molècules que donen resposta en reaccionar amb un determinat anticòs, descartant-se la resta.

Donades aquestes xifres sobre nous compostos sintetitzats i el potencial interès d'aconseguir noves molècules amb propietats concretes, no és d'estranyar l'acceptació donada a les tècniques de

disseny molecular assistit per ordinador (CADD, *Computer-Aided Drug Design*). L'objectiu de dites tècniques és aconseguir determinar les propietats d'una estructura, la qual no té perquè haver estat sintetitzada prèviament, de forma totalment virtual, de manera que es poden estalviar els costos derivats de la recerca experimental, o com a mínim reduir-los en base a encarar la recerca sobre aquelles molècules on el mètode teòric dóna més possibilitats.

Per a tal de dur a terme la seva tasca, el CADD està fonamentat per les anomenades Relacions Quantitatives Estructura-Propietat (QSAR, *Quantitative Structure-Activity Relationships*). Aquesta disciplina, recentment estructurada i racionalitzada, pretén establir relacions matemàtiques senzilles entre unes propietats donades per a conjunts de molècules amb paràmetres derivats de l'estructura de les mateixes, aconseguint així poder descriure tals propietats en funció dels paràmetres estructurals escollits. Dits paràmetres poden ser utilitzats posteriorment per a predir les propietats considerades per a noves molècules on tal característica no és coneguda, a partir de l'ús dels paràmetres que es deriven de les estructures dels nous compostos. En l'actualitat, i segons el camp d'estudi, les tècniques QSAR¹ es divideixen en tres subgrups majoritaris:

- QSAR pròpiament dit, on s'avaluen activitats i respostes biològiques.
- QSPR (*structure-property*), on es consideren propietats fisicoquímiques.
- QSTR (*toxicity-relationships*), que comprèn les relacions entre la toxicitat de les molècules amb les seves estructures. Tot i que la toxicitat és un cas concret de propietat biològica, on el subjecte de l'estudi mor o resulta seriosament perjudicat, la seva importància en els camps medioambientals o de la salut les diferencia de les QSAR.

4.2. Del Segle XIX al XXI, breu cronologia del QSAR

4.2.1. Els Inicis

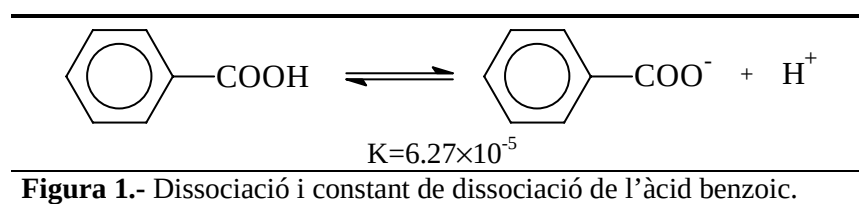
Els primers indicis del que es pot considerar un intent de relacionar una resposta biològica d'una sèrie de compostos amb una propietat fisico-química molecular es remunta a mitjans del segle XIX, concretament al 1863. En aquell temps, Cros [1], afiliat a la Universitat d'Estrasburg, va observar com la toxicitat de certs alcohols en mamífers augmentava de manera inversament proporcional a la seva solubilitat en aigua. Posteriorment al 1865, Crum-Brown i Fraser [2] elucidaren una dependència entre els que ells anomenaren activitats fisiològiques, resposta a la

¹ Tot i que com es diu, les tècniques es bifurquen segons el camp d'aplicació, tanmateix el nom QSAR s'emprarà de forma general al llarg d'aquest capítol, podent-se substituir per QSPR o QSTR si es vol concretar en un àmbit determinat.

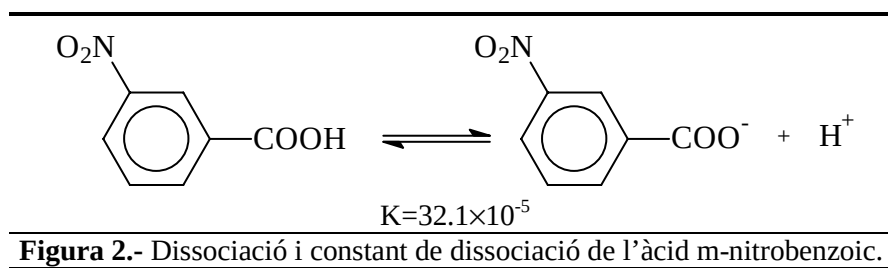
injecció en cobaies de derivats de la morfina, atropina i estricnina, i les estructures químiques de les substàncies injectades. Poc després aparegué l'estudi de Richet sobre la toxicitat de certs compostos orgànics oxigenats i la seva dependència inversament proporcional a la solubilitat en aigua [3]. Ja a tocar al canvi de segle, Meyer [4] i Overton [5], de manera totalment independent, postularen un coeficient de distribució de solubilitats entre aigua i certs lípids, amb el qual establiren relacions lineals amb l'acció narcòtica d'alguns compostos; val a dir que tal coeficient de distribució evolucionaria fins al concepte de coeficient de distribució octanol-aigua, més conegut com a $\log P$, àmpliament emprat avui en dia. Ja entrats al segle XX, Ferguson [6] va realitzar un estudi extensiu sobre la relació entre la toxicitat de sèries homòlogues de molècules i algunes de les seves respectives propietats físiques, com per exemple, la solubilitat en aigua, distribució entre dues fases, capil·laritat i pressió de vapor.

4.2.2. Les relacions d'energia lliure

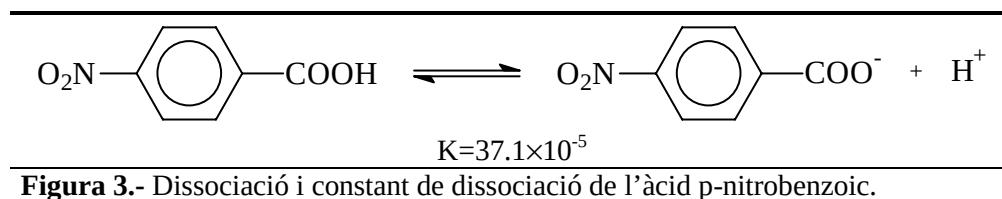
A final de la dècada dels 30, Hammett [7,8] va iniciar el que esdevindria el camp dels efectes electrònics, al relacionar constants de ionització i reactivitat de sèries de benzoàcids segons el tipus de substitució i la seva posició² en l'anell benzènic. Si es considera la dissociació següent:



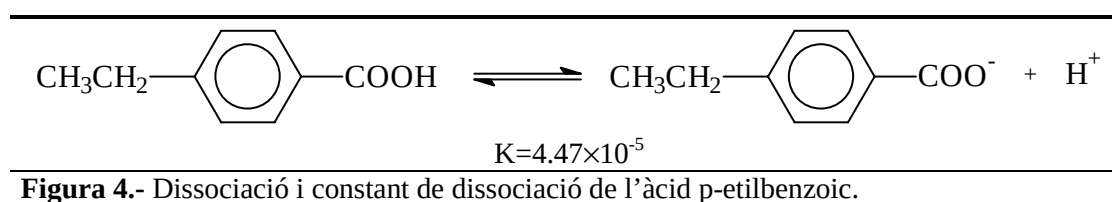
Hammett observà que l'addició de substituents a l'anell aromàtic de l'àcid benzoïc tenia un efecte ordenat i quantitatiu en la constant de dissociació. Si, per exemple s'observa la reacció següent:



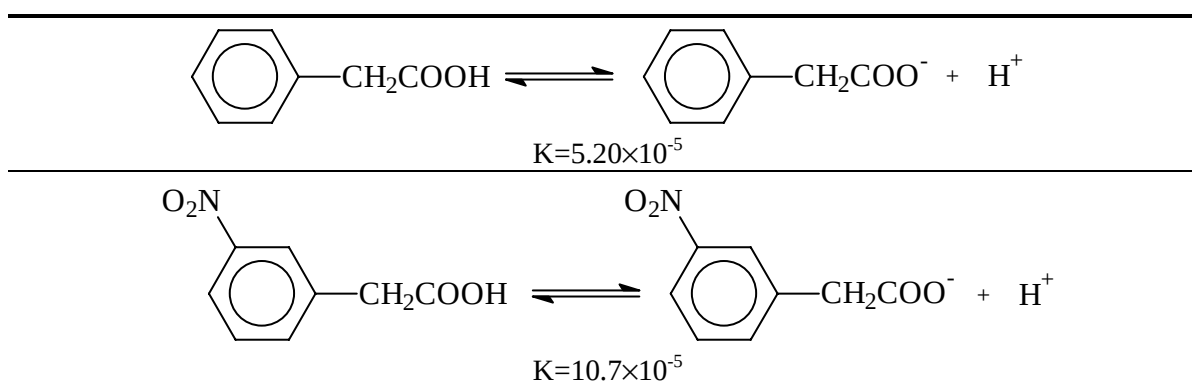
un grup nitro a la posició meta incrementa la constant de dissociació, degut a que el grup nitro és electró-acceptor, o d'altra manera, estabilitza la càrrega negativa creada en la dissociació. Si es considera el grup nitro en posició para, tal i com es mostra a la figura 3:



la constant d'equilibri és encara major, indicant un increment en la capacitat d'acceptació electrònica. D'altra banda, si es considera un grup etil en posició para, com a la figura 4:



en aquest cas, la constant de dissociació és menor que la del compost original, indicant que el grup etil és electró-donador, o dit d'altra manera, desestabilitza la càrrega negativa generada en el procés de dissociació. Hammett també observà que els substituents tenen un efecte similar en la dissociació d'altres àcids i bases. Per exemple, si es consideren les reaccions d'alguns àcids fenilacètics mostrats a la figura 5:



² Només es consideraren les posicions *meta* i *para*, ja que la substitució en *orto* provocava efectes estèrics difícils de parametritzar per l'època.

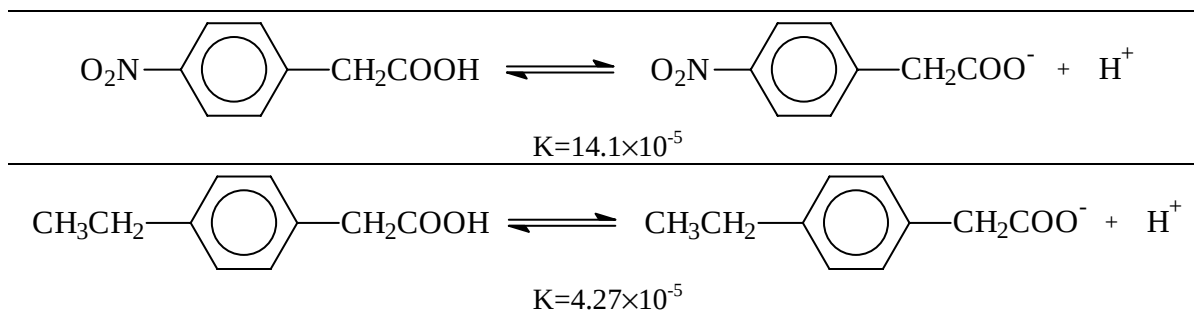


Figura 5.- Dissociacions i constants de dissociació d'àcids fenilacètics.

l'efecte estabilitzador del grup nitro afavoreix la dissociació, en major mesura en posició para que en meta, tal i com s'observa en l'àcid benzoïc. Anàlogament, el grup metil desplaça la reacció a reactius. Aquests equilibris acostumen a representar-se gràficament com es mostra a la següent figura 6:

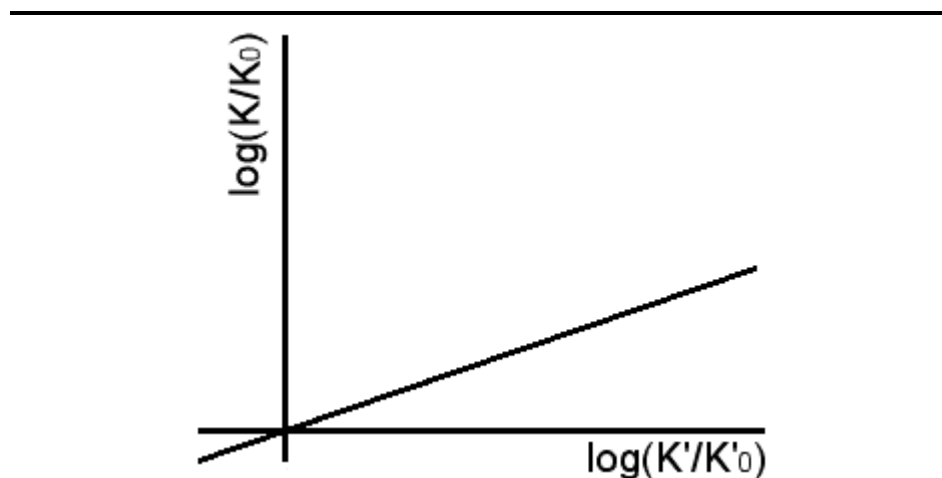


Figura 6.- Exemple de gràfic de l'evolució de les constants d'equilibri amb diverses substitucions. K_0 i K'_0 representen les constants d'equilibri dels compostos no-substituïts, i K i K' les dels compostos substituïts. Els valors en l'eix de les X són calculats a partir de les constants de dissociació dels derivats de l'àcid benzoïc i els de l'eix de les Y a partir d'una altra sèrie.

Degut a que tal relació és lineal, es pot escriure la següent equació:

$$\log \frac{K}{K_0} = \rho \log \frac{K'}{K'_0}, \quad (1)$$

on ρ és el pendent de la gràfica. Els valors de l'eix de les X de la figura 6 són referits als derivats de l'àcid benzoïc i es representen pel símbol σ . Llavors es pot escriure:

$$\log \frac{K}{K_0} = \rho \log \sigma, \quad (2)$$

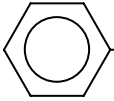
El pendent ρ és una constant de proporcionalitat pertanyent a un equilibri en concret, i relaciona l'efecte dels substituents en aquest equilibri a l'efecte dels substituents en l'equilibri de dissociació de l'àcid benzoic. Llavors, si l'efecte dels substituents es proporcionalment major que en la reacció de l'àcid benzoic $\rho > 1$, i menor que 1 en cas contrari. Per definició, ρ per l'àcid benzoic és la unitat.

σ és el descriptor dels substituents. La magnitud de σ dona la força relativa en l'acceptació o donació electrònica dels substituents. σ és positiva si són electró-acceptors, i negativa si són electró-donadors. Per definició, σ per $-H$ és zero.

Aquestes relacions, desenvolupades per Hammett, reben el nom de *relacions d'energia lliure* degut a la seva similitud amb l'equació que relaciona l'energia lliure a una constant d'equilibri: $\Delta G = -RT \ln K$, que expressada en paraules significa que l'energia lliure és proporcional al neperià de la constant d'equilibri. Aquestes relacions d'energia lliure també s'anomenen *extratermodinàmiques* degut a que poden ser formulades via paràmetres termodinàmics, tot i que cap principi estableix que tals relacions siguin vertaderes.

Per a una millor comprensió d'aquestes relacions, és convenient considerar alguns valors de ρ i σ , tal i com es mostren a les següents taules 1 i 2:

Taula 1.- Exemples de valors de ρ per a diferents equilibris.

	NH_3^+	$\rho=2.90$
	OH	$\rho=2.23$
	CH_2COOH	$\rho=0.49$
	CH_2CH_2COOH	$\rho=0.21$

En els equilibris de l'anilina i el fenol, el H^+ dissociat està en una posició la qual està a una distància d'enllaç de l'anell benzènic, mentre que a l'àcid benzoic, tal protó es troba a un grup separat per dues distàncies; llavors es comprensible que els substituents de l'anell benzènic exerceixin un efecte més notori en la dissociació de l'anilina i el fenol que no en l'àcid benzoic, donant un ρ major que 1. Per altra banda, en els àcids fenilacètic i fenilpropioníc, el protó es troba en grups que disten 3 i 4

distàncies d'enllaç de l'anell benzènic, resultant en un menor efecte per part dels substituents en la seva dissociació, i per tant, ρ és menor que la unitat.

Taula 2.- Exemples de valors de σ per a diferents substituents.

Substituent	σ_m	σ_p
-H	0.0	0.0
-NO ₂	0.71	0.78
-Cl	0.37	0.23
-OCH ₃	0.12	-0.27

Els valors positius de σ per al grup nitro indiquen la seva capacitat electró-acceptora. Per a la comprensió dels valors de σ pel grup nitro en posicions meta i para s'han de considerar els mecanismes de l'acceptació i la donació electrònica. Un grup nitro en posició meta accepta electrons degut a un efecte inductiu provocat per l'electronegativitat dels àtoms constituents. Si tan sols es considerés la inducció, llavors s'esperaria un efecte menor en la posició para, tanmateix aquest valor és major i resulta de la combinació dels efectes inductius i de ressonància. Pel clor, l'electronegativitat de l'àtom provoca un efecte inductiu d'acceptació electrònica, i només tal efecte és possible. El grup metoxi pot ser electró-donador o acceptor segons la seva posició. L'electronegativitat de l'oxigen resulta, en posició meta, en un efecte inductiu d'acceptació d'electrons; tanmateix, en posició para, tal efecte inductiu és minoritari, i l'efecte de ressonància per la donació electrònica resulta globalment majoritari. Tal i com s'ha dit, l'H es considera el zero de referència. La literatura ha reflexat l'interès pels valors d'aquesta constant, publicant-se nombroses taules per a múltiples substituents [9,10] i diferents aplicacions [11,12]. En alguns casos, els valors de σ són aplicables a molts tipus diferents d'equilibris, i en altres, s'han derivat valors de σ per a equilibris específics, el qual és particularment interessant si es volen tractar substituents en posició orto.

Aquest estudi de Hammett fou el primer en *particionar* la molècula, i explicar la seva activitat a partir de tan sols fragments d'aquest, enlloc de referir-la a la totalitat. Altres contribucions d'aquest estil, on es consideren els substituents es deuen al treball de Taft sobre els efectes estèrics [13] (la coneguda E_s de Taft), o els estudis dels efectes de camp i ressonància de Swain [14]. Diferents estudis es duren a terme en anys posteriors en aquesta línia [15-20], malgrat no gaudiren del reconeixement atorgat als treballs anteriors.

Un particionat diferent de la molècula fou proposat al 1964 per Free i Wilson [21], postulant que l'activitat biològica d'un conjunt molecular es pot relacionar amb l'addició de

substituents, tenint en compte el seu nombre, tipus i posició en l'esquelet comú. D'aquesta manera es podia discretitzar l'activitat biològica com a una simple suma de contribucions; tanmateix, un dels principals problemes d'aquest model era la impossibilitat de tractar noves molècules amb característiques no presents en el conjunt emprat per a la construcció del model (també anomenat d'entrenament, de l'anglès *training set*). Els estudis de Free-Wilson foren presos com a punt de partida per Bocek i Kopecký [22,23], els quals construïren models de Free-Wilson amb la inclusió de termes creuats per tenir en compte les probables interaccions entre els diferents substituents propers entre ells. Posteriorment, Fujita i Ban [24] simplificaren l'equació de Free-Wilson construint models per a molècules substituïdes a partir d'estimacions sobre l'estructura base.

4.2.3. Anàlisi de Hansch

Els models QSAR basats en les constants de Hammett, tal i com s'ha comentat, utilitzen les propietats electròniques dels substituents com a descriptors moleculars. Tot i l'èxit aconseguit en diversos estudis, algunes dificultats van aparèixer al aplicar relacions de tipus Hammett a sistemes biològics, indicant la necessitat d'introduir nous descriptors estructurals.

Muir, un botànic de la Universitat de Pomona, estudià l'activitat biològica d'un conjunt de compostos semblants als àcids indolacètic i fenoxiacètic, els quals actuen de reguladors del creixement de les plantes. Quan va intentar correlacionar les estructures dels compostos amb les seves respectives activitats, va consultar el seu company de química, Corwin Hansch. Emprant únicament sigmes de Hammett, per a tenir en compte els efectes electrònics dels substituents, els resultats no van portar a models QSAR significatius. Tanmateix, Hansch va suggerir que l'ús de la lipofilicitat, expressada com el coeficient de partició octanol-aigua ($\log P$), podia ser rellevant en l'activitat biològica [25]. Avui en dia, es reconeix que aquest paràmetre proporciona una mesura de la *biodisponibilitat* d'un compost, el qual determina en part, la quantitat de compost que penetra les barreres cel·lulars. Així s'assimila el sistema octanol-aigua com a un model de membrana cel·lular (normalment lipofílica, tal i com l'octanol és) i entorn de transport (normalment hidrofílic, i constituït majoritàriament per aigua).

D'aquesta manera es van desenvolupar models de correlació emprant aquest paràmetre estructural, la lipofilicitat, amb diverses activitats. En alguns casos, un model d'una sola variable era suficient, prenent la forma de:

$$\log\left(\frac{1}{C}\right) = a \log P + b, \quad (3)$$

on C és la concentració molar d'un compost, el qual provoca una resposta estàndard, com per exemple el conegut LD_{50} ³. Amb altres conjunts de dades, s'observà que les correlacions milloraven al combinar els paràmetres electrònics de Hammett amb la mesura de la lipofilicitat de Hansch, emprant equacions híbrides com:

$$\log\left(\frac{1}{C}\right) = k_1\pi + k_2\sigma + k_3, \quad (4)$$

on σ és la sigma de Hammett del substituent considerat i π es defineix [26], anàlogament a σ com a:

$$\pi = \log\left(\frac{P_X}{P_H}\right). \quad (5)$$

En altres casos, relacions parabòliques eren necessàries entre la resposta biològica i la lipofilicitat, i s'observà que es podia ajustar emprant el terme $(\log P)^2$ en el model QSAR. La inclusió d'aquest terme fou justificat en base a que moltes membranes cel·lulars han de ser travessades pels compostos fins a arribar al centre actiu, i aquells amb major lipofilicitat es trobaran localitzats a les primeres membranes que es trobin inicialment.

En l'actualitat, i prenent el treball de Hansch [27] com a punt de partida, els models QSAR es construeixen amb una varietat molt àmplia de descriptors, considerant-se que l'activitat biològica és el resultat de la combinació de múltiples factors, els quals actuen de forma independent i s'ajunten emprant un model de correlació multilíneal. Alguns dels descriptors més utilitzats són:

- *Descriptors geomètrics*: relacionats amb la forma i mida molecular. Els més emprats dins d'aquesta categoria són el volum molecular [28,29], l'àrea molecular superficial [29-34] i l'àrea superficial parcialment carregada [35,36]. Un cas particular de descriptor geomètric força estès és la variable discreta en funció de la presència o absència d'un substituent, sobretot emprat en models QSAR complexos.
- *Descriptors polars*: relacionats amb les forces de dispersió, representada normalment per la refractivitat molar [10,37-39], entesa com el volum molar corregit per un índex de refracció [40].
- *Descriptors electrònics i energètics*: representen diferents paràmetres electrònics presents en les estructures. Apart de la coneguda sigma de Hammett, també es troben dades tabulades i derivades de càlculs mecan quàntics i semiempírics. Dins d'aquest grup de descriptors, els més emprats són les energies dels orbitals frontera HOMO i

³ Per LD_{50} s'entén concentració letal (de l'anglès *Lethal Concentration*). Altres indicadors són el DL_{50} (Death Level) o IC (Inhibition Concentration)

LUMO, la diferència dels quals proporciona una idea del caràcter nucleofílic o electrofílic molecular [38].

- *Descriptors estèrics*: representen la massa molecular i posen de relleu l'efecte de la variació dels substituents d'un esquelet molecular comú, com pot ser la E_s de Taft [13,41], o per a descriure la importància de l'amplada o llargada d'un grup (paràmetres Verloop) [42].

4.2.4. Estudis Topològics

De forma independent al desenvolupament de les relacions extratermodinàmiques anteriorment exposades, també sortí la topologia molecular com a font de descriptors per a la construcció de models QSAR, entre d'altres aplicacions.

La topologia [43-49] postula que la informació bàsica sobre els compostos orgànics està continguda a la seva taula de connectivitats, la qual és una representació compacte del tipus d'àtom, enllaços i connexions que formen la molècula. Tot i que els primers pilars sobre els que es fonamenta la topologia es remunten al segle XIX, la concepció moderna es deu al treball de Wiener [43], continuat pel de Kier i Hall [44] i Randić [45]. La següent figura 7 mostra un exemple de taula de connectivitats.

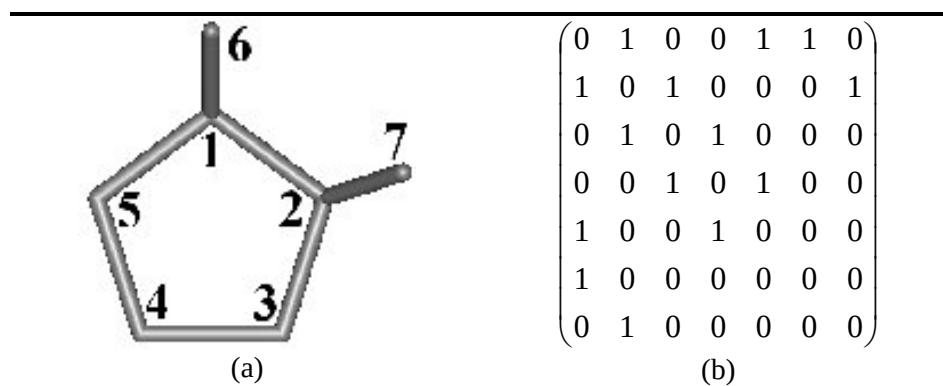


Figura 7.- Taula de connectivitats per al cis 1,2-dimetilciclopentà. (a) estructura, (b) taula de connectivitats, on 1 significa enllaç i 0 en cas contrari.

El fet que la taula de connectivitats siguin usualment emprades en la codificació de molècules en bases de dades, permet que aconseguir, d'una manera ràpida i accessible, descriptors moleculars mitjançant els anomenats *índexs topològics* [43-54], que són les reduccions numèriques de les taules de connectivitat i distàncies, els quals han donat resultats satisfactoris en les relacions QSAR [48,49].

4.2.5. Mètodes espaials

Un dels problemes principals de les tècniques anteriorment exposades és la dificultat, o en alguns casos total impossibilitat, de tractar diferents conformacions o de tenir en compte l'estereoquímica en les estructures moleculars tractades. D'aquí que l'aparició de metodologies que tractaven aquests paràmetres, de forma més o menys complexa, però sistemàtica, suposà un autèntica transformació en les tècniques QSAR. Usualment, aquests models involucren el càlcul de diferents magnituds en l'espai que envolta les estructures estudiades, com per exemple potencials electroestàtics, forces de London, densitat de càrrega o energia d'interacció amb un compost de test (normalment anomenat *probe*). Per coherència en la comparació de les magnituds calculades en les diferents molècules, cal que aquestes estiguin alineades segons algun criteri, normalment en base a un farmacòfor comú, derivat d'estudis lligand-receptor.

La primera tècnica d'aquest tipus, que data de l'any 1988, fou el resultat de la investigació de Crammer [55], la qual gaudeix d'una gran projecció en l'actualitat. Des de l'aparició d'aquest mètode, altres tècniques han aparegut o s'han adaptat per a la creació de models QSAR. Alguns exemples són el CoMSIA [56], CoMMA [57], Compass [58], MS-WHIM [59], E-State [60], PARM [61] i les Mesures de Semblança Molecular Quàntica (MQSM) aplicades a QSAR, les quals es detallaran en el següent capítol.

4.2.6. El disseny molecular assistit per ordinador

El disseny de molecular, o de fàrmacs, assistit per ordinador (CADD) consisteix una de les més recents aplicacions dels ordinadors com a eines en el procés de disseny de nous compostos. Tot i que l'ordinador no és cap mena de substitut en la comprensió clara dels sistema a estudiar, si que ofereix informació i proporciona un enfoc diferent sobre la 'química' o 'biologia' del problema a tractar.

En la major part de les aplicacions del CADD s'intenta trobar aquell lligand (molècula d'interès) que interaccionarà favorablement amb un receptor, el qual representa al centre actiu. La unió amb tal receptor pot incloure interaccions de tot tipus, tals com electroestàtiques, hidrofòbiques i de pont d'hidrogen. A més a més, les energies de solvatació són rellevants, ja que cal una desolvatació parcial o total abans de la unió lligand-centre.

Aquesta tècnica normalment optimitza l'ajust entre els dos sistemes, tanmateix una unió computacionalment òptima no garanteix que l'activitat del lligand s'incrementi, o que d'altra banda

redueixi els efectes secundaris. De fet, al no tenir en compte la farmacocinètica del lligand, el resultat final pot no ser molt realista.

Les aproximacions realitzades en CADD depenen de la quantitat d'informació disponible sobre el receptor i el propi lligand. Idealment, es disposaria de informació estructural tridimensional del receptor i del complex lligand-receptor a partir de difracció de raig X o ressonància magnètica nuclear; tanmateix, aquestes dades estan rarament a l'abast. Per l'altra banda, la total indisponibilitat de dades experimentals sobre les quals basar els càlculs proporcionen llibertat d'acció, gràcies a la manca de restriccions que imposarien les dades empíriques.

Segons la informació disponible, es pot aplicar disseny basat en el lligand o en el receptor. Els mètodes basats en el lligand són aplicables quan l'estructura del receptor és desconeguda, però es disposa d'un conjunt de molècules, en les quals s'ha determinat l'activitat d'interès. Per a major eficiència, es disposaria de compostos estructuralment similars que cobreixin un ampli rang de valors de l'activitat. En tècniques de mapejat de centres, s'intenta reconèixer el farmacòfor, entès com a un patró comú en el conjunt molecular, representat com a un conjunt de grups funcionals en l'espai tridimensional, el qual és complementari a la geometria⁴ del centre actiu.

Al aplicar el mètode basat en el lligand s'ha de realitzar una anàlisi conformacional, la magnitud del qual dependrà de la flexibilitat dels compostos estudiats. Una possibilitat és prendre les geometries que donin la mínima energia segons algun criteri de càlcul. Entre d'altres possibilitats, es pot minimitzar tan sols una sola molècula (normalment la més rígida o la més activa) i sobreposar la resta sobre aquesta amb les adequades restriccions. Finalment, al tenir totes les estructures sobreposades es pot generar el farmacòfor. El patró derivat pot ser llavors emprat per a la generació de nous compostos, afegint o traient grups funcionals en posicions rellevants o d'interès.

Cal dir que en l'aplicació d'aquest mètode, s'assumeix que els confòrmers de mínima energia són aquells que s'uniran més favorablement al centre actiu, quan de fet no existeix cap raó per excloure confòrmers de més energia.

Els mètodes basats en el receptor s'apliquen quan existeix un model verídic del centre actiu, bé sigui per raig X, ressonància magnètica nuclear o modelatge homòleg. Amb la geometria del centre actiu, el problema del disseny de lligands que interaccionaran favorablement al centre es trasllada a l'acoblament (de l'anglès *docking*).

⁴ Val a dir que no només existeixen complementarietats geomètriques, sinó també d'altres tipus, com són les electroestàtiques i les hidrofòbiques.

4.3. Sobreposició en QSAR

Dins de les múltiples metodologies existents destinades a l'obtenció de descriptors moleculars, hi ha aquelles que requereixen l'alineament molecular a fi de calcular una funció intermolecular o, entre d'altres possibles càlculs, establir comparacions de propietats fisicoquímiques.

En els primers estudis QSAR, i fins a l'aparició dels mètodes espacials, els descriptors corresponents a una molècula s'obtenien directament i única, amb més o menys complexitat, a partir de l'estructura. Descriptors moleculars tipus log P (teòric), sigma de Hammett, tot el rang d'índexos topològics, ... no requereixen cap mena de sobreposició amb d'altres molècules per a realitzar el càlcul.

Amb l'aparició d'aquestes noves tècniques, va aparèixer la necessitat de mètodes de sobreposició i de diferents criteris en que basar-se. Les dues línies principals de sobreposició molecular són:

- aquelles basades en les posicions i tipus d'àtoms (atom type): enfoquen l'alineació molecular en base a la coincidència de fragments, grups o subestructures comunes de la molècula.
- aquelles basades en camps (field type): orientades al màxim recobriment de diferents camps moleculars, com són els electrostàtics o els estèrics.

En el capítol 5, el tema de la sobreposició molecular, i en concret una proposta duta a terme des del nostre grup, s'aborda amb més profunditat.

4.4. Tècniques estadístiques d'ajust

En aquest sub-apartat, es donaran breus introduccions a diferents tècniques estadístiques, sovint emprades per a relacionar descriptors moleculars, generats a partir de qualsevol metodologia, amb una propietat molecular d'interès. Tot i la gran quantitat de protocols estadístics existents, la discussió es centrarà bàsicament en la regressió multilíneal, l'anàlisi de components principals, el mínims quadrats parcials i les xarxes neurals. Al final es donarà una breu pinzellada a d'altres tècniques.

4.4.1. Introducció

Usualment en els mètodes QSAR, els descriptors moleculars ($\mathbf{X}(n \times m)$, on $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m\}$) s'intenten correlacionar amb una, tot i que també poden ser vàries, variable de resposta ($\mathbf{y}(n)$). L'objectiu de l'anàlisi és normalment incrementar el coneixement sobre un sistema biològic, o efectuar prediccions sobre compostos no avaluats, o fins i tot no sintetitzats. Les conclusions que es poden donar després d'una anàlisi de regressió depenen del model triat [62]. Si es creu que la relació és acceptable emprant un model lineal dins de les variables de regressió, un possible model podria ser:

$$\mathbf{y} = b_0 + b_1\mathbf{x}_1 + b_2\mathbf{x}_2 + \mathbf{e}. \quad (6)$$

En l'anterior equació, les $\{b_i\}$ són constants desconegudes anomenades coeficients de regressió, i l'objectiu de la anàlisi és justament donar-ne una estimació. El nombre de mètodes de regressió disponibles en l'actualitat és gran i en QSAR, sobretot en la última dècada, la regressió multilíneal [62,63] s'ha desplaçat a favor de mètodes de projecció, com l'anàlisi de components principals o els mínims quadrats parcials [63,64].

És prou ben conegut en anàlisis de regressió que un pretractament adequat de les dades pot arribar a ser crucial per a la qualitat del resultat. Els descriptors moleculars acostumen a diferir en rang, variació i mida. Conseqüentment, previ a l'anàlisi de regressió s'aplica un escalat a cada descriptor, transformant-los a mitjana zero i desviació estàndard la unitat. Matemàticament:

$$\mathbf{x}'_i = \frac{\mathbf{x}_i - \overline{\mathbf{x}_i}}{\sigma(\mathbf{x}_i)}. \quad (7)$$

on $\mathbf{x}_i$ és el vector columna original de descriptors, $\overline{\mathbf{x}_i}$ n'és la mitjana aritmètica, $\sigma(\mathbf{x}_i)$ és la desviació estàndard mostral i $\mathbf{x}'_i$ és el vector transformat resultant. Gràficament, aquest procés es pot representar com es mostra a la figura 8:

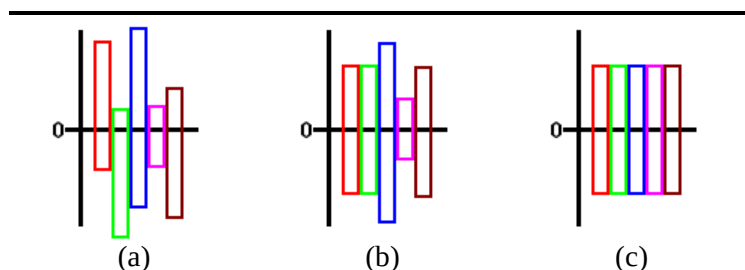


Figura 8.- Efecte de l'autoescalat sobre el conjunt de dades. (a) dades originals, (b) dades centrades a la seva mitjana, i (c) dades escalades finals.

4.4.2. Regressió multilinear

A l'hora d'establir una relació entre $\mathbf{X}(n \times m)$ i $\mathbf{y}(n)$, la regressió multilinear (MLR, de l'anglès *Multiple Linear Regression*) [62] era, fins fa poc, el mètode indiscutiblement triat. En la MLR, s'assumeix que $\mathbf{X}$ és de rang complert i que els x_{ij} estan mesurats amb error negligible. El model algebraic de la MLR es defineix a la següent equació:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}, \quad (8)$$

on $\mathbf{X} = [\mathbf{x}_0 | \mathbf{x}_1 | \mathbf{x}_2 | \dots | \mathbf{x}_m]$, $\mathbf{b}^T = [b_0, b_1, b_2, \dots, b_m]$ i $\mathbf{e}$ és el vector error. Val a dir que la primera columna de $\mathbf{X}$ està composta tan sols per constants, les quals després del centrat, esdevé zero. Quan $\mathbf{X}$ és de rang complert, la solució de mínims quadrats és:

$$\mathbf{b} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}. \quad (9)$$

Una de les desavantatges òbvies de la MLR en QSAR es dona quan el nombre de molècules s'aproxima o supera al de descriptors ($n \approx m$), ja que llavors $\mathbf{X}$ pot no ser de rang complert i $(\mathbf{X}^T \mathbf{X})^{-1}$ no està definida, i $\mathbf{b}$ no pot ser estimat. Aquest problema és conegut amb el nom de *colinearitat* o dependència lineal entre els descriptors [62]. La colinearitat pot ser detectada abans de dur a terme la regressió, podent-se evitar, suprimint descriptors, problemes en la resolució de (9) [62-68]. Tanmateix, la colinearitat fou la raó principal de la pèrdua de popularitat de la MLR dins del camp del QSAR.

4.4.3. Anàlisi de Composants Principals

En l'anàlisi de components principals (PCA, de l'anglès *Principal Component Analysis*) [68,69], els descriptors originals es substitueixen per components principals (PCs), els quals són combinacions lineals de les columnes de $\mathbf{X}$. L'extracció dels PCs a partir de $\mathbf{X}$, o descomposició de $\mathbf{X}$, s'expressa algebraicament com a:

$$\mathbf{X} = \mathbf{t}_1 \mathbf{p}_1^T + \mathbf{t}_2 \mathbf{p}_2^T + \dots + \mathbf{t}_k \mathbf{p}_k^T + \mathbf{e}. \quad (10)$$

Així, l'objectiu del PCA és simplement la descomposició de $\mathbf{X} = [\mathbf{x}_1 | \mathbf{x}_2 | \dots | \mathbf{x}_m]$ en k components vectors *score* ($\mathbf{T} = [\mathbf{t}_1 | \mathbf{t}_2 | \dots | \mathbf{t}_k]$) i vectors *loading* ($\mathbf{P} = [\mathbf{p}_1 | \mathbf{p}_2 | \dots | \mathbf{p}_k]$), on $k < m$. Prèviament al PCA, normalment s'autoescalen les columnes de $\mathbf{X}$.

Cada PC extret és ortogonal als PCs prèviament generats ($\mathbf{t}_i^T \mathbf{t}_j = 0$, per $i \neq j$) i descriu un percentatge decreixent de la variança en la matriu $\mathbf{X}$ original. Els vectors loading també són ortogonals ($\mathbf{p}_i^T \mathbf{p}_j = 0$), i són escalats a la unitat ($\mathbf{P}^T \mathbf{P} = \mathbf{I}_k$). Es pot demostrar també que els vectors $\mathbf{p}_l$ ($l \in (1, k)$) són vectors propis de $\mathbf{X}^T \mathbf{X}$ amb valors propis $\boldsymbol{\lambda} = [\lambda_1, \lambda_2, \dots, \lambda_k]$. Matemàticament es compleix que $\mathbf{X}^T \mathbf{X} \mathbf{p}_l = \mathbf{p}_l \lambda_l$. Els valors propis mostren quin percentatge de la variança explica cada PC. De forma similar, es pot mostrar com els scores $\mathbf{t}_l$ ($l \in (1, k)$) representen els vectors propis de $\mathbf{X} \mathbf{X}^T$, escalats a $\lambda_l^{0.5}$.

Els primers PCs es poden considerar com a una representació vàlida de $\mathbf{X}$ degut a que la variança residual s'assumeix com a *soroll de fons* insignificant. Els vectors loading obtinguts són importants per la comprensió dels descriptors de $\mathbf{X}$ rellevants i quins són redundants, i per a entendre els vectors scores, els quals informen sobre les semblances i diferències entre els objectes estudiats en el model.

A la pràctica, en el càlcul de PCA s'empra sovint l'algorisme NIPALS [64,70] quan es calcula component a component. La segona component es calcula a partir dels residuals restants de la primera component ($\mathbf{X}_1 = \mathbf{X}$):

$$\mathbf{X}_2 = \mathbf{X}_1 - \mathbf{t}_1 \mathbf{p}_1^T. \quad (11)$$

Com que les components scores ($\mathbf{T}$) són ortogonals i contenen la major part de la variança de $\mathbf{X}$, aquestes són adequades com a variables de regressió per a $\mathbf{y}$ emprant MLR. D'aquesta manera, s'introdueix el següent model:

$$\mathbf{y} = q_1 \mathbf{t}_1 + q_2 \mathbf{t}_2 + \dots + q_k \mathbf{t}_k + \mathbf{e}, \quad (12)$$

on les q_i són els coeficients de regressió que descriuen la relació entre les variables de resposta ($\mathbf{y}$) i les k components score ($\mathbf{T}$). Anàlogament a la MLR, la solució de mínims quadrats en l'estimació de $\mathbf{q}$ és:

$$\mathbf{q} = (\mathbf{T}^T \mathbf{T})^{-1} \mathbf{T}^T \mathbf{y}, \quad (13)$$

on $\mathbf{q}$ és l'estimació obtinguda pels coeficients de regressió. Si el vector $\mathbf{q}$ de l'anterior equació pogués ser expressat en coeficients de regressió, $\mathbf{b}$, com a l'equació (9), la interpretació del model es simplificaria. Els coeficients $\mathbf{b}$ poder ser estimats [64] com:

$$\mathbf{y} = \mathbf{X} \mathbf{b} = \mathbf{T} \mathbf{q}. \quad (14)$$

Substituint $\mathbf{T}$ per $\mathbf{X} \mathbf{P}$ (on $\mathbf{P}$ és la matriu que conté els k vectors loading), una possible solució és:

$$\mathbf{b} = \mathbf{P} \mathbf{q}, \quad (15)$$

però degut a la quasi singularitat de $\mathbf{X}^T\mathbf{X}$, la solució no és única.

Una senzilla il·lustració del PCA es troba a la següent figura 9, en la qual es mostra la primera composant principal d'un conjunt de dades repartides en dos dimensions.

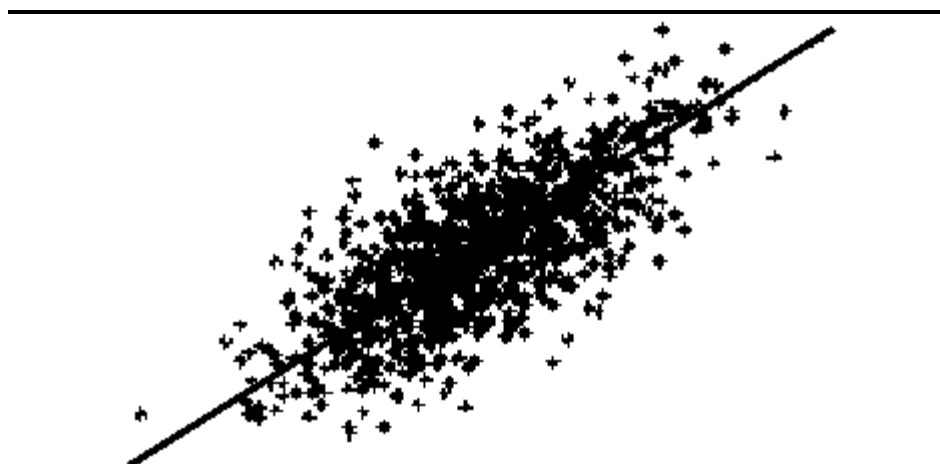


Figura 9.- PCA d'un núvol de dades bidimensionals. La línia mostra la direcció de la primera composant principal, la qual proporciona la òptima (en sentit de mínims quadrats) reducció de dues a una dimensió.

4.4.4. Mínims quadrats parcials

L'anàlisi de PCA descriu l'estructura latent de $\mathbf{X}$, la qual es pot utilitzar com a regressors de $\mathbf{y}$. En el mínims quadrats parcials (PLS, de l'anglès *Partial Least Squares*), la variable de resposta $\mathbf{y}$, s'inclou en el procés de descomposició i es busca un vector de loading, com és el vector dels pesos $\mathbf{w}_l$ ($l \in (1, k)$), que maximitzi $\mathbf{w}_l^T \mathbf{X}_{l-1}^T \mathbf{y}$ sota la restricció $\mathbf{w}_l^T \mathbf{w}_l = 1$ ($\mathbf{X}_{l-1}$ conté els residuals del càlcul de la composant anterior, tal i com s'explica a continuació).

Aquesta és la definició general de PLS [63-66], on *parcial* indica que la solució de mínims quadrats s'aplica quan $\mathbf{w}$ es determina tal i com s'ha definit, i no de forma general [71]. En la descripció que segueix es tractarà només el procés de descomposició quan hi ha només una variable de resposta (PLS1); tanmateix el PLS pot tractar simultàniament diverses variables de resposta⁵ (p.e. PLS2).

⁵ Un cas pràctic on es poden avaluar múltiples variables de resposta és el cas de la correlació de l'activitat biològica d'un fàrmac i de la seva toxicitat associada.

L'algorisme original de PLS [70,72] comença per estimar el vector dels pesos $\mathbf{w}_l$ per la component l -èsima com el vector que maximitza l'expressió $\mathbf{w}_l^T \mathbf{X}_{l-1}^T \mathbf{y}$:

$$\mathbf{w}_l = \frac{\mathbf{X}_{l-1}^T \mathbf{y}}{\|\mathbf{X}_{l-1}^T \mathbf{y}\|}. \quad (16)$$

El vector $\mathbf{t}$ es determina com:

$$\mathbf{t}_l = \frac{\mathbf{X}_{l-1} \mathbf{w}_l}{\mathbf{w}_l^T \mathbf{w}_l}, \quad (17)$$

i com que, per definició, $\mathbf{w}_l^T \mathbf{w}_l = 1$, l'anterior expressió es simplifica a $\mathbf{t}_l = \mathbf{X}_{l-1} \mathbf{w}_l$. El vector de loadings $\mathbf{p}_l$, necessari per al càlcul dels residuals del nou model, s'obté per regressió de $\mathbf{X}$ amb $\mathbf{t}_l$:

$$\mathbf{p}_l = \frac{\mathbf{X}_{l-1}^T \mathbf{t}_l}{\mathbf{t}_l^T \mathbf{t}_l}. \quad (18)$$

Per a que puguin ser possibles les estimacions de $\mathbf{y}$ a partir de $\mathbf{t}_l$, és necessari el coeficient de regressió q_l per a la component l -èsima, el qual es determina per regressió entre $\mathbf{y}$ i $\mathbf{t}_l$:

$$q_l = \frac{\mathbf{y}^T \mathbf{t}_l}{\mathbf{t}_l^T \mathbf{t}_l}. \quad (19)$$

Finalment, els nous residuals $\mathbf{X}_l$ es calculen per a treure l'efecte de la component prèvia:

$$\mathbf{X}_l = \mathbf{X}_{l-1} - \mathbf{t}_l \mathbf{p}_l^T. \quad (20)$$

De forma anàloga al PCA, els coeficients de regressió $\mathbf{b}_{\text{PLS}}$ són útils per a la interpretació del model, així com per a la realització de prediccions sobre compostos externs ($\mathbf{X}_{\text{TEST}}$) com $\mathbf{y} = \mathbf{X}_{\text{TEST}} \mathbf{b}_{\text{PLS}}$. Els coeficients $\mathbf{b}_{\text{PLS}}$ es calculen després de les k components com:

$$\mathbf{b}_{\text{PLS}} = \mathbf{W}(\mathbf{P}^T \mathbf{W})^{-1} \mathbf{q}. \quad (21)$$

on $\mathbf{W}$ és $(\mathbf{w}_1 | \mathbf{w}_2 | \dots | \mathbf{w}_k)$, $\mathbf{P} = (\mathbf{p}_1 | \mathbf{p}_2 | \dots | \mathbf{p}_k)$ és i $\mathbf{q}^T = (q_1, q_2, \dots, q_k)$.

Aquest algorisme també és anomenat l'algorisme del PLS ortogonalitzat, ja que els scores estimats i els vectors dels pesos són ortogonals ($\mathbf{t}_i^T \mathbf{t}_j = 1$ i $\mathbf{w}_i^T \mathbf{w}_j = 1$ on $i \neq j$). Si el nombre k de components extretes iguala al nombre de descriptors continguts en $\mathbf{X}$, llavors la solució PLS és idèntica a la solució donada per MLR.

El segon algorisme de PLS, presentat per Martens [64], difereix del de Wold en la no ortogonalitat dels scores i dels pesos. En conseqüència, els coeficients interns de regressió $\mathbf{q}$ no es

poden calcular component a component, sinó que totes les components han de ser calculades simultàniament:

$$\mathbf{q} = (\mathbf{T}^T \mathbf{T})^{-1} \mathbf{T}^T \mathbf{y}. \quad (22)$$

En l'algorisme no ortogonalitzat, els residuals es calculen restant $\mathbf{t}_l \mathbf{w}_l^T$ de la component prèvia, en oposició a l'equació (20) de l'algorisme de Wold. Tanmateix, les prediccions són idèntiques en ambdós algorismes, tot i que en l'algorisme de Martens, l'equació (21) es simplifica a:

$$\mathbf{b}_{\text{PLS}} = \mathbf{W} \mathbf{q}. \quad (23)$$

Un excel·lent tutorial sobre l'algorisme de PLS, el qual s'examinen totes les etapes del procés i permet una fàcil implementació, es troba a la referència 63.

4.4.5. Xarxes Neurs Artificial

Les xarxes neurals artificials [73-75] (ANN, de l'anglès *Artificial Neural Network*) és un sistema inspirat en el funcionament del cervell humà. Aquesta disciplina també és coneguda per altres noms, tals com procés de distribució paral·lela, neurocomputació, sistemes d'intel·ligència natural o algorismes d'aprenentatge. En qualsevol cas, l'objectiu és simular, a partir de maquinari especialitzat o de sofisticat programari, com són les múltiples capes dels simples elements de procés anomenats *neurones*. Cada neurona està unida a un nombre concret de neurones veïnes amb coeficients variables de connectivitat, els quals representen la força de tals connexions. L'aprenentatge s'aconsegueix ajustant aquests coeficients per a assolir que la xarxa proporcioni, com a sortida, els resultats apropiats.

Els components més bàsics de les neurones artificials estan modelats a partir de l'estructura del cervell. Tanmateix, algunes estructures en les ANN són bastant rares, si no inexistents, en el model biològic. Tot i això, i degut a les grans semblances, bona part de la nomenclatura s'ha manllevat de la neurobiologia.

La neurona biològica present en el cervell humà és un tipus de cèl·lula molt específica, la qual ens proporciona les habilitats de recordar, pensar i aplicar prèvies experiències a les nostres accions. Cadascuna d'aquestes neurones es pot arribar a connectar a 200000 més. La *força* i el *potencial* del cervell prové de la quantitat de neurones present, i de les múltiples connexions entre elles. Totes les neurones biològiques tenen 4 components principals: dendrites, soma, àxon i sinàpsis. Bàsicament, una neurona biològica rep informació d'altres, les combina d'alguna forma, practica una operació no lineal en el resultat, i llavors dona el resultat final.

Les neurones artificials simulen les quatre funcions bàsiques de les naturals. Les artificials són molt més simples, i la xarxa creada acostuma a tenir poques neurones i connexions al ser comparades amb la xarxa existent en un cervell.

El disseny d'una ANN pot arribar a ser llarg, degut a les possibles múltiples possibilitats, les quals es descarten normalment per prova i error. El disseny d'una ANN consisteix en:

- disposar les neurones en varies capes.
- decidir el tipus de connexions entre neurones de diferents capes, així com entre les neurones d'una mateixa capa.
- determinar com la neurona rep la informació, com la manipula (funció de pas) i com la treu.
- assignar les forces de les connexions de la xarxa permetent que la xarxa *aprengui* a partir de pesos de connexions a determinar per un conjunt de dades d'entrenament.

4.4.6. Altres metodologies

Tot i que no tractades en el període que abarca aquest treball, existeixen multitud de tècniques estadístiques⁶ apart de les exposades prèviament, presentant a continuació un recull parcial amb les seves característiques principals.

- Regressions no lineals [76-79]: addicionen termes elevats a diferents potències a fi d'augmentar la qualitat de l'ajust. N'hi ha de linearitzades, no linearitzades i polinomials.
- Descomposició PARAFAC [80,81]: semblant al PCA, on es manté l'estructura n -dimensional de les dades originals, descomposant-se la matriu (o tensor) original de dades en n vectors de loading.
- Regressió multilinear PLS [82]: extensió del model PLS2[64] en combinació amb la descomposició PARAFAC.
- Regressió multimodal de covariances principals [71,83]: combinació de PCA sobre la matriu dels descriptors amb regressions simultànies en les variables de resposta.

⁶ Recordant una vella cita de Bohr: 'Si el problema que estàs estudiant necessita estadística, canvia de problema'.

4.5. Validació d'un model QSAR

Un cop s'ha construït el model cal determinar tant si és vàlid com si es comporta de manera robusta. A tal efecte, existeixen múltiples eines per a lavaluació dels models proposats. Aquestes eines es poden classificar en dues categories segons els tipus de dades que es manegen, existint la validació *interna* i l'*externa*.

4.5.1. Validació interna

La validació interna correspon a l'avaluació del model tenint en compte les pròpies dades emprades en la seva construcció, i sense l'addició de nous elements. La qualitat del model es pot avaluar internament segons els següents criteris:

- Qualitat de l'ajust.
- Probabilitat d'obtenir aleatòriament el mateix resultat.
- Obtenció, per validació creuada, de valors de la variable de resposta propers als originals.
- Test d'aleatorietat

La *qualitat de l'ajust* acostuma a avaluar-se mitjançant el coeficient de correlació r , el qual avalua les diferències quadràtiques entre els valors originals de la variable de resposta amb els obtinguts mitjançant el model proposat. El coeficient de correlació existeix en l'interval $[-1,1]$, i a mesura que aquestes diferències es redueixen (millor ajust per part del model), el valor absolut de r creix, i decreix en cas contrari. Matemàticament, r té la següent expressió:

$$r = \frac{n \sum_{i=1}^n y_i y_{i,c} - \sum_{i=1}^n y_i \sum_{i=1}^n y_{i,c}}{\sqrt{\left(n \sum_{i=1}^n y_i^2 - \left(\sum_{i=1}^n y_i \right)^2 \right) \left(n \sum_{i=1}^n y_{i,c}^2 - \left(\sum_{i=1}^n y_{i,c} \right)^2 \right)}}, \quad (24)$$

on n és el nombre d'objectes, y_i representa un valor de la variable de resposta original i $y_{i,c}$ un valor d'aquesta variable ajustat. Val a dir que el coeficient de correlació acostuma a donar-se com el seu valor elevat al quadrat (r^2).

Tot que pot portar a pensar que r (o r^2) dóna ja força informació sobre la validesa del model, l'experiència demostra gairebé el contrari. Si se suposa una sèrie de paràmetres linealment

independents $\mathbf{X}$ ($n \times n$) linealment independents a correlacionar amb les n $\mathbf{y}$, el valor de r serà sempre de +1, el qual suposa la resolució d'un sistema d' n equacions amb n incògnites (sistema compatible determinat), sense cap mena de validesa en el camp del QSAR. La negació d'aquesta afirmació portaria a l'afirmació de que qualsevol conjunt de n descriptors, linealment independents, és plenament vàlid per a la descripció d'un sistema molecular.

Perfilant més aquest últim punt, es veu que l'addició de nous paràmetres en les regressions comporta sempre un increment en el valor de r , a no ser que tal nou paràmetre sigui una constant o una combinació lineal de paràmetres previs, el qual no tindria cap efecte. Aquest increment en el valor de r , tal i com es veurà al parlar de la validació creuada, és beneficiós fins al punt en que l'addició de nous descriptors porta a la sobreparametrització (de l'anglès *overfitting*), és a dir, ja no s'està descrivint a la variable original de resposta, sinó parametritzant el soroll de fons.

Un procediment, val a dir molt enginyós, de comprovar si s'han emprat massa paràmetres en la correlació és calcular la *probabilitat de obtenir atzarosament el mateix resultat*. Una possible manera de dur a terme aquesta tasca és per *força bruta*, és a dir, simular computacionalment milers de correlacions aleatòries i mirar quin percentatge dóna una correlació similar a l'obtingut amb els descriptors originals. Tanmateix, s'ha publicat recentment un mètode que proporciona directament la probabilitat d'obtenir a l'atzar un coeficient de correlació [84], donats un nombre de punts i de paràmetres emprats. Si després d'un ajust s'observa que la probabilitat d'obtenir a l'atzar el mateix resultat és alta, cal rebutjar el model obtingut.

A part d'un bon ajust, es desitja també una bona capacitat de predicció en el model proposat, en altres paraules, que sigui capaç de proporcionar estimacions vàlides, en un marge d'error acceptable, de molècules no incloses en la generació del model. Un mètode àmpliament estès per a determinar la capacitat de predicció d'un model quan no es disposa de dades externes és la *validació creuada* [85] (de l'anglès *cross-validation*). Aquest procediment consisteix en extreure un nombre de compostos del conjunt original, construir el model amb les dades restants, i emprar el conjunt prèviament extret com a test per a predir-ne la seva propietat. El procés es repeteix tantes vegades com calgui fins a que totes les molècules tinguin $\binom{n-1}{m-1}$ prediccions, on n és el nombre de molècules del conjunt, i m el nombre de molècules que s'extreuen. En el cas de $m=1$, llavors cada molècula serà predita 1 cop, per $m=2$ serà predita $n-1$ vegades, i així progressivament.

La *profunditat* de l'estudi de validació creuada es fixa segons el nombre de molècules que s'extreuen. Així, si s'extreu una molècula cada cop es té el *Leave-One-Out* (LOO, o deixar-ne un

fora), dues porta al *Leave-Two-Out* (LTO, o dos fora), o en general *Leave-many-Out* (LmO, o vàries fora). El procés més emprat és el LOO, tot i que la potència computacional ja permet actualment la realització pràctica de validacions creuades de profunditats més grans. Recentment ha aparegut un treball de Besalú [86], on es tracta la fàcil i pràctica realització de LmO sense haver de recalculer els models en MLR, podent dedicar el temps computacional estalviat a un estudi estadístic dels resultats de predicció, o permetent la búsqueda de descriptors òptims per resultats de predicció a profunditats més elevades de validació creuada.

A partir dels resultats d'una validació creuada, es poden obtenir dos coeficients que determinen quant bona és la capacitat de predicció del model: el coeficient de predicció q^2 i el coeficient de correlació de validació creuada r_{cv} .

El coeficient de predicció avalua la capacitat de predicció a partir dels residuals de predicció. L'expressió matemàtica de q^2 és:

$$q^2 = \frac{SSY - PRESS}{SSY}, \quad (25)$$

on SSY és la suma dels quadrats de la resposta biològica (de l'anglès *Sum of Squares of Y*) i

s'expressa com a $SSY = \sum_{i=1}^n (y_i - \bar{y})^2$, on $\bar{y}$ és la mitjana de les activitats; i el $PRESS$ és la suma dels quadrats dels errors de predicció (de l'anglès *PRediction Error Sum of Squares*), expressat matemàticament com $PRESS = \sum_{i=1}^n (y_i - y_{i,p})^2$. El valor de q^2 està definit en l'interval $(-\infty, 1]$,

podent-se donar tres situacions possibles:

- $q^2=0$: denota capacitat de predicció nul·la. El fet de proposar la mitjana de les activitats, o en general una constant arbitrària, com a prediccions portaria al mateix resultat.
- $q^2<0$: denota una capacitat de predicció nefasta. De fet, proposar una constant arbitrària com a prediccions portaria a millors resultats. Acostuma a ser indicació de sobreparametrització.
- $q^2>0$: denota capacitat de predicció per part del model, la qual, a més propera a 1, més alta. Segons diversos autors, un valor de q^2 superior a 0.5-0.6 és pot considerar ja satisfactori.

El coeficient de correlació de validació creuada r_{cv} correspon a analitzar la correlació entre les variables de resposta originals i predites, de igual manera que es fa l'equació (24), canviant les variables ajustades per les predites. L'interval de r_{cv} està definit igualment en $[-1,1]$, indicant millor capacitat de predicció quant més proper s'està a $+1$. Cal anar en compte si s'eleva r_{cv} al quadrat, ja que es perd el rastre del signe original, i pot donar-se el cas que les prediccions inverteixin la tendència de les respostes

(objectes amb activitat alta siguin predits amb activitats baixes, i viceversa), el qual donaria un valor de r_{cv} en valor absolut alt, però negatiu, que al ser elevat al quadrat, seria igualment elevat, però contradictori. Una forma de veure si la tendència és la correcte és observar en l'equació de regressió entre les y i les y_p (idealment de pendent 1 i ordenada a

l'origen 0) el signe del pendent de la recta, el qual es correspon amb el de r_{cv} .

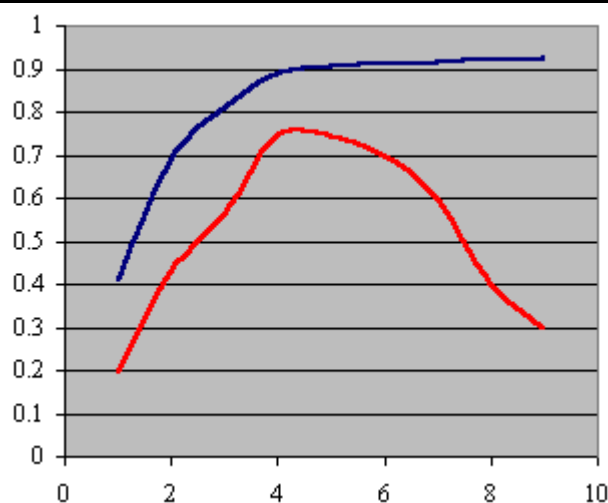


Figura 10.- Evolució de r^2 (en blau) i q^2 (en vermell) segons el nombre de descriptors emprats. Dades extretes de la referència 87 i arxius.

Els valors d'aquests dos últims paràmetres pot servir per a decidir la quantitat de descriptors a emprar en el model òptim. Mentre que el coeficient d'ajust r augmenta amb l'increment progressiu del nombre de paràmetres, r_{cv} i q^2 no necessàriament ho fan, sinó que acostumen a augmentar fins a la inclusió de cert nombre de descriptors, assoleixen un màxim i a partir d'aquí, el seu valor disminueix, indicant que ja s'ha sobreparametritzat el model per inclusió de massa descriptors. La figura 10 mostra gràficament la progressió normal d'aquests paràmetres.

Tal i com es veu a l'anterior figura, fins a l'ús de 5 descriptors tant la qualitat de l'ajust com la capacitat de predicció augmenten, tanmateix a partir d'aquí, l'addició de nous descriptors empobreix notablement la capacitat de predicció, mentre que l'ajust millora de forma lenta. En el cas concret de l'estudi mostrat [87], es considerarà que l'ús de 4 descriptors era òptim, donant la màxima capacitat de predicció al model.

Un últim indicador de la qualitat del model és la desviació estàndard entre els valors originals de la propietat i els ajustats o els predits. La desviació estàndard és una suma ponderada de les diferències quadràtiques entre ambdues magnituds, i matemàticament s'expressa com:

$$\sigma = \left(\frac{\sum_{i=1}^n (y_{i,c} - y)^2}{n} \right)^{1/2}, \quad (26)$$

on $y_{i,c}$ pot canviar-se per $y_{i,p}$ per a referir-se als valors predits enlloc dels calculats per l'ajust. A més baixa la σ^7 , menors són els errors en els ajustos (o les prediccions), o dit d'altra manera, millor és el model.

Finalment, s'acostuma a realitzar el *test d'aleatorietat* (de l'anglès *Random Test*), el qual consisteix en permutar a l'atzar les activitats del vector de resposta un cert nombre de vegades, realitzar el càlcul d'ajust i de capacitat de predicció amb aquest nou vector reordenat i monitoritzar els valors de r (o r^2) i q^2 (o r^2_{cv}). Un model correcte ha de donar una clara separació entre els valors d'ajust i predicció originals i els valors obtinguts del random test, d'altra manera, s'indica que existeix una sobreparametrització o que la globalitat del protocol emprat en la construcció del model és capaç de correlacionar qualsevol conjunt de dades.

Cal diferenciar entre el test d'obtenir el mateix ajust a l'atzar del random test. El primer tracta la probabilitat atzarosa de obtenir el mateix resultat, és a dir, tenir sort. El segon, apart d'incloure aquesta possibilitat, inclou al propi mètode de correlació, és a dir, que el mètode no sigui forçat a trobar sempre una correlació entre dos conjunts de dades, i que existeixi una clara relació entre els descriptors moleculars i les propietats estudiades. El fet de no superar un random test és raó suficient per a rebutjar el model obtingut⁸.

Els random tests es poden dur a terme conceptualment de dues maneres diferents:

- Amb total llibertat al mètode per a escollir i manipular les dades.
- Restringint part de la informació, per exemple, que els descriptors a utilitzar en la regressió siguin aquells que proporcionen un model òptim. Aquesta manera de fer dona millors resultats en el random test, però li resta bona part de la credibilitat, ja que si

⁷ Tot i que tenen el mateix símbol, no cal confondre la desviació estàndard amb la sigma de Hammett.

⁸ Una nota de precaució s'ha de donar en aquest aspecte. Si força valors permutats donen valors de r^2 i q^2 elevats, llavors cal rebutjar el model; però si són pocs aquests punts, cal comprovar que no corresponguin a permutacions que tan sols hagin bescanviat les activitats de molècules amb propietat i descriptors semblants, les quals portarien a resultats similars als originals. Una manera d'evitar això és mirar la correlació entre el vector permutat i l'original, i cas que hi hagi correlació significativa, tornar a permutar sense continuar els càlculs per al vector permutat.

entren en el joc aquells descriptors òptims per a l'activitat estudiada, difícilment correlacionaran amb la propietat reordenada. Tanmateix, aquesta manera és la preferida en la comunitat del QSAR.

La següent figura mostra tres exemples de random test.

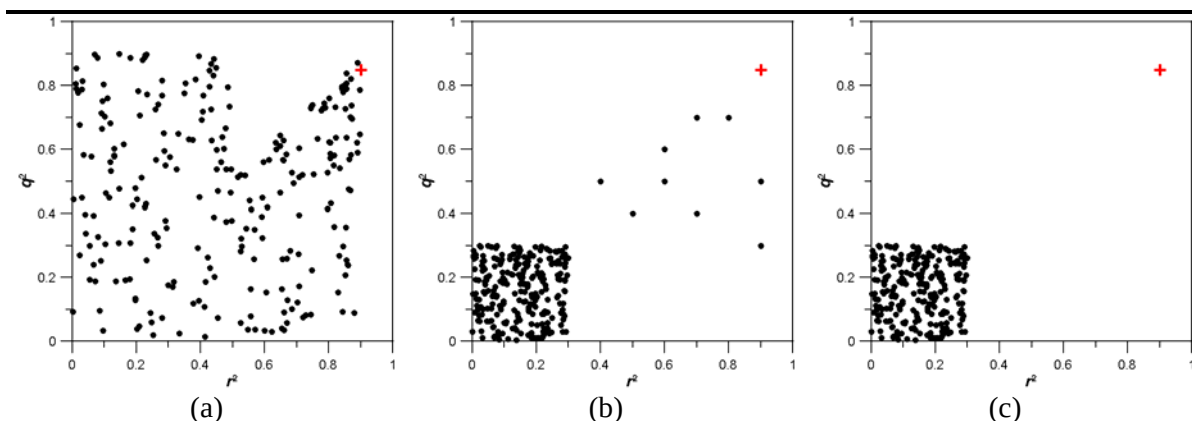


Figura 11.- Exemples esquemàtics de random-test, on + són els valors de r^2 i q^2 originals, i • els valors obtinguts per permutacions dels valors de l'activitat de resposta. (a) random test fallit, (b) random-test amb dades atzaroses properes i (c) random-test superat satisfactòriament. (exemples generats artificialment)

Tal i com es veu a la figura 11, una clara separació entre els punts originals i els separats existeix quan el random test es superat amb èxit. Tanmateix, quan el test no es supera, els punts s'acosten, i en alguns casos poden arribar a superar, als valors originals, i en casos d'aquest estil, el model obtingut no és vàlid. En el cas intermig de la figura 11, on el gruix dels punts no dona models satisfactoris, però alguns d'ells s'acosten als valors originals, s'hauria de examinar manualment les dades per veure a quines permutacions corresponen, i si seria o no lícit treure-les.

4.5.2. Validació externa

La prova definitiva de la validesa d'un model, i en última instància de la capacitat de predicció de la que disposa, és el resultat que proporciona amb un conjunt d'objectes extern, el qual no hagi sigut inclòs en cap etapa del procés de construcció del model. L'aplicació del model generat sobre els descriptors d'aquest conjunt d'objectes donarà unes prediccions sobre aquests compostos, i si es disposa de dades d'activitat, es podrà determinar la seva validesa.

Els paràmetres que quantifiquen la qualitat de predicció sobre conjunt extern poden ser els mateixos que qualifiquen internament la predicció (q^2 , r_{cv} , σ de predicció), substituint els valors *realment* predits allà on calgui, sent el més utilitzat la suma dels errors quadràtics de predicció *SSPE* (de l'anglès *Sum of Squares Prediction Errors*):

$$SSPE = \sum_i^{n_t} (y_{p,t} - y_t)^2 . \quad (27)$$

on n_t és el nombre de molècules del conjunt extern de test, $y_{p,t}$ és la predicció i y_t és el valor real de la variable de resposta.

Un dels problemes principals d'emprar aquest mètode de validació és la disponibilitat de prou dades per a fer dos conjunts significatius. Si el conjunt que s'analitza no és molt gran, distribuir-lo en dos pot resultar estadísticament inviable al no poder construir un conjunt d'entrenament que contingui un nombre o quantitat de característiques suficient. Un possible alternativa per a simular un conjunt d'entrenament consisteix en el mètode ITS (de l'anglès *Internal Test Set*) [88], en el qual s'extreu una molècula del conjunt, considerada com a test, i la amb resta de molècules es construeix un model de predicció. Posteriorment, la molècula extreta es prediu amb els descriptors òptims obtinguts amb les molècules restants. El procés es repeteix amb totes les molècules. En el sentit metodològic, aquest procediment és un Leave-One-Out, tanmateix difereix en la total absència d'intervenció de la pròpia molècula extreta en el càlcul, obtenint-se autèntiques prediccions .

4.6. *Tria de Descriptors en els models QSAR*

Ja per acabar aquesta introducció a la construcció de models QSAR, fer un incís en la selecció de descriptors. Aquest aspecte, a l'igual que en altres apartats, és força ampli i sovint depenent de la tècnica estadística que s'utilitza en conjunció o posteriorment.

La tria de descriptors és una reducció prèvia de la informació sobre les molècules a estudiar, a fi d'alleugerir la tasca computacional i optimitzar-ne els resultats, ja que l'ús de **X** sencera pot resultar una tasca molt feixuga, sobretot en mètodes basats en camps moleculars, on cada molècula pot tenir associada milers de descriptors.

A continuació es detallen dos dels mètodes emprats durant el període que abarca la realització d'aquest treball: el NSS i el MPVM.

Els *Símbols de Sumes Aniuades* [89-92] (de l'anglès *Nested Summation Symbols*) són un protocol computacional per a definir bucles de profunditat variable en temps d'execució sense haver de canviar el codi font d'un programa. Aquest procediment permet obtenir, donat un conjunt de m

descriptors sobre el qual se'n volen emprar k , les $\binom{m}{k}$ combinacions. A cada tirada, es monitoritzen els valors estadístics d'interès, quedant-se amb el model que doni els millors resultats. Tot i que aquest mètode és conceptualment molt senzill, l'explosió combinatòria que genera pot arribar-lo a fer computacionalment molt costós a k s relativament petites.

El *Most Predictive Variable Method* [93] (MPVM o mètode de les variables més predictives) és un protocol per a determinar els millors descriptors d'acord amb la seva correlació individual amb la variable de resposta. Tanmateix, es demostra que el mètode és equivalent a, per a cada descriptor, realitzar una regressió amb la variable de resposta, i posteriorment, la matriu de descriptors s'ordena d'acord amb la magnitud del coeficient de correlació obtingut. Llavors, els models QSAR es construïran emprant els descriptors ordenats, prenent els k requerits segons el model [94]. També basat en la correlació per parelles, el mètode CORCHOP [95] elimina aquells descriptors, els quals es correlacionen sobre un llindar pre-establert. D'aquest mètode existeixen diferents variants [96-98], incloent també eliminació per relacions no-lineals [99-101].

Altres mètodes de tria de descriptors es basen en la combinació conjunta d'una tècnica estadística d'ajust, com la MLR [102], anàlisis discriminants [103], PLS [104-106], i més recentment amb algorismes evolutius i genètics específics, com els k -veïns més propers [107] o les xarxes neurals [108,109].

Tot i que estrictament fan el contrari, hi ha mètodes d'eliminació de descriptors redundants o que aporten poca informació, que també ajuden a triar els descriptors òptims. Aquest mètodes es basen en l'eliminació de descriptors que són combinacions lineals d'altres, aquells que tenen una coeficient de variació petit o que, en el PCA, la variança associada és baixa.

4.7. Referències

1. Citat a: Borman, S. *New QSAR techniques eyed for environmental assesments*. Chem. Eng. News, 68, 1990, pp 20-23.
2. Crum-Brown, A. and Fraser, T.R. *On the connection between chemical constitution and physiological action. Part 1. On the physiological action of the ammonium bases, derived from strychia, brucia, thebaia, codeia, morphia and nicotia*. Trans. Royal Soc. Edinburgh, 25, 1868-9, pp 257-274.
3. Richet, C. *Compt. Rend. Soc. Biol. (Paris)*, 45, 1893, pp 775-776.
4. Meyer, H. *Theorie der Alkoholnarkose, welche Eigenschaft dir Anästhetica bedingt ihre narkotische Wirkung*. Arch. Exp. Pathol. Pharmacol., 42, 1899, pp 109-118.
5. Overton, E. *Studien über dir Narkose*. Gustav Fischer, Jena, 1901.
6. Ferguson, J. *The use of chemical potentials as an indicator of toxicity*. Proc. Royal Soc. London, B127, 1939, pp 387.
7. Hammett, L.P. *The effect of structures upon the reactions of organic compounds. Benzene derivatives*. J. Am. Chem. Soc., 59, 1937, pp 96.
8. Hammett, L.P. *Physical organic chemistry*. McGraw-Hill, New York, 1940.
9. Hansch, C., Leo, A. and Taft, R.W. *A survey of Hammett substituent constants and ressonance and field parameters*. Chem. Rev., 91, 1991, pp 165-195.
10. Hansch, C., Leo, A. and Hökman, D. *Exploring QSAR – hydrophobic, electronic and steric constants*. American Chemical Society, Washington D.C., 1995.
11. Seydel, J.K. *Prediction of in vitro activity of sulfonamides, using Hammett constants or spectrophotometric data of the basic amines for calculation*. Mol. Pharmacol., 2, 1966, pp 259-265.
12. Hansch, C. *Drug research of the luck of the draw*. J. Chem. Ed., 51, 1969, pp 360-365.
13. Taft, R.W. J. Am. Chem. Soc., 74, 1952, pp 3120-3128.
14. Swain, C.G., Lupton Jr., E.C. *Field and ressonance components of substituent effects*. J. Am. Chem. Soc., 90, 1968, pp 4328-4335.
15. Bruice, T.C., Kharasch, N., Winzler, R.J. Arch. Biochem. Biophys., 62, 1956, pp 305.
16. Zahradník, R., Chvapil, M. *Experientia*, 16, 1960, pp 511-512.
17. Zahradník, R. *Experientia*, 18, 1962, pp 534-536.
18. Zahradník, R. *Arch. Int. Pharmacodyn.*, 125, 1962, pp 311-329.
19. Chvapil, M., Zahradník, R., Cmuchalova, B. *Arch. Int. Pharmacodyn.*, 125, 1962, pp 330.
20. Hansen, E.R. *Acta Chem. Escand.*, 16, 1962, pp 1593.
21. Free, S.M. and Wilson, J.W. *A mathematical contribution to structure-activity studies*. J. Med. Chem., 7, 1964, pp 395-399.
22. Bodek, K., Kopecký, J., Krivucova, M. and Vlachova, D. *Chemical structure and biological activity of p-disubstituted derivatives of benzene*. *Experientia*, 20, 1964, pp 667-678.
23. Kopecký, J., Bocek, K. and Vlachova, D. *Nature*, 207, 1965, pp 667-678.
24. Fujita, T. and Ban, T. *Structure-activuty studies of phenylethylamines as substrates of biosynthetic enzymes of sympathetic transmmitters*. J. Med. Chem., 14, 1971, pp 148-152.
25. Hansch, C. *A quantitative approach to biochemical structure-activity relationships*. *Acct. Chem. Res.*, 2, 1969, pp 232-239.
26. Fujita, T., Iwasa, J. and Hansch, C. *A new substituent constant, π , derived from partition coefficients*. J. Am. Chem. Soc., 86, 1964, pp 5175-5180.
27. Hansch, C. and Fujita, T. *ρ - σ - π analysis. A method for the correlation of biological activity and chemical structure*. J. Am. Chem. Soc., 86, 1964, pp 1616-1626.
28. Bondi, A. *van der Waals Volumes and Radii*. J. Phys. Chem., 68, 1964, pp 441-451.

29. Pearlman, R.S. *Physical chemical properties of drugs*. Vol. 10, Yalkowsky, S.H., Sikula, A.A. and Valvani, S.C. (eds.), Dekker, New York, 1980.
30. Pearlman, R.S. *Quantum Chem. Prog. Exchange Bull.*, 1, 1981, pp 15.
31. Pearman, R.S. *Partition coefficient determination and estimation*. Dunn, W.J., Block, J.H. and Pearman, R.S. (eds.), Pergamon, New York, 1986.
32. Camilleri, P., Watts, A. and Boraston, J.A. *J. Chem. Soc. Parkin. Trans.*, 2, 1988, pp 1699.
33. Lee, B. and Richards, F.M. *The interpretation of protein structures: Estimation of static accessibility*. *J. Mol. Biol.*, 55, 1971, pp 379-400.
34. Pearman, R.S. SAREA. Quantum Chemistry Program Exchange, University of Indiana, Bloomington, IN, Program Number 432.
35. Grigoras, S. *J. Comput. Chem.*, 11, 1990, pp 493.
36. Howel, J., Rossi, A., Wallace, D., Hiraki, K. and Hoffmann, R. *FORTICON 8*. Quantum Chemistry Program Exchange, University of Indiana, Bloomington, IN, Program Number 469.
37. Ramsden, C.A. (ed.) *Comprehensive medicinal chemistry*. Vol. 4, Pergamon, Oxford, 1990.
38. Kubiny, H. *QSAR: Hansch analysis and related approaches*. VCH, Weinheim, 1993.
39. Hansch, C. and Leo, A. *Exploring QSAR. Fundamentals and applications in chemistry and biology*. American chemical society, Wasington D.C., 1995.
40. Martin, Y.C. *Quantitative drug design: a critical introduction*. Marcel Dekker, New York, 1978.
41. Taft, R.W. *Steric effects in organic chemistry*. Newman, M.S. (ed.), Wiley, New York, 1956.
42. Verloop, A., Hoogenstraaten, W. and Tipker, J. *Development and Application of New Steric Substituent Parameters in Drug Design*. In *Drug Design*. Vol. VII. Ariens, E.J. (ed.), Academic Press, New York, 1976, pp 165-207.
43. Wiener, H. *Structural determination of paraffin boiling points*. *J. Chem. Phys.*, 69, 1947, pp 17-20.
44. Kier, L.B., Hall, L.H., Murray, W.J., Randic, M. *Molecular connectivity. I. Relationship tot nonspecific local anaesthesia*. *J. Pharm. Sci.*, 64, 1975, pp 1971-1974.
45. Randic, M. *On the characterization of molecular branching*. *J. Am. Chem. Soc.*, 97, 1975, pp 6609-6615.
46. Trinajstic, N. *Chemical graph theory*. CRC Press, Boca Raton, 1992.
47. Hosoya, H. *A Newly proposed quantity Characterizing the topological nature of structural isomers of saturated hydrocarbons*. *Bull. Chem. Soc. Jpn.*, 44, 1971, pp 2332-2339.
48. Kier, L.B. and Hall, L.H. *Molecular connectivity in chemistry and drug research*. Academic, New York, 1976.
49. Kier, L.B. and Hall, L.H. *Molecular connectivity in structure-activity analysis*. Research studies press, Letchworth, 1986.
50. Brown, R.D. and Martin, Y.C. *Use of Structure-Activity Data To Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection*. *J. Chem. Inf. Comput. Sci.*, 36, 1996, pp 572-584.
51. Brown, R.D. and Martin, Y.C. *The Information Content of 2D and 3D Structural Descriptors Relevant to Ligand-Receptor Binding*. *J. Chem. Inf. Comput. Sci.*, 37, 1997, pp 1-9.
52. Doman, T.N., Cibulskis, J.M., Cibulskis, M.J., McCray, P.D. and Spangler, D.P. *Algorithm5: A Technique for Fuzzy Similarity Clustering of Chemical Inventories*. *J. Chem. Inf. Comput. Sci.*, 36, 1996, pp 1195-1204.
53. Klopman, G., Balthasar, D.M. and Rosenkranz, H.S. *Application of the computer-automated structure evaluation (CASE) program to the study of structure-biodegradation relationships of miscellaneous chemicals*. *Environ. Toxicol. Chem.*, 12, 1993, pp 231-240.
54. Rosenkraz, H.S. and Klopman, G. *Relationships between electronegativity and genotoxicity*. *Mutation Res.*, 328, 1995, pp 215-227.

55. Crammer III, R.D., Paterson, D.E. and Bunce, J.D. *Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins.* J. Am. Chem. Soc., 110, 1988, pp 5959-5967.
56. Klebe, G., Abraham, U. and Mietzner, T. *Molecular similarity indices in comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity.* J. Med. Chem., 37, 1994, pp 4130-4146.
57. Silverman, B.D. and Platt, D.E. *Comparative molecular moments analysis (CoMMA): 3D-QSAR without molecular superposition.* J. Med. Chem., 39, 1996, pp 2129-2140.
58. Jain, A.N., Koile, K. and Chapman, D. *Compass: predicting biological activities from molecular surface properties. Performance comparisons on a steroid benchmark.* J. Med. Chem., 37, 1994, pp 2315-2327.
59. Bravi, G., Gancia, E., Mascagni, P., Pegna, M., Todeschini, R. and Zaliani, A. *MS-WHIM, new 3D theoretical descriptors derived from molecular surface properties: a comparative 3D QSAR study in a series of steroids.* J. Comput.-Aided Mol. Des., 11, 1997, pp 79-92.
60. Kellogg, G.E., Kier, L.B., Gaillard, P. and Hall, L.H. *E-State fields: applications to 3D QSAR.* J. Comput.-Aided Mol. Des., 10, 1996, pp 513-520.
61. Chen, H., Zhou, J. and Xie, G. *PARM: a genetic evolved algorithm to predict bioactivity.* J. Chem. Inf. Comput. Sci., 38, 1998, pp 243-250.
62. Myers, R.H. *Classical and modern regression with applications.* PWS-KENT Publishing company, Boston, Massasuchetts, 1990.
63. Geladi, P. and Kowalsky, B.R. *Partial Least Squares: a tutorial.* Anal. Chem. Acta, 185, 1986, pp 1-17.
64. Martens, H. and Næs, T. *Multivariate calibration.* John Wiley & Sons, New York, 1989.
65. Wold, S., Sjöström, M. and Eriksson, L. *Partial least squares projections to latent structures (PLS) in chemistry.* In *Encyclopedia of Computational Chemistry.* Schleyer, P.R., Allinger, N.L., Clark, T., Gasteiger, J., Kollman, P.A., Schaefer III, H.F. and Screiner, P.R. (eds.). John Wiley and Sons Ltd., Chichester, 1994, Vol. 4, 2006-2021.
66. Wold, S., Johansson, E. and Cocchi, M. *PLS-Partial Least-Squares Projections to Latent Structures.* In *3D QSAR in Drug Design.* Kubinyi, H. (ed). ESCOM, Leiden, 1993, Chapter 5.
67. Besley, D.A., Kuh, E. and Welsch, R.E. *Regression diagnostics.* Wiley, New York, 1980.
68. Jolliffe, I.T. *Principal Component Analysis.* Springer-Verlag, New York, 1986.
69. Kendall, M. *Multivariate Analysis.* Charles Griffin&Co., London, 1975.
70. Wold, H. *Soft modeling by latent variable, the non-linear iterative partial least squares (NIPALS) algorithm.* In *Perspectives in probability and statistics.* Gani, J. (ed.), Academic Press, London, 1975, pp 117-142.
71. Smilde, A.K. Comments on multilinear PLS. J. of Chemometrics, 11, 1997, pp 367-377.
72. Wold, S., Ruhe, A., Wold, H. and Dunn III, W.J. *The collinearity problem in linear regression. The partial least squares ({PLS}) approach to generalized inverses.* J. Sci. Stat. Comput., 3, 1984, pp 735-743.
73. Novič, M., Nikolovska-Coleska, Z. and Solmajer, T. *Quantitative structure-activity relationship of flavonoid p56 protein tyrosine kinase inhibitors. A neural network approach.* J. Chem. Inf. Comput. Sci., 37, 1997, pp 990-998.
74. Tetko, I.V., Alessandro, E., Villa, P. and Livingstone, D.J. *Neural network studies. 2. Variable selection.* J. Chem. Inf. Comput. Sci., 36, 1996, pp 794-803.
75. Duprat, A.F., Huynh, T. and Dreyfus, G. *Toward a principled methodology for neural network design and performance evaluation in QSAR. Application to the prediction of log P.* J. Chem. Inf. Comput. Sci., 38, 1998, pp 586-594.
76. Bard, Y. *Non-linear parameter estimation.* Academic Press, New York, 1974.
77. Bukietyński (ed.) *Mathematical programming.* Akademia Ekonomiczna, Wrocław, 1976 (en polonès).

78. Tabak, D. and Kuo, B.C. *Optimal control by mathematical programming*. Prentice-Hall Inc., Englewood Cliffs, N.J., 1971.
79. Wilde, D.J. *Optimum seeking methods*. Prentice-Hall Inc., Englewood Cliffs, N.J., 1965.
80. Carroll, J. and Chang, J.J. *Analysis of individual differences in multidimensional scaling with an n -way generalization of the Eckart-Young decomposition*. *Psychometrika*, 35, 1970, pp 283-319.
81. Harshman, R.A. *Foundations of the PARAFAC procedure: models and conditions for an explanatory multimodal factor analysis*. UCLA Working Papers on Phonetics, 16, 1970, pp 1-84.
82. Bro, R. *Multiway calibration. Multilinear PLS*. *J. of Chemometrics*, 10, 1996, pp 47-61.
83. De Jong, S. and Kiers, H.A.L. *Principal covariates regression Part I: theory*. *Chemom. and Intell. Lab. Syst.*, 14, 1992, pp 155-164.
84. Article proba ponec
85. Allen, D. M. *The relationship between variable selection and data augmentation and a method for prediction*. *Technometrics*, 16, 1974, pp 125-127.
86. Besalú, E. *Fast computation of cross-validated properties in full linear leave-many-out procedures*. IT-IQC-01-02.
87. Robert, D., Gironés, X. and Carbó-Dorca, R. *Quantification of the influence of single-point mutations on haloalkane dehalogenase activity: a molecular quantum similarity study*. *J. Chem. Inf. Comput. Sci.*, 40, 2000, pp 839-846.
88. Besalú, E., and Vera, L. *Internal Test Sets (ITS) Method: A new cross-validation technique to assess the predictive capability of QSAR models*. *SAR QSAR Environ. Res.*, in press.
89. Carbó, R. and Besalú E. *Nested Summation Symbols and Perturbation Theory*. *J. Math. Chem.*, 13, 1993, pp 331-342.
90. Carbó, R. and Besalú, E. *Definition, Mathematical Examples and Quantum Chemical Applications of Nested Summation Symbols and Logical Kronecker Deltas*. *Comput. Chem.*, 18(2), 1994, pp 117-126.
91. Carbó, R. and Besalú, E. *Definition and Quantum Chemical Applications of Nested Summation Symbols and Logical Functions: Pedagogical Artificial Intelligence Devices for Formulae Writing, Sequential Programming and Parallel Implementation*. *J. Math. Chem.*, 18, 1995, pp 37-72.
92. Carbó, R. and Besalú, E. *Application of Nested Summation Symbols to Quantum Chemistry: Formalism and Programming Techniques*. In *Strategies and Applications in Quantum Chemistry: from Astrophysics to Molecular Engineering*. DeFranceschi, M. and Ellinger, Y. (eds.). Kluwer Academic Publishers, Amsterdam, 1996, pp. 229-248.
93. Cuadras, C. M., and Arenas, C. *A distance based regression model for prediction with mixed data*. *Commun. Stat. Theor. Method.*, 19, 1990, pp 2261-2279.
94. Besalú, E., and Vera, L. *On the optimal selection of principal components in QSPR studies*. *J. Math. Chem.*, 1, 2001, pp 21-34.
95. Livingstone, D.J., and Rahr, E. *Corchop – An interactive routine for the dimension reduction of large QSAR data sets*. *Quant. Struct.-Act- Relat.*, 8, 1989, pp 103-108.
96. Kikuchi, O. *Systematic QSAR procedures with quantum chemical descriptors*. *Quant. Struct.-Act. Relat.*, 6, 1987, pp 179-184.
97. Gute, B.D., and Basak, S.C. *Use of topostructural, topochemical, and geometric parameters in the prediction of vapor pressure: a hierarchical QSAR approach*. *SAR QSAR Environ. Res.*, 7, 1997, pp 117-131.
98. Stanton, D.T. *Evaluation and use of BCUT descriptors in QSAR and QSPR studies*. *J. Chem. Inf. Comput. Sci.*, 39, 1996, pp 11-20.
99. Tetko, I.V., Villa, A.E.P., and Livingstone, D.J. *Neural networks studies. 2. Variable selection*. *J. Chem. Inf. Comput. Sci.*, 36, 1996, pp 794-803.

100. Kovalishyn, V.V., Tetko, I.V., Luik, A.I., Kholodovych, V.V., Villa, A.E.P., and Livingstone, D.J. *Neural networks studies. 3. Variable selection in the cascade-correlation learning architecture*. J. Chem. Inf. Comput. Sci., 38, 1998, pp 651-659.
101. Manallak, D.T., and Livingstone, D.J. *Neural networks in drug discovery: have they lived up to their promise?* Eur. J. Med. Chem., 34, 1999, pp 195-208.
102. Shi, L.M., Fan, Y., Myers, T.G., O'Connor, P.M., Paull, K.D., Friend, S.H., and Weinstein, J.N. *Mining the NCI anticancer drug discovery databases: genetic function approximation to QSAR study of anticancer ellipticine analogues*. J. Chem. Inf. Comput. Sci., 39, 1998, pp 189-199.
103. Partek Pro 2000. Partek Incorporated, *Partek analysis and recognition technologies*, 1993-1999.
104. Hasegawa, K., Miyashita, Y., and Funatsu, K. *GA strategy for variable selection in QSAR studies: GA-based PLS analysis of calcium channel antagonists*. J. Chem. Inf. Comput. Sci., 37, 1997, pp 306-310.
105. Hasegawa, K., Miyashita, Y., and Funatsu, K. *GA strategy for variable selection in QSAR studies: GA-PLS and D-optimal designs for predictive QSAR model*. J. Mol. Struct. (Theochem), 425, 1998, pp 255-262.
106. Cho, S.J., Zheng, W., and Tropsha, A. *Rational combinatorial library design. 2. Rational design of targeted combinatorial peptide libraries using chemical similarity probe and the inverse QSAR approaches*. J. Chem. Inf. Comput. Sci., 38, 1998, pp 259-268.
107. Zheng, W., and Tropsha, A. *Novel variable selection quantitative structure-property relationships approach based on the k-nearest neighbor principle*. J. Chem. Inf. Comput. Sci., 40, 2000, pp 185-194.
108. So, S.S., and Karplus, M. *Genetic neural networks for quantitative structure-property relationships: improvement and application of benzodiazepine affinity for benzodiazepine/GABA A receptor*. J. Med. Chem., 39, 1996, pp 5246-5256.
109. Wessel, M.D., Jurs, P.C., Tolan, J.W., and Muskal, S.M. *Prediction of human intestinal absorption of drug compounds from molecular structure*. J. Chem. Inf. Comput. Sci., 38, 1998, pp 726-735.

5 Algorisme de Sobreposició Topo-Geomètrica (TGSA)

- **Resum:** En aquest capítol es dona un repàs a diferents tècniques de sobreposició molecular, i en particular es tracta un nou mètode d'alineament estructural desenvolupat per a disposar les estructures en base al nombre màxim de subestructures conjuntes comunes a ambdues. Fonamentat sobre característiques topològiques i geomètriques, aquest mètode es troba implementat en un programa informàtic (TGSA99), explicat detalladament en l'article adjunt *J. Comp. Chem.*, 22, 2001, 255–263, i en la nova revisió (TGSA2001) explicada en aquest capítol.
- **Paraules Clau:** Sobreposició, alineament, subestructura comuna, TGSA, MQSM.

5.1 Introducció

La sobreposició tridimensional és una tasca abstracta¹, la qual implica el posicionament d'un objecte en relació a un altre. Les tècniques computacionals que sobreposen molècules serveixen com a punt focal per a la interpretació i comprensió de dades moleculars. La sobreposició juga un paper força important en la correlació de gran varietat de fonts d'informació, des de l'energètica, fisicoquímica, biològica o farmacocinètica.

L'energia d'un sistema està dictada pel balanç entre les forces nuclears i electròniques, la solvatació i el reconeixement molecular per part d'un ent biològic, així com és dependent de la conformació i de diverses interaccions subtils que la molècula exerceix amb el seu entorn. Cadascuna d'aquestes propietats està directament lligada a l'ordenació espacial de la molècula, llavors es pot suposar que si àtoms de característiques semblants es posicionen a les mateixes ubicacions, s'esperarà que aquests exerceixin un efecte resultant similar.

Les diverses tècniques en que la sobreposició s'utilitza per a la interpretació de dades varien com les mateixes dades a interpretar. Pels mètodes que descriuen les dades en termes de característiques moleculars, àtoms o fragments, grups funcionals o atributs físics, la sobreposició n'esdevé el component comú. D'altres tècniques, com l'anàlisi comparatiu de camps moleculars [1] o aquelles basades en la semblança tridimensional [2-4], estan gairebé totes basades en l'alineament molecular. Les semblances es poden emprar en disseny de compostos, en la seva optimització i en la seva identificació en noves classes.

El conjunt de sobreposicions resultant per a una sèrie molecular acostuma a generalitzar-se en forma de *farmacòfor*, o dit d'aquella regió molecular comuna a tota la sèrie a la qual, i degut a la seves característiques d'interacció, s'atribueix l'activitat de les molècules.

Tanmateix, la sobreposició molecular no és tan sols un idea emprada en tècniques QSAR o derivades, també s'utilitza en altres camps. L'alineament és molt emprat en cristal·lografia de raigs X, recerca estructural en proteïnes o en la quantificació de la distorsió ocasionada en un esquelet comú per l'efecte d'un substituent [5,6]. Una altra branca que fa un ús intensiu de la sobreposició molecular és el camp de les bases de dades moleculars [7-19], el qual, gràcies a diferents avenços computacionals, permet fer recerques en conjunts que compten els membres de la base per milers o milions; d'aquí que aquest camp demani un mètode de sobreposició, no només efectiu, sinó ràpid i eficient.

Tot i que conceptualment la noció de sobreposició molecular és senzilla, i fins i tot, donades dues molècules relativament semblants, evident, l'aplicació computacional pràctica s'enfronta al problema combinatori: el nombre de maneres en que dues molècules es poden sobreposar pot esdevenir molt gran, convertint el problema en quelcom inherentment, des d'un punt de vista computacional, complex. Si al mateix temps es té en compte la possibilitat de diferents conformacions a les molècules, el problema esdevé encara més complicat, incrementant-se la complexitat a mesura que augmenta la mida i nombre d'enllaços flexibles de les molècules a estudiar.

Conseqüentment, la majoria de les tècniques que han estat utilitzades amb èxit, redueixen el problema, bé via dirigint la búsqueda per procediments heurístics (selecció a l'atzar de diferents alineaments) o bé per medis de mostreig molecular.

¹ S'entén abstracta per 'no realista', en el sentit que una molècula no pot físicament col·locar-se en el mateix espai que una altra.

Les tècniques de sobreposició molecular es poden dividir en dues classes majoritàries: aquelles basades en àtoms o fragments d'una molècula a alinear amb els corresponents de l'altra (*atom-based*, de l'anglès basada en els àtoms), i aquelles en que s'utilitzen camps moleculars, volums i/o superfícies (*field-based*, de l'anglès basada en els camps).

5.2 Metodologies de sobreposició

5.2.1 Aproximacions manuals

Quan s'estudien sèries moleculars relativament homòlogues, es poden obtenir sobreposicions raonables variant els enllaços d'una molècula fent manualment rotacions d'aquests, quan no són rígids, i translacions de les estructures en una terminal gràfica. Per exemple, en la figura 1 es mostren dos antagonistes de l'angiotensina:

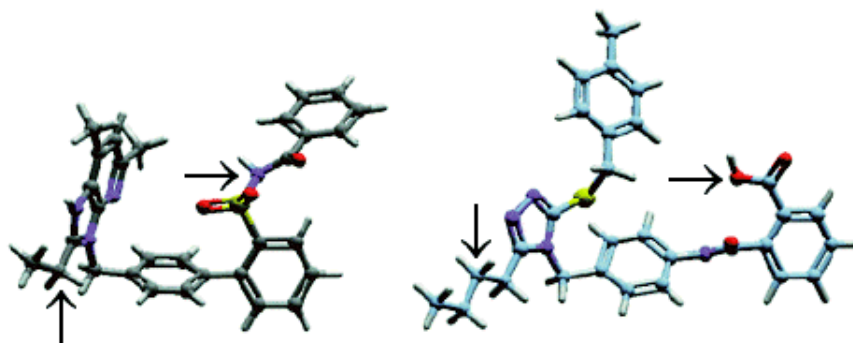


Figura 1.- Inhibidors de l'angiotensina: la nevirapina i la α -APA. Els àtoms senyalats ($\leftarrow$) corresponen al farmacòfor, tal i com descriu Prendergast [20]. Les estructures foren minimitzades amb el camp de forces MMFF [21] del programa Batchmin [22].

Les molècules foren sobreposades per rotacions dels enllaços flexibles, minimitzant la distància entre els àtoms farmacofòrics, senyalats a la figura 1. Les estructures es re-minimitzaren amb el mateix camp de forces i sobreposades per un procediment de mínims quadrats. El resultat de la sobreposició es mostra a la figura 2:

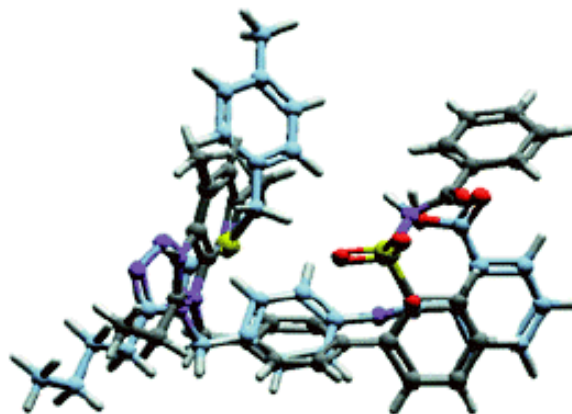


Figura 2.- Resultat de la sobreposició manual de dos inhibidors de l'angiotensina.

Tanmateix, aquestes aproximacions manuals estan força limitades. És força difícil per a molts químics trobar les sobreposicions alternatives o jutjar-validar objectivament la qualitat de l'alineament obtingut. Per a assistir en aquests mètodes interactius, s'han desenvolupat tècniques de retroalimentació (de l'anglès *feedback*) i classificació [23], les quals faciliten l'establiment d'una mesura objectiva de la bondat de l'alineament. Altres formes d'automatització adrecen la sobreposició atòmica explícitament per la incorporació de camps de forces. La distància entre els grups que coincideixen a les dues molècules s'assumeix idealment com a zero, i les desviacions existents s'expressen en forma d'una funció de penalització, la qual es minimitza conjuntament amb la valència i els termes no enllaçats [24].

5.2.2 Algorismes genètics

Una altra aproximació al problema, la qual emprava tècniques evolutives, està constituïda pels algorismes genètics [25]. Aquests procediments necessiten una funció d'ajust, la qual avalua la qualitat d'una població donada i s'utilitza per a seleccionar o construir la participació de la següent població. Degut a la natura probabilística d'aquestes tècniques, les solucions obtingudes acostumen a ser bones, tot i que no es pot assegurar que cap resultat sigui l'òptim, requerint-se múltiples càlculs amb punts inicials diferents. A la pràctica, en l'ús d'algorismes genètics en la sobreposició molecular, això no representa una dificultat, excepte que la flexibilitat molecular i la mida del conjunt molecular a estudiar pot limitar l'estudi que es pot realitzar en termes raonables de temps.

5.2.3 Tècniques atom-based

Diferents tècniques força eficients han estat desenvolupades per a examinar totes les parelles atòmiques dins d'una molècula. Si la distància entre qualsevol àtom, o vèrtex, es considera en relació amb la resta de centres, llavors es té el que es coneix com a *graph de màxima connectivitat*, tal i com es mostra a la figura 3 per a 2 molècules:

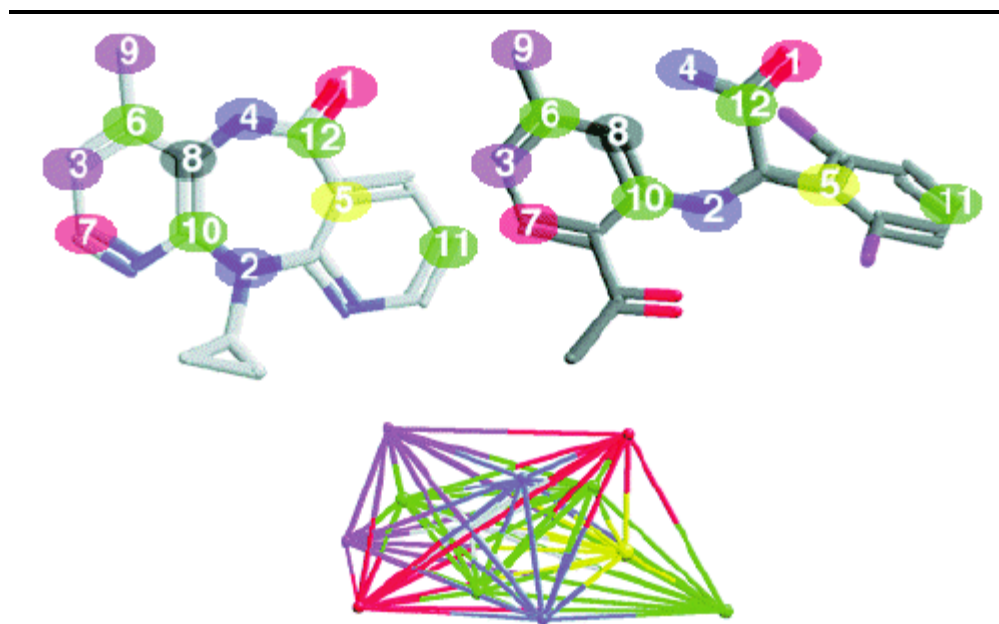


Figura 3.- Cliqué per a la nevirapina i la α -APA [25]. La tolerància de les arestes és de 0.5 Å. Les correspondències atòmiques estan indicades amb colors coincidents.

Cada vèrtex fa $n-1$ connexions amb la resta de vèrtexs, on n és el nombre de correspondències entre dues molècules. Aquest mètode ha estat implementat satisfactòriament en diferents programes, tals com Disco o SQ [26-29]. Per a assistir en la sobreposició, una possibilitat és la projecció de propietats físiques en els vèrtexs i arestes dels graphs enumerats. Així, l'alineament esdevé la identificació de les correspondències entre els graphs, sient la màxima correspondència coneguda com a *cliqué* (figura 3). Aquest conjunt de correspondències es pot convertir en sobreposició utilitzant un ajust per mínims quadrats dels graphs.

Les correspondències basades en els àtoms poden ser insuficients per a descriure les dades a estudiar. Per a compensar aquesta mancança, les descripcions atòmiques es poden corregir amb la inclusió de funcionalitzacions o subestructures que representen el comportament d'un grup sencer,

com per exemple, un anell aromàtic pot representar un grup hidrofòbic, o un carbonil pot emular la interacció direccional de parells solitaris.

5.2.4 Tècniques field-based

Tot i que les millores descrites per als mètodes atom-based, quan les molècules analitzades són estructuralment diverses, la seva utilitat decreix inevitablement al no trobar correspondències. Com a alternativa, s'han desenvolupat mètodes que analitzen camps moleculars, volums o superfícies. La mètrica més emprada avui en dia [30] és la semblança molecular quàntica (descrita amb profunditat al capítol 6), definida per Carbó [31], com a:

$$Z_{AB}(\Omega, \theta) = \iint \rho_A(\mathbf{r}_1) \Omega(\mathbf{r}_1, \mathbf{r}_2) \rho_B(\mathbf{r}_2, \theta) d\mathbf{r}_1 d\mathbf{r}_2, \quad (1)$$

on A i B són les molècules estudiades, $\rho_A(\mathbf{r}_1)$ i $\rho_B(\mathbf{r}_2)$ són les respectives funcions densitat, $\Omega(\mathbf{r}_1, \mathbf{r}_2)$ és un operador bielectrònic definit positiu, θ representa la transformació de les coordenades de B respecte les de A , ja que la mesura és depenent de la posició relativa d'ambdues molècules a l'espai, i Z_{AB} és la mesura de semblança quàntica resultant.

Originalment definida per a densitats electròniques, la semblança s'ha generalitzat per a poder avaluar la similitud de molècules a partir de qualsevol tipus de camp. Per a facilitar el procés de càlcul de la semblança, el qual pot resultar molt costós segons la formulació seguida per a descriure el camp, es poden emprar expansions centrades en àtoms per a aproximar la densitat electrònica, el potencial electroestàtic, la super-

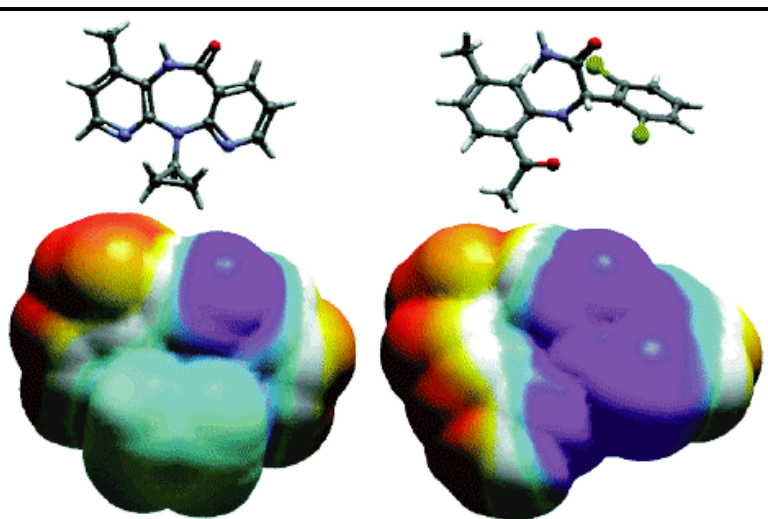


Figura 4.- Mapejat de la densitat electrònica sobre la superfície de van der Waals per a les molècules de nevirapina i α -APA. Les regions vermelles i grogues corresponen a zones deficientes en electrons, mentre que les blavoses corresponen a zones riques.

fície estèrica o el volum, tot i que també s'han emprat representacions basades en xarxes de punts (de l'anglès *grids*), tal i com es mostra a la figura 4 [32-36], donant la sobreposició, en base al recobriment de zones coincidents, que es mostra a la figura 5:

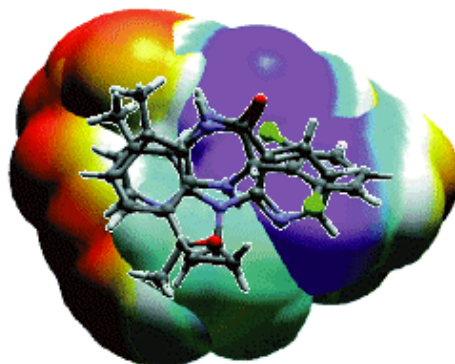


Figura 5.- Sobreposició de la nevirapina i la α -APA en base al recobriment de zones coincidents.

Aquestes tècniques, tot i que funcionen bé, són significativament més costoses en termes computacionals. Mentre que aquestes tècniques són insensibles a canvis en l'estructura atòmica i en la seva posició, poden tenir dificultats en la diferenciació entre regions contingudes en la molècula. Aquesta consideració és important, ja que sovint tan sols un fragment molecular és el responsable de l'activitat biològica a estudiar. Per a evitar aquestes possibles inconsistències, la metodologia s'ha estès a la consideració de semblances locals, les quals proporcionen solucions alternatives [37].

5.2.5 Altres tècniques

Existeixen altres tècniques de sobreposició més aproximades. Tot i que les seves representacions de les propietats moleculars no disposen directament sobre les superfícies o volums moleculars, s'ha demostrat que es comporten correctament i que computacionalment esdevenen més eficients. Alguns exemples són:

- *Tècniques basades en la simetria:* El mètode SPERM representa les propietats moleculars en un conjunt fix de coordenades basades en un icosaèdre tasselat (reconstruït a base de triangles), veure figura 6 (vèrtexs vermells), i en el centre de les cares d'aquest (figura 6, vèrtexs grisos) [38]. Tenint en compte el volum molecular i aprofitant la simetria de l'icosaèdre, es permet fer un alineament de forma ràpida,

minimitzant les diferències entre les propietats assignades als vèrtexs per a ambdues molècules. La principal avantatge d'aquesta tècnica és la velocitat. La semblança molecular calculada amb aquest mètode és suficientment ràpida per a avaluar grans conjunts moleculars, i a més a més, les semblances calculades amb aquesta tècnica permeten una clara classificació, permetent el seu ús en recerques en bases de dades estructurals. La tècnica tampoc està lliure de limitacions, les quals recauen en les aproximacions, com que exemple l'escala de l'icosàedre a triar, i que a mesura que augmenta la mida de la molècula, la discretització de les propietats en els vèrtex aporta menys informació.

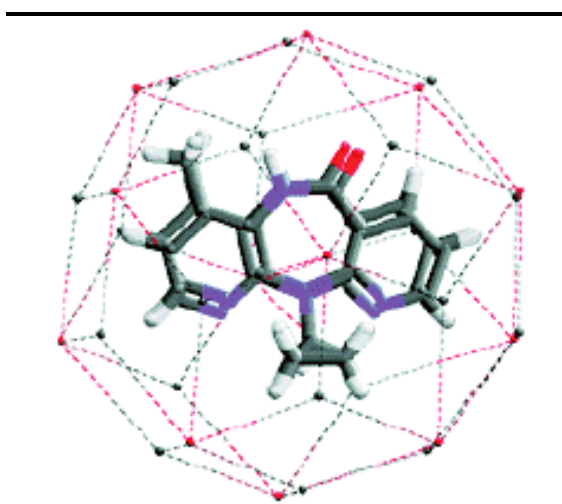


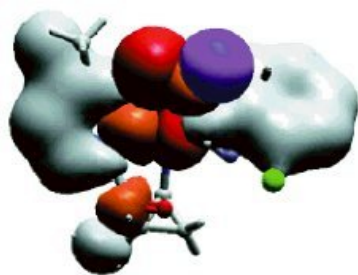
Figura 6.- Exemple de isosahedre tasselat per a la molècula de nevirapina. Els punts vermells en representen els vèrtexs i els grisos els centres de la cara.

- *Pells moleculars*: moltes de les interaccions importants pel reconeixement o resposta biològica són de curt abast, els quals es manifesten majoritàriament a la superfície molecular. Per a representar aquest concepte, es va desenvolupar la noció de *pell molecular* [39], definida com el volum obtingut de la resta d'una superfície de radi menor d'una de major, basant llavors la sobreposició en la optimització del recobriment dels volums de les *pells*. Aquest mètode té una particular avantatge sobre els mètodes de volum o de camps, la qual és la seva eficiència en molècules de mides diferents. Un exemple de pell molecular per a la molècula de nevirapina es mostra a la figura 7.

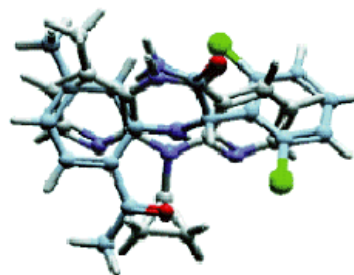


Figura 7.- Exemple de pell molecular per a la molècula de nevirapina. El gruix de la pell molecular, diferència entre els radis extern i intern, és de 0.25Å.

- *Tècniques de recobriment regional:* altres mètodes que funcionen bé amb molècules de diferent mida representen les propietats moleculars a partir del recobriment de funcions decaients. Programes com SEAL o SQ utilitzen funcions gaussianes per a descriure aquests solapament [25,40]. De manera diferent a l'ús de funcions gaussianes per a descriure el volum, aquestes tècniques utilitzen les gaussianes per a calcular el recobriment regional, ponderant les contribucions d'una regió dins la molècula. Ambdós efectes estèric i electrònic es tracten d'aquesta manera. El caràcter electrònic d'una molècula es caracteritza per set contribucions: catiònica, aniònica, pont d'hidrogen-donora, pont d'hidrogen-acceptora, polar (incloent donora i acceptora), hidrofòbica i de polarització. A la figura 8a es mostra el recobriment regional per a les molècules de nevirapina i α -APA. La sobreposició resultant es mostra a la següent figura 8b. Com a resultat, amb aquestes tècniques es poden obtenir diferents sobreposicions, segons el nombre de contribucions existents o emprades en els càlculs, afavorint-se les contribucions hidrofòbiques o electròniques.



(a)



(b)

Figura 8.- Sobreposició de recobriments regionals per a les molècules de nevirapina i α -APA obtingudes amb el programa SQ [28,29,40]. (a) regions sobreposades: àrees hidrofòbiques (blanc), regions donores d'enllaç d'hidrogen (vermell), regions acceptores d'enllaç d'hidrogen (blau) i regions de polarització (taronja). (b) sobreposició final de les estructures.

5.3 Sobreposició topo-geomètrica

5.3.1 Introducció

En el nostre grup de recerca, dedicat majoritàriament al desenvolupament de les tècniques en Mesures de Semblança Quàntica Molecular, així com diferents aspectes de implementació, la sobreposició molecular ha estat tractada extensament, ja que, tal i com s’ha comentat anteriorment a l’apartat 5.2.4, aquestes mesures són dependents de les posicions relatives de les estructures estudiades.

Inicialment, en els primers treballs on ja intervenien molècules poliatòmiques, les estructures se sobreposaven, com a primera aproximació, en base a l’alineació dels moments dipolars [41]. Tanmateix aquest procediment tenia diversos punts negatius² (només cal pensar en la sobreposició del benzè i el metà, ambdós amb $\mu = 0$).

En el següent procediment es considerarà l’alineació en base a la pròpia mesura, sobreposant-se les molècules en aquella posició en que la mesura avaluada esdevenia màxima [32]. A tal efecte, el procés seguit efectuava una buscada exhaustiva entre diverses disposicions, avaluant la mesura de semblança (amb el cost computacional que se’n deriva) a cada pas. Aquest procediment dóna un criteri de sobreposició rigorós entre dues molècules, amb l’avantatge de poder obtenir diversos alineaments alternatius en base a emprar diferents operadors en l’equació (1). D’aquesta manera, donades dues distribucions electròniques i un operador, la sobreposició esdevé fixada inequívocament en aquella que maximitza la integral de semblança.

Tot i donar un criteri rigorós de sobreposició, la regla de la màxima semblança té l’inconvenient de la pèrdua de sentit químic en aquells casos on les molècules implicades contenen àtoms pesants, com per exemple el Brom o el Iode. Aquests elements constitueixen centres de densitat (de l’anglès *density wells*), causant a l’algorisme la sobreposició en base a aquests substituents, prescindint de al resta d’un possible esquelet comú compostat per múltiples àtoms lleugers. A títol d’exemple, la següent figura 9 mostra dos casos de sobreposició basat en la màxima semblança:

² Cal dir que en aquest treball s’especifica que aquesta metodologia és vàlida tan sols en molècules polars.

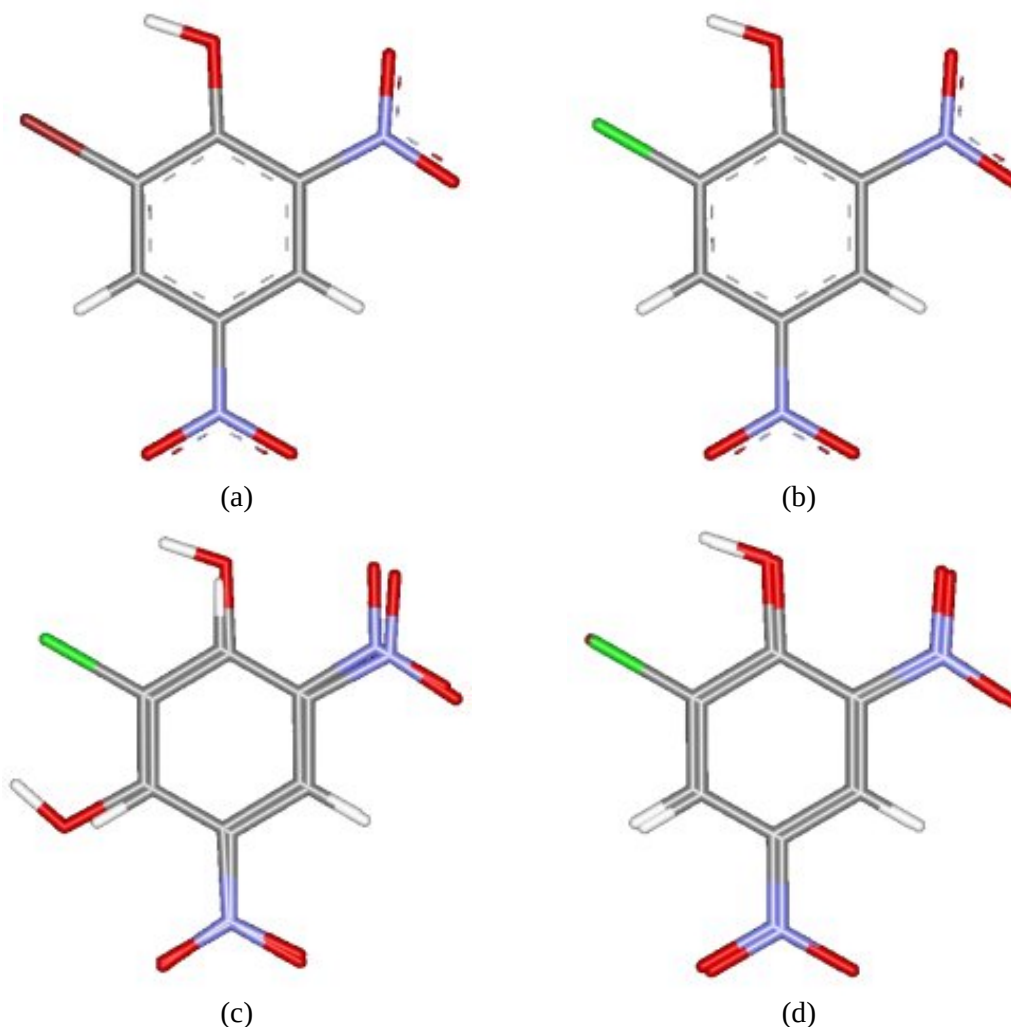
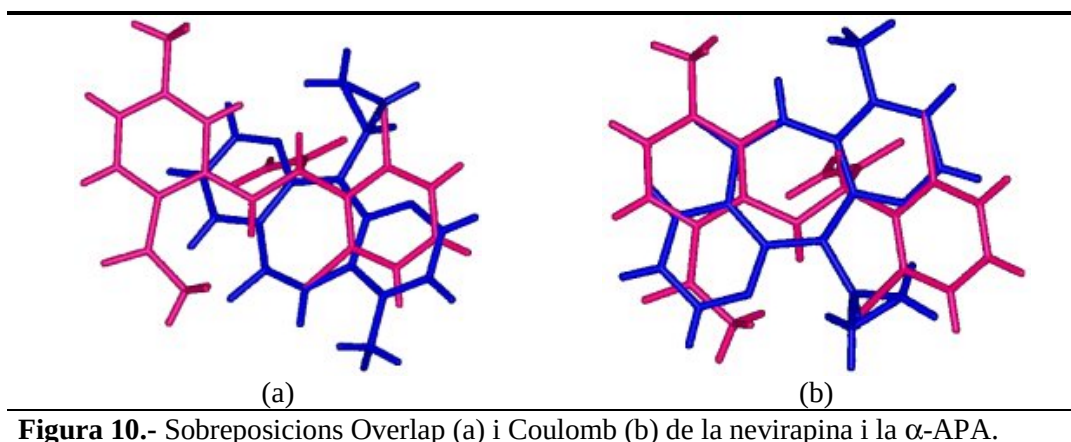


Figura 9.- Exemples de sobreposició basats en la regla de la màxima semblança. (a) i (b): molècules de test (1-bromo,4,6-dinitrofenol i 1-cloro,4,6-dinitrofenol), (c): resultat de la sobreposició basada en l'operador *Overlap*, (d): resultat de la sobreposició basada en l'operador *Coulomb*.

Tal i com es pot veure a l'anterior figura 9c, la sobreposició *Overlap* disposa en primer lloc els àtoms pesats (Clor i Brom, $Z_{\text{Cl-Br}} \approx 3000$, en front a per exemple $Z_{\text{C-C}} \approx 30$, $Z_{\text{N-N}} \approx 50$ o $Z_{\text{O-O}} \approx 80$) i seguidament considera que la sobreposició de dos grups nitro assoleixen la màxima semblança, tot i que no estan en les mateixes posicions (grup 4-nitro de (a) amb el grup 6-nitro de (b)). La sobreposició basada en *Coulomb* ja dona una disposició més intuïtiva, tot i que s'observa clarament un desplaçament de la subestructura comuna per a compensar la diferència entre les distàncies d'enllaç C-Br i C-Cl, els quals es sobreposen exactament.

En el cas de les molècules exposades en els exemples anteriors, la nevirapina i la α -APA, la sobreposició que s'obté es mostra a la figura 10, on s'observa una simple tendència a sobreposar àtoms el més pesats possibles.



A partir d'aquí, es va proposar la creació d'un nou mètode d'alineament molecular senzill i amb pocs requeriments de maquinari, basat només en distàncies i tipus atòmics, i destinat a trobar la màxima subestructura comuna entre dues molècules. Aquest nou procés s'anomenà 'Algorisme de Sobreposició Topo-Geomètrica' (TGSA, de l'anglès *Topo-Geometrical Superposition Algorithm*) [42], ja que com es veurà en els següents apartats, s'inclou la topologia molecular en forma de comparacions entre distàncies d'enllaç.

A continuació s'exposarà detalladament el mètode³ seguit per aquest algorisme per a sobreposar les estructures, des de la lectura dels fitxers que codifiquen la geometria molecular, fins a l'alineament final. A fi de clarificar el procés, cada etapa rellevant serà il·lustrada amb les dues molècules anteriorment sobreposades per la regla de la màxima semblança. Així mateix, es presentaran diferents exemples, conjuntament amb diferents paràmetres derivats del càlcul.

5.3.2 Metodologia

El primer pas, previ a la rutina de sobreposició, és la lectura de les coordenades moleculars i dels nombres atòmics. A tal efecte, s'han codificat diverses rutines destinades a llegir diferents formats, tals com les sortides de programes comercials, com l'Ampac [43] o Gaussian (tant la sortida estàndard com la FChk) [44], o típiques d'intercanvi de dades, com el format MDL Mol o l'XYZ [45]. Un cop en memòria, les coordenades moleculars són reordenades en base a nombre

³ El procediment exposat en aquest capítol difereix lleugerament del presentat en l'article original degut a diverses millores derivades de l'aplicació pràctica i extensions del mètode. Aquesta nova versió es troba implementada en el programa TGSA2001.

atòmic decreixent, a fi de determinar de forma senzilla en nombre d'àtoms no hidrogen de la molècula, els quals no són considerats en la resta del procés per abaratir els requeriments de càlcul, i fer córrer els diferents bucles només pels àtoms pesats, evitant així l'ús d'operadors condicionals.

Suposant dos conjunts de dades moleculars en memòria ($A=[a_1, a_2, \dots, a_n]$ i $B=[b_1, b_2, \dots, b_m]$), l'algorisme comprèn les següents etapes:

- posar l'àtom a_i al centre de coordenades
- posar l'àtom b_i al centre de coordenades
 - posar un àtom a_j , el qual està unit a a_i , a l'eix positiu de les Z
 - posar un àtom b_j , el qual està unit a b_i , a l'eix positiu de les Z
 - si la distància a_i-a_j és comparable a b_i-b_j :
 - posar un àtom a_k , el qual està unit a a_i i/o a_j , al pla YZ
 - posar un àtom b_k , el qual està unit a b_i i/o b_j , al pla YZ
 - si les distàncies a_i-a_k i a_j-a_k són comparables a b_i-b_k i b_j-b_k
 - avaluació del nombre d'àtoms del mateix tipus sobreposats i índex d'ajust (de l'anglès *fit*)

Aquest procés es repeteix per a tots els àtoms presents a les molècules, i aquella alineació que maximitzi el nombre d'àtoms sobreposats minimitzant l'índex de fit serà la triada per l'algorisme.

Determinar si dues distàncies interatòmiques són comparables es calcula en base al valor absolut de la seva diferència. Si aquesta és igual o inferior a un llindar fixat per l'usuari (per defecte 0.10 a.u.) aquesta es considera comparable.

El nombre d'àtoms sobreposats també disposa d'un llindar de tolerància (també definible externament i fixat per defecte a 0.30 a.u.), ja que òbviament és improbable aconseguir sobreposicions *computacionalment exactes* de diversos àtoms.

L'índex de fit es calcula a partir del quocient entre el producte de les sumes de distàncies interatòmiques a A i B, i la suma de distàncies intermoleculars entre A i B. Així, l'índex de fit, C_{AB} , s'expressa com a suma de mòduls al quadrat:

$$d_{AB} = \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} |\mathbf{x}_{i,A} - \mathbf{x}_{j,B}|^2, \quad (2)$$

on n són els nombres d'àtoms i les $\mathbf{x}$ les coordenades moleculars (en a.u.). Llavors, C_{AB} es construeix com a:

$$C_{AB} = \left(\frac{(d_{AA}d_{BB})^{1/2}}{d_{AB}} \right)^{1/2}. \quad (3)$$

D'aquesta manera C_{AB} avalua la qualitat de la sobreposició amb un índex a l'interval [0,1] passant compte de les subestructures no sobreposades i de les distorsions moleculars. Donant-se un millor alineament a mesura C_{AB} que s'apropa a 1, on el cas ideal de $C_{AB}=1$ indicaria que les molècules estudiades són idèntiques. D'altra banda, a mesura que l'índex s'allunya més de la unitat, més dolenta és la sobreposició obtinguda.

5.3.3 Exemple il·lustrat: el programa TGSA2001

En aquest apartat es descriurà el funcionament pràctic de l'algorisme implementat al programa TGSA2001, il·lustrat amb captures de pantalla de la interfície gràfica del programa i la seva aplicació a les dues molècules emprades per l'exemple de la regla de la màxima semblança: 2-bromo,4,6-dinitrofenol i 2-cloro,4,6-dinitrofenol.

En iniciar el programa TGSA2001, es mostra la següent pantalla:

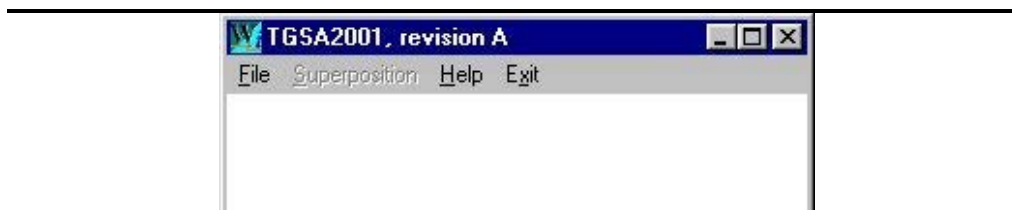


Figura 11.- Pantalla principal del programa TGSA2001.

En executar-se, el programa dóna tres opcions possibles: *File*, per a obrir fitxers, processar-los i re-iniciar el programa, *Help*, per a mostrar ajuda, i *Exit*, per a sortir del programa. L'opció *Superposition* no es troba activa al principi ja que no hi ha estructures moleculars en memòria. Per afegir-n'hi, es seleccionarà l'opció *File*:

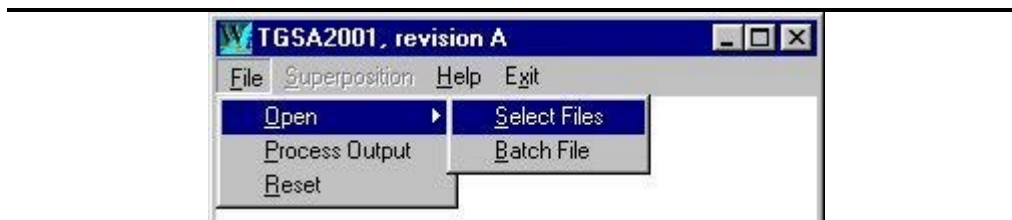


Figura 12.- Opcions desplegadas per l'opció *File*.

L'opció *File* dona tres opcions possibles: *Open*, per a seleccionar els fitxers d'interès, *Process Output*, per a formatar la sortida del programa per a ser entrada d'altres, i *Reset*, per a fer un buidat de memòria i poder tornar a començar sense re-executar el programa. Cal veure que l'opció *Open* es subdivideix en dues opcions més: *Select Files* i *Batch File*. La diferència entre ambdues és que en la primera, la selecció dels fitxers és manual, mentre que en *Batch*, un fitxer de text conté ja els noms dels fitxers d'interès. Tot i que la presència d'aquestes dues opcions pot semblar irrellevant, ja que és possible fer múltiples seleccions sense necessàriament anar afegint fitxer a fitxer, certes configuracions de Windows alteren l'ordre en que els fitxers s'han triat, podent-se donar resultats no esperats si no es segueix aquesta modificació. Així, quan el control de l'ordre és rellevant, o el nombre de fitxers a seleccionar és gran, l'opció *Batch* esdevé interessant. Si es selecciona *Select Files*, s'obre el següent diàleg:

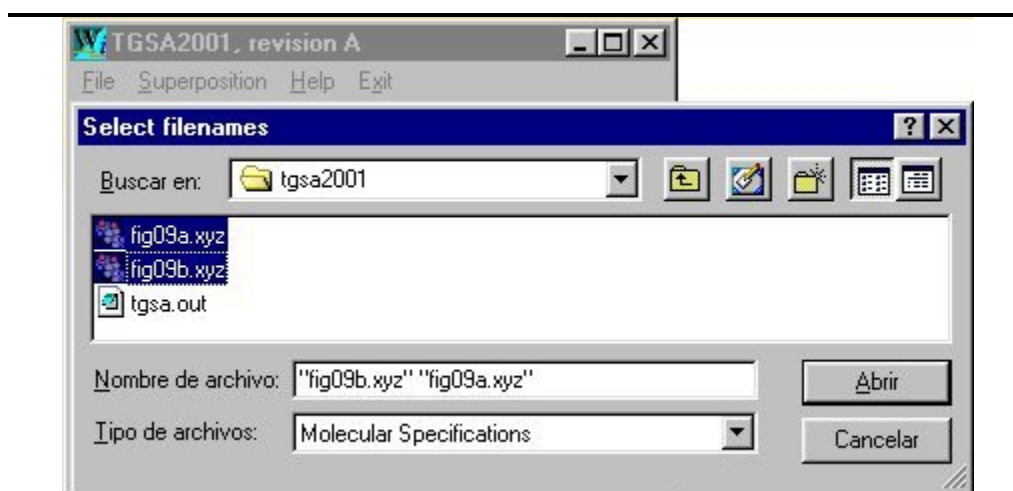


Figura 13.- Selecció manual de fitxers amb l'opció *Select Files*.

On es poden seleccionar manualment els fitxers, en aquest cas, les molècules emprades a la Figura 9. Un cop triats els fitxers, el programa efectua una lectura de les dades fitxer a fitxer, per a comprovar si són llegibles. Així mateix, es guarda en un vector lògic la presència o absència de diferents àtoms al llarg de tota la sèrie de molècules. La identificació del format es fa en base a l'extensió del nom del fitxer, així per exemple *fig09.xyz* denota format XYZ. Altres extensions reconegudes són:

- QFI i LOG: fitxers estàndard de Gaussian.
- FCK i FCH: fitxers checkpoint formatats de Gaussian.
- OUT: fitxers Ampac.
- MOL: fitxers MDL molfile.

Val a dir que no es verifica si el format és correcte, així si un fitxer porta una extensió equivocada o no reconeguda, el programa terminarà amb error per intentar llegir en un format no correcte o no tenir implementada tal rutina de lectura.

Si la lectura de dades s'ha dut a terme sense errors, el diàleg *Superposition* del menú principal s'activa, el qual al ser seleccionat dóna el següent diàleg:

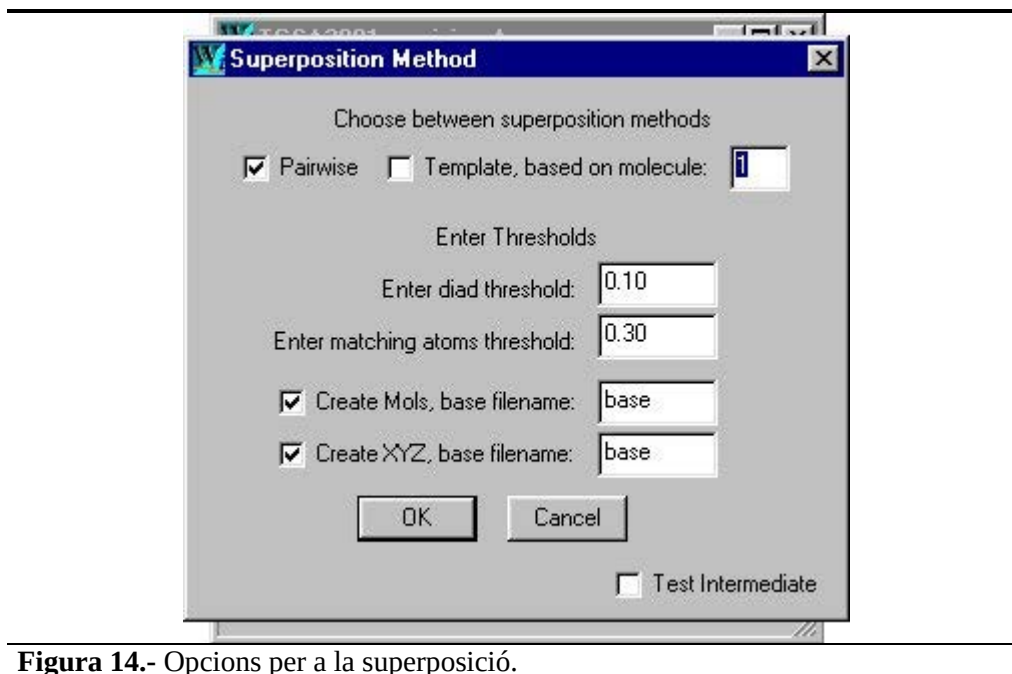


Figura 14.- Opcions per a la superposició.

Primer cal seleccionar si interessa una sobreposició *Pairwise* o *Template*. Per sobreposició pairwise s'entén fer les $n \cdot (n-1)/2$ sobreposicions per parelles, derivades d'un conjunt de n molècules. La sobreposició template realitza les $(n-1)$ sobreposicions sobre una molècula triada en el diàleg de text lateral. En aquest últim cas, el que realment es realitza és $(n-1)$ sobreposicions per parelles, on una d'elles és sempre la molècula triada, i després, a posteriori, es van guardant els resultats, donant finalment una sobreposició múltiple. Els següents diàlegs de text permeten variar els valors per defecte dels llindars de tolerància en la comparació de distàncies interatòmiques (*diad threshold*) i d'àtoms sobreposats (*matching atoms threshold*). Les següents caselles permeten escollir si es vol una sortida de les sobreposicions pairwise, el nom de base d'aquests, i en quins formats de sortida (MDL molfile o XYZ). Cas de ser seleccionades, s'obtenen fitxers de la sèrie *nom_de_base_XXX_YYY.(mol/xyz)*, on *nom_de_base* és el nom triat a les caselles de text laterals i XXX i YYY corresponen als dígitos identificatius de les molècules en l'ordre seleccionat. Finalment, l'opció *Test Intermediate* proporciona sortides a fitxers de les diferents sobreposicions intermitges

que es generen durant el càlcul. Degut a la natura combinatòria del procés, el nombre de fitxers de sortida d'aquesta opció pot ser enorme.

Finalment abans d'iniciar el càlcul, el programa pregunta sobre la realització de càlculs de semblança molecular quàntica sobre les molècules un cop sobreposades. Això es du a terme amb el següent diàleg:

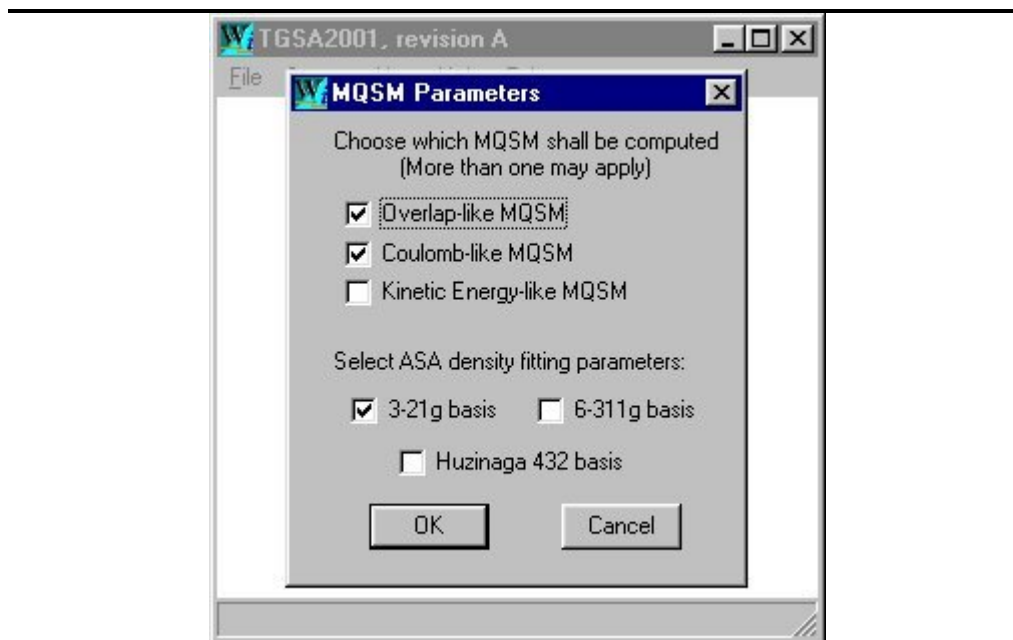


Figura 15.- Opcions per al càlcul de semblança molecular quàntica.

On es permet triar els tipus de mesures a realitzar i la base ajustada promolecular a utilitzar. Cal dir que, en aquesta versió, no es revisa si els ajustos per a un àtom en particular estan definits dins de la base escollida, podent-se donar un error d'execució. A títol d'exemple, no existeixen paràmetres pel iode dins dels ajustos per a les bases 3-21G ni 6-311G.

Un cop carregades les molècules d'exemple, s'inicia el procés de sobreposició. A títol d'exemple, una de les distàncies interatòmiques, dins del conjunt d'àtoms que estan enllaçats, considerades semblants és la que es mostra a la figura següent. Una parella d'àtoms es considera que està enllaçada si la distància que els separa és igual o inferior al 110% de la suma dels seus respectius radis covalents.

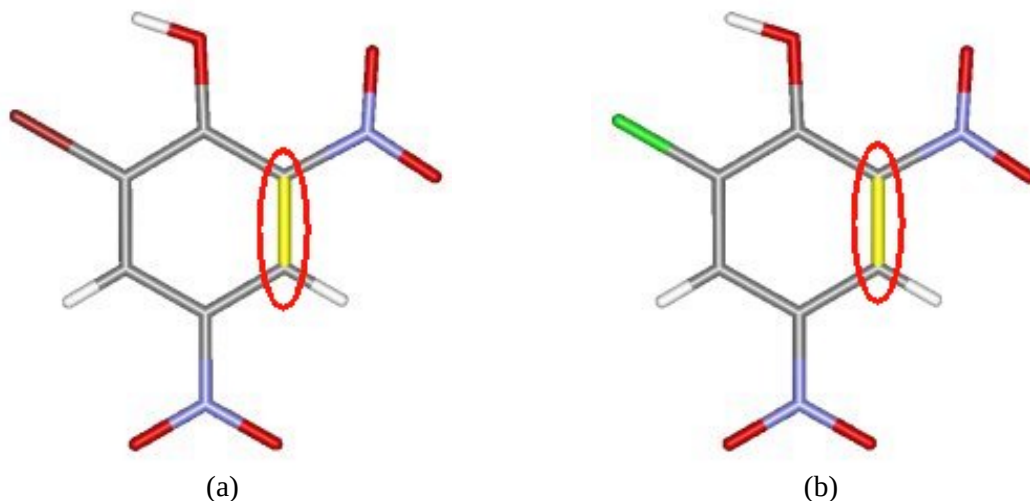


Figura 16.- Parelles d'àtoms de distància considerada com a semblant (0.735 u.a. en ambdós casos).

I una de les tríades atòmiques considerades com a semblants:

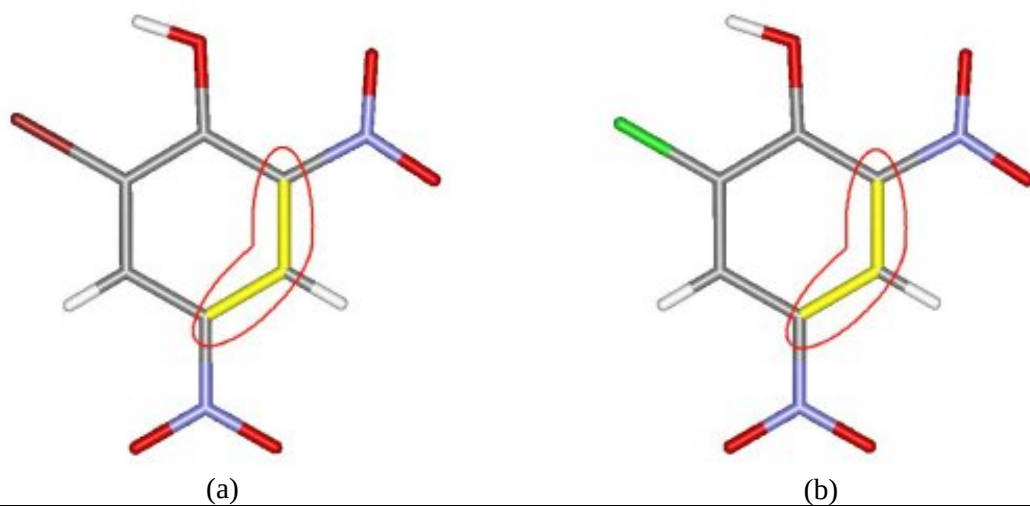


Figura 17.- Tríades d'àtoms de distància considerada com a semblant (el nou enllaç de la tríada també té una longitud de 0.735 u.a. i la distància entre els àtoms no enllaçats directament és de 1.270 u.a. en ambdós casos).

A partir dels tres àtoms seleccionats, ja es pot fer una sobreposició, ja que la fixació de tres punts ja provoca que la resta de la molècula quedi determinada. Així doncs, l'alineament final d'acord a l'algorisme del TGSA és:

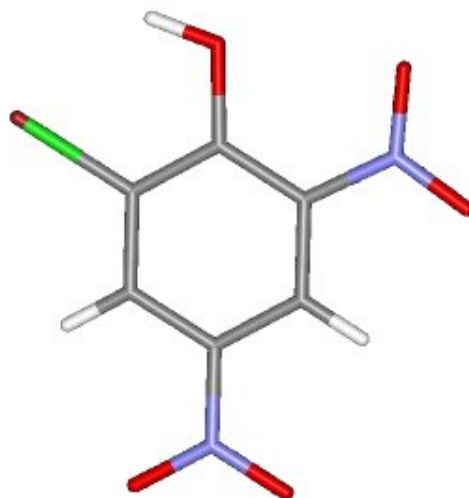


Figura 18.- Sobreposició final del 2-bromo,4,6-dinitrofenol i el 2-cloro,4,6-dinitrofenol.

A l'anterior figura 18, es pot observar com l'esquelet 4,6-dinitrofenòlic ha estat exactament sobreposat, excepte per minúscules fluctuacions derivades de la substitució brom/clor. També, aquests últims no apareixen sobreposats, sinó desplaçats, ja que no formen part de l'esquelet comú molecular. Algunes dades que es deriven del càlcul es mostren a continuació:

- nombre de distàncies interatòmiques semblants: 388.
- nombre de tríades processades: 1032, o el que és el mateix, s'han realitzat 1032 sobreposicions, sent la mostrada a la figura 18 l'escollida.
- nombre d'àtoms sobreposats: 13 (l'esquelet comú, excepte els hidrògens, que no entren en el procediment).
- índex d'ajust: 0.999984.
- temps de càlcul: 3.0 segons (en un PIII-1000MHz / Win98SE)

La simple recerca per *força bruta* de la sobreposició òptima sense considerar les restriccions de distàncies semblants de l'algorisme portaria, en aquest sistema compost per 14 àtoms no-hidrogen, a processar 33124 parelles d'àtoms i 4769856 tríades. La imposició de restriccions ha reduït aquest nombre a 1032, amb l'estalvi computacional que comporta. Alguns exemples d'alineaments processats i descartats pel mètode es mostren a la següent figura 19:

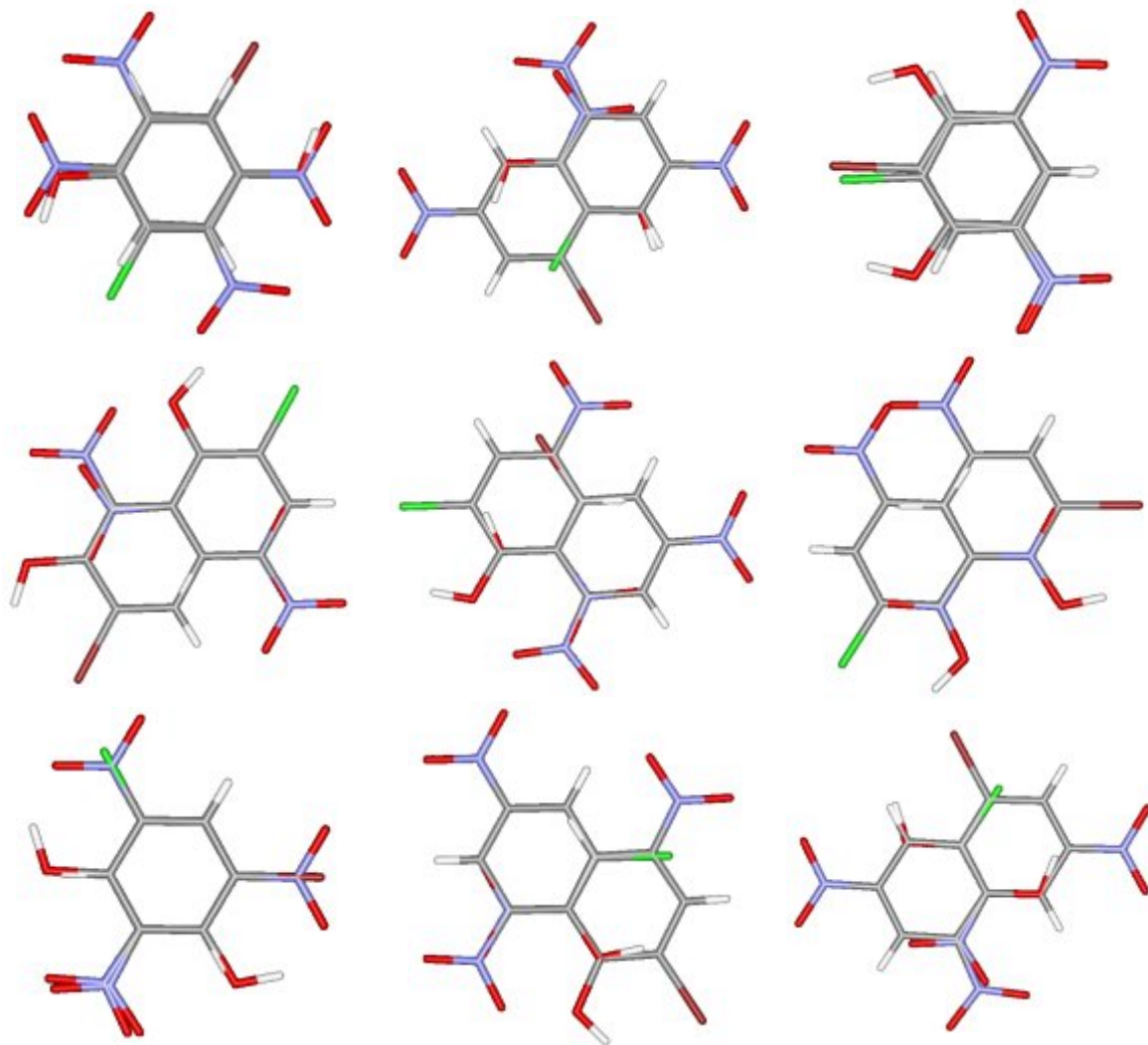


Figura 19.- 9 exemples d'alineaments dels 1032 descartats pel mètode.

Tal i com es veu en els exemples de la figura 19, en tots els casos de sobreposició es troba alguna subestructura comuna. Existeixen diverses disposicions on l'anell benzènic es sobreposa, però amb els diversos substituents rotats, també sobreposant un grup nitro que no equival a la mateixa posició, o bé un enllaç aromàtic amb un substituent. Degut a que sobreposen pocs àtoms i a que les distàncies intermoleculars augmenten amb tals alineaments, són aviat descartades per sobreposicions que ja localitzen l'esquelet comú, de les quals també n'existeixen diverses d'equivalents.

La següent figura 20 mostra la solució TGSA per a les molècules utilitzades al principi del capítol, la nevirapina i l' α -APA:

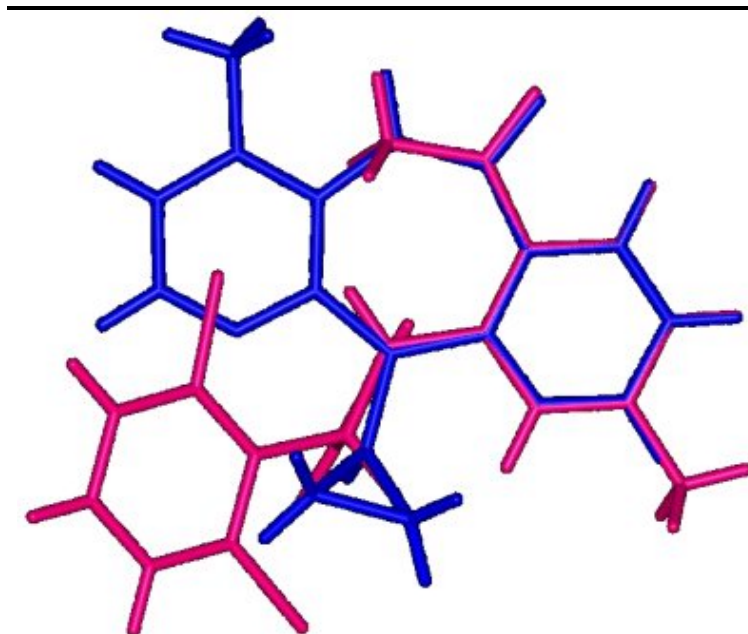


Figura 20.- Sobreposició TGSA de la nevirapina (blau) i la α -APA(fúcsia).

Tal i com es veu a la figura anterior, l'algorisme aconsegueix reconèixer un anell aromàtic (part dreta) i una funcionalització cetònica (part superior). Els paràmetres derivats del càlcul es mostren a continuació:

- nombre de distàncies interatòmiques semblants: 716.
- nombre de tríades processades: 732.
- nombre d'àtoms sobreposats: 8.
- índex d'ajust: 0.947476.
- temps de càlcul: 7.0 segons (en un PIII-1000MHz / Win98SE)

5.3.4 Altres exemples gràfics

En aquest apartat es mostraran 3 exemples de conjunts moleculars sobreposats emprant el programa TGSA2001. Els tres conjunts consisteixen en els 20 aminoàcids essencials, la geometria dels quals ha estat agafada del programa PC Spartan [46], 20 derivats de l'artemisinina (compostos antimalarials [47]), la geometria dels quals ha estat optimitzada amb el camp de forces Sybil incorporat al paquet PC Spartan [46], i finalment, un conjunt de 46 dihidropiridines [48], optimitzades amb l'Ampac 6.01 [49] a nivell AM1 [50]. En els tres casos presentats es mostra un exemple de sobreposició d'una parella de molècules, així com la sobreposició del conjunt total en

base a una estructura. Addicionalment, també es proporcionen dades d'interès estadístics, com el temps de sobreposició, àtoms sobreposats i nombre de díades i tríades processades.

- *Conjunt dels 20 aminoàcids essencials.*

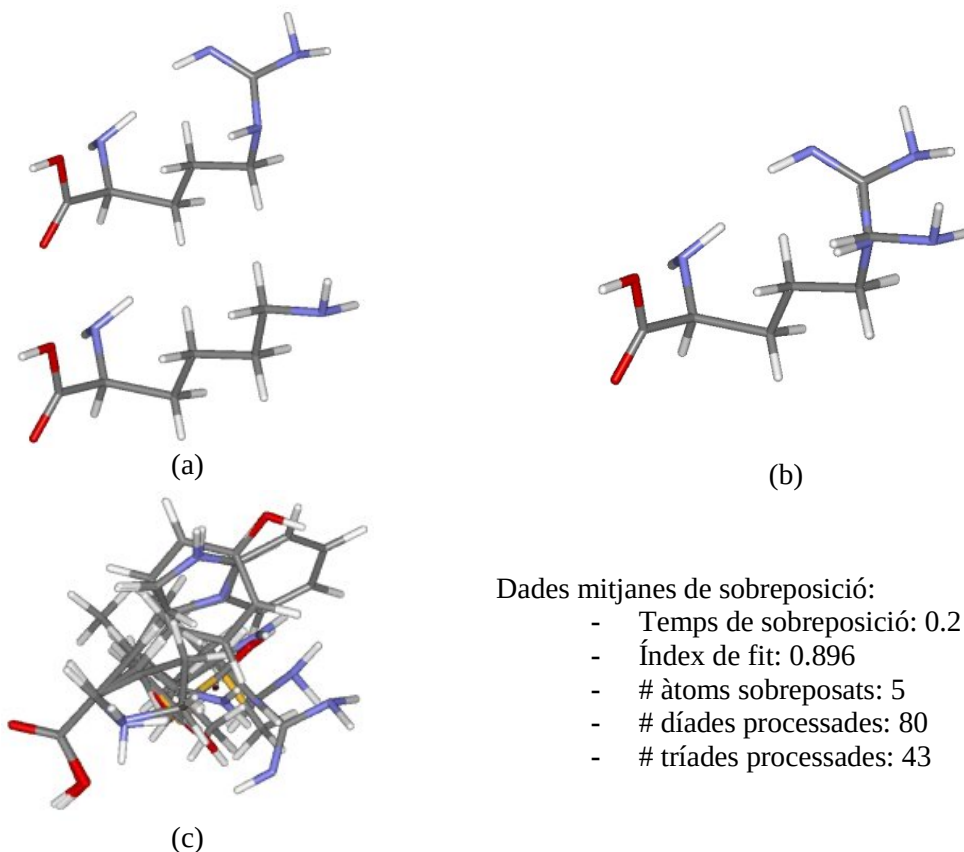


Figura 21.- Alineaments en el conjunt de 20 derivats mitjanes de la sobreposició per parelles. (a) Estructures de l'arginina i la lisina. (b) Sobreposició de l'arginina i la lisina. (c) Sobreposició del conjunt dels 20 aminoàcids en base a la molècula d'alanina.

Aquest senzill exemple dels aminoàcids permet veure clarament l'efectivitat del procediment. Tal i com es pot veure a la Figura 21, l'estructura $-\text{CH}(\text{COOH})\text{NH}_2$ present als aminoàcids fa gairebé indistingibles les molècules presentades a la part (b) de la figura. Al sobreposar la totalitat del conjunt sobre l'alanina (part (c)), l'única part distingible és aquest esquelet comú aminoacid.

- Conjunt de 20 antimalarials derivats de l'artemisinina.

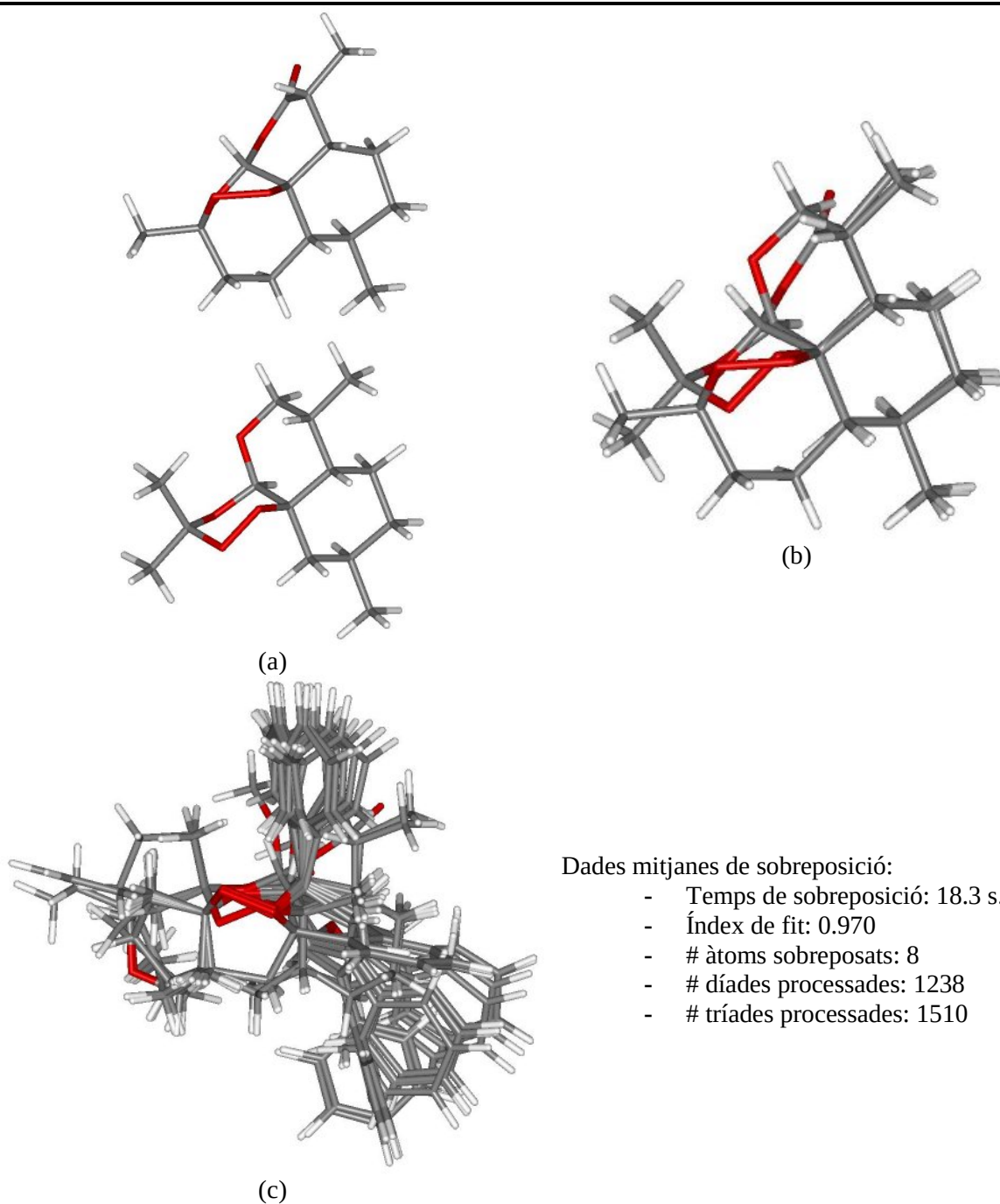


Figura 22.- Alineaments en el conjunt de 20 antimalarals derivats de l'artemisinina i dades mitjanes de la sobreposició per parelles. (a) Estructures de l'artemisinina i un derivat. (b) Sobreposició de l'artemisinina amb el derivat. (c) Sobreposició del conjunt dels 20 compostos en base a l'artemisinina.

- *Conjunt de 46 dihidropiridines.*

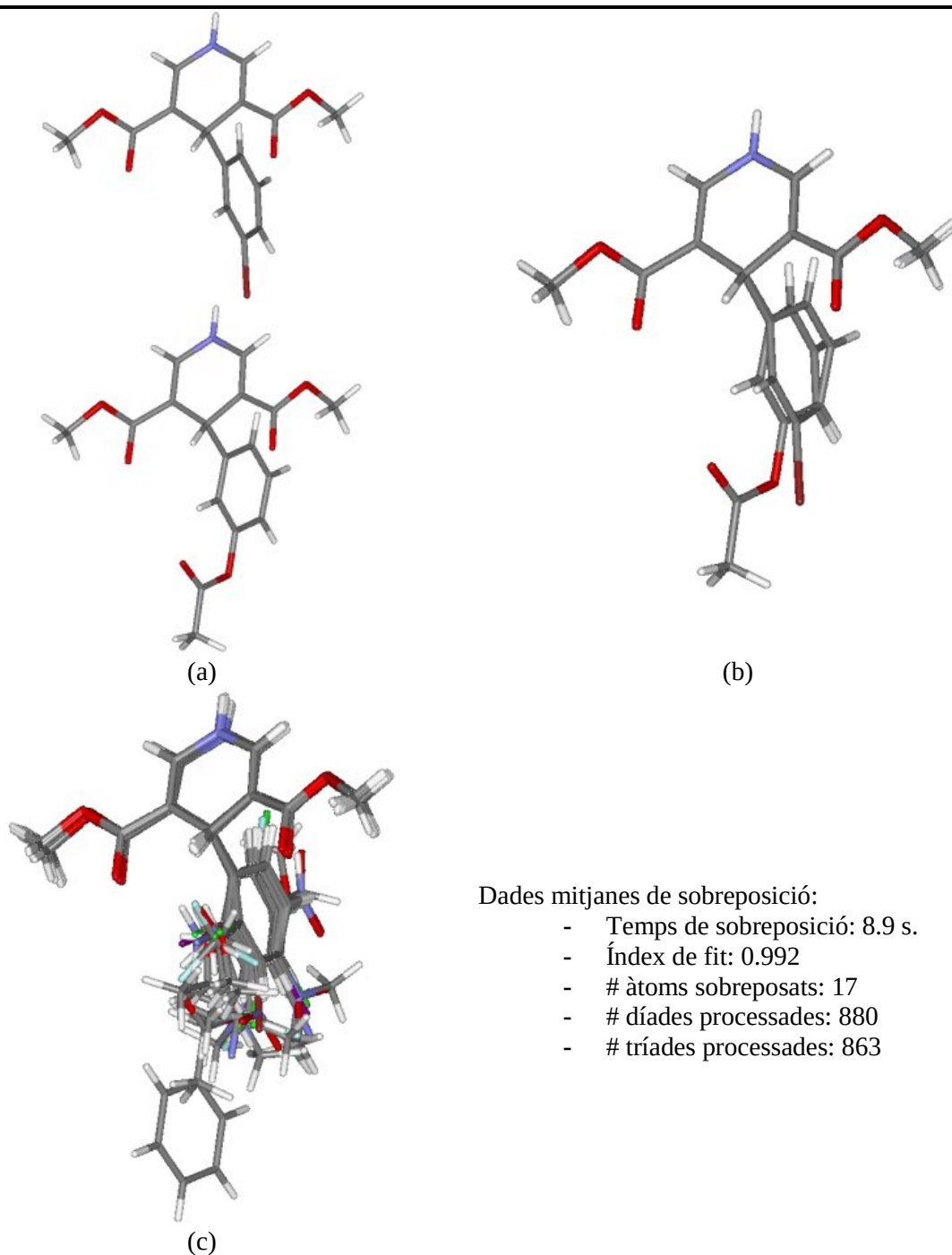


Figura 23.- Alineaments en el conjunt de les 46 dihidropiridines i dades mitjanes de la sobreposició per parelles. (a) Estructures de dos derivats. (b) Sobreposició dels dos derivats. (c) Sobreposició del conjunt de les 46 dihidropiridines en base a la molècula no substituïda.

Tal i com es pot veure en els dos exemples anteriors, formats de molècules amb un esquelet comú considerable, la sobreposició que hom esperaria és aconseguida. En l'exemple dels derivats antimalarials, la sobreposició està majoritàriament basada en l'anell comú 1,2,4-trioxà, present a totes les molècules tractades, però tanmateix, es pot veure a la Figura 22b com, a diferència de la sobreposició del conjunt total (c), l'alineament trobat està basat en l'anell lateral, el qual maximitza l'índex de fit.

En l'exemple de les dihidroxipiridines, l'esquelet comú ja és prou gran, i és prou uniforme al llarg de totes les molècules de la sèrie, per a resultar indistingible en les dues molècules sobreposades en la Figura 23b, i en el cas de la sobreposició conjunta (c), s'observen poques distorsions en aquest, així com també s'alinea bona part de l'anell benzènic enllaçat.

En els tres casos presentats es pot veure no només com els esquelets comuns de les sèries de molècules han estat adequadament detectats i que les sobreposicions han estat adequadament dirigides, sinó també com el temps de càlcul necessari per a realitzar-les resulta força petit, donat el nombre d'àtoms, i de díades i tríades processades en total.

5.4 Conclusions

S'ha desenvolupat un mètode de sobreposició de propòsit general, el qual aconsegueix els propòsits de proporcionar tant un alineament molecular coherent amb la intuïció química, així com proveir aquest amb mínims requeriments computacionals. Basat tan sols en distàncies i nombres atòmics, aquest procediment evita l'ús d'altres paràmetres, abaratint el temps de càlcul.

La seva aplicació en el camp de les mesures semblança molecular quàntica [31], tal i com s'expressa en l'equació (1), dona la possibilitat d'accelerar el seu càlcul al evitar repetir la mesura a cada intent de sobreposició, duta a terme en el procés anterior [32]. Així mateix, al proporcionar un alineament diferent, la mesura resultant reflectirà la semblança referida a la màxima subestructura comuna entre ambdues molècules, enlloc de proporcionar el valor màxim, possiblement esbiaixat si les estructures contenen àtoms pesats.

Així mateix, el desenvolupament d'una interfície gràfica dona al possible usuari més facilitat d'ús del programa, així com evita parcialment l'aparició d'errors d'execució deguts a la introducció de paràmetres erronis.

Com a apunt final, cal dir que s'està treballant ja en pròximes millores en el programa. Aquestes, tot i que algunes d'elles encara en estat força embrionari, inclouen:

- Millores en l'aspecte de la interfície, així com reconeixement de més formats moleculars.
- Incorporació de flexibilitat a les estructures, permetent-se la rotació angles de torsió i variacions en distàncies i angles d'enllaç, possibilitant l'alineament de més àtoms comuns.
- Permetre l'execució del programa en paral·lel. Aquest objectiu permetria una execució, en principi, tantes vegades més ràpida com *nodes*⁴ treballin alhora. Així mateix, la paral·lelització s'hauria de poder dur a terme, segons les possibilitats, de modes diferents:
 - Diferents nodes calculen diferents sobreposicions.
 - Diferents nodes calculen parcialment una sola sobreposició.
 - Diferents nodes controlen com diferents sobreposicions es calculen parcialment en nodes de nivells inferiors.

5.5 Contribucions i articles adjunts

- Gironés, X., Robert, D., and Carbó-Dorca, R. *TGSA: A Molecular Superposition Program Based on Topo-Geometrical Considerations*. J. Comp. Chem., 22, 2001, pp 255-263. Inclós en aquest volum.
- Carbó-Dorca, R., Amat, L., Besalú, E., Gironés, X., and Robert, D. *Quantum molecular similarity: theory and applications to the evaluation of molecular properties, biological activities and toxicity*. In *The fundamentals of molecular similarity*. Kluwer Academic / Plenum Publishers, New York, 2001, pp 197-320.
- Gironés, X., Robert, R., Carbó-Dorca, R. *Molecular Superposition: A Topological - Geometrical Approach*. IV Girona Seminar on Molecular Similarity. Juliol 1999.

⁴ Per node s'entén bàsicament processador, tot i que també en l'àmbit de la supercomputació es pot entendre com a part d'un conjunt de màquines, les quals contenen en si mateixes diversos processadors.

5.6 Referències

1. Crammer III, R.D., Paterson, D.E., and Bunce, J.D. *Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins.* J. Am. Chem. Soc., 110, 1988, pp 5959-5967.
2. Carbó-Dorca, R., and Besalú, E. *A general survey of quantum similarity.* J. Mol. Struct. (Theochem), 451, 1998, pp 11-23.
3. Bravi, G., Gancia, E., Mascagni, P., Pegna, M., Todeschini, R., and Zalianni, A. *MS-WHIM, new 3D theoretical descriptors derived from molecular surface properties: A comparative 3D QSAR study in a series of steroids.* J. Comput.-Aided Mol. Des., 11, 1997, pp 79-89.
4. Kellogg, G.E., Kier, L.B., Gaillard, P., and Hall, L.H. *The E-State Fields. Applications to 3D QSAR.* J. Comput.-Aided Mol. Design, 10, 1996, pp 513-520.
5. Haneef, B.I., Moss, D.S., Standford, M.J., and Borkakoti, N. *Restrained Structure-Factor Least-Squares Refinement of Protein Structures Using a Vector Processing Computer.* Acta Crystallogr., A41, 1985, pp 426-433.
6. Sippl, M.J., and Stegbuchner, H. *Superposition of Three Dimensional Objects: A Fast and Numerically Stable Algorithm for the Calculation of the Matrix of Optimal Rotation.* Comput. Chem., 15, 1991, pp 73-78
7. Allen, F.H., Bellard, S., Brice, M.D., Cartwright, B.A., Doubleday, A., Higgs, H., Hummelink, T., Humelink-Peters, B.G., Ken-nard, O., Motherwell, W.D.S., Rodgers, J.R., and Watson, D.G. *The Cambridge Crystallographic Data Centre: computer-based search, retrieval, analysis and display of information.* Acta Crystallogr., B35, 1979, pp 2331-2339.
8. Artymiuk, P.J., Bath, P.A., Grindley, H.M., Pepperrell, C.A., Poirrette, A.R., Thorner, D.A., Wild, D.J., Willet, P., Allen, F.H., and Taylor, R. *Similarity searching in databases of three-dimensional molecules and macromolecules.* J. Chem. Inf. Comput. Sci., 32, 1992, pp 617-630.
9. Attias, R. *DARC Substructure Search System: A New Approach to Chemical Information.* J. Chem. Inf. Comput. Sci., 23, 1983, pp 102-108.
10. Brint, A.T., and Willet, P. *Algorithms for the identification of three-dimensional maximal common substructures.* J. Chem. Inf. Comput. Sci., 27, 1987, pp 152-158.
11. Crandell, C.W., and Smith, D.H. *Computer assisted examination of compounds for common 3D substructures.* J. Chem. Inf. Comput. Sci., 23, 1983, pp 186-197.
12. Cringean, J.K., and Lynch, M.F. *Subgraphs of reduced chemical graphs as screens for substructure searching of specific chemical structures.* J. Inf. Sci., 15, 1989, pp 211-222.
13. Gushurst, A.J., Nourse, J.G., Houndshell, W.D., Leland, B.A., and Raich, D.G. *The Substance Module: The Representation, Storage, and Searching of Complex Structures.* J. Chem. Inf. Comput. Sci., 31, 1991, pp 447-454.
14. Hagadone, T. R. *Molecular substructure similarity searching : Efficient retrieval in two-dimensional structure databases.* J. Chem. Inf. Comput. Sci., 32, 1992, pp 515-521.
15. Rasmussen, E.M., Downs, G.M., and Willett, P. *Automatic classification of chemical structure databases using a highly parallel array processor.* J. Comput. Chem., 9, 1988, pp 378-386.
16. Willet, P. *An algorithm for chemical superstructure searching.* J. Chem. Inf. Comput. Sci., 25, 1985, pp 114-116.
17. Willet, P. *Textual and chemical information processing using parallel computer hardware.* J. Inf. Sci., 15, 1989, pp 223-236.
18. Willet, P., Winterman, V., and Bawden, D. *Implementation of nearest neighbour searching in an online chemical structure search system.* J. Chem. Inf. Comput. Sci., 26, 1986, pp 36-41.
19. Willet, P., Winterman, V., and Bawden, D. *Implementation of non-hierarchic cluster analysis methods in chemical information systems: selection of compounds for biological testing and clustering of substructure search output.* J. Chem. Inf. Comput. Sci., 26, 1986, pp 109-118.

20. K. Prendergast, K. Adams, W. L. Greenlee, R. B. Nachbar, Patchett, A.A., and Underwood, D.J. *Derivation of a 3D pharmacophore model for the angiotensin-II site one receptor*. J. Comput.-Aided Molec. Des., 8, 1994, pp 491-512.
21. Halgren, T. A. *Merck Molecular Force Field: I. Basis, Form, Scope, Parameterization and Performance of MMFF94*. J. Comp. Chem., 17, 1996, pp 490-519.
22. Mohamadi, F., Richards, N.G.J., Guida, W.C., Liskamp, R., Lipton, M., Caufield, C., Chang, G., Hendrickson, T., and Still, W.C. *MacroModel - an Integrated Software System for Modeling Organic and Bioorganic Molecules Using Molecular Mechanics*. J. Comput. Chem., 11, 1990, pp 440-467.
23. Lejeune, J., Michel, A., and Vercauteren, D.P. J. Comput. Chem., 7, 1986, pp 739-744.
24. McMartin, C., and Bohacek, R.S. *Flexible matching of test ligands to a 3D pharmacophore using a molecular superposition force field: Comparison of predicted and experimental conformations of inhibitors of three enzymes*. J. Comput.-Aided Molec. Des., 9, 1995, pp 237-250.
25. Ren, J. Esnouf, R., Garman, E., Somers, S., Ross, C., Kirby, I., Keeling, J., Darby, G., Jones, Y., Stuart, D., and Stammers, D. *High resolution structures of HIV-1 RT from four RT-inhibitor complexes*. Nature Struct. Biol., 2, 1995, pp 293-302.
26. Clark, D.E., and Westhead, D.R. *Evolutionary algorithms in computer-aided molecular design*. J. Comput.-Aided Molec.Design, 10, 1996, pp 337-358.
27. Martin, Y.C., Bures, M.G., Danaher, E.A., DeLazzer, J., Li-co, I., and Pavlik, P.A. *A Fast New Approach to Pharmacophore Mapping and It's Application to Dopaminergic and Benzodiazepine Agonists*. J. Comput.-Aided Molec. Des., 7, 1993, pp 83-102.
28. Miller, M.D. *SQ A program for producing rapid molecular superimpositions*. ACS 207th National Meeting and Exposition, Computers in Chemistry Section, San Diego California, 1993.
29. Miller, M.D. *SQ A program for the rapid production of multimolecular superposition*. ACS 210th National Meeting and Exposition, Medicinal Chemistry Section, Chicago, Illinois, 1995.
30. Miller, M.D. *Molecular Superposition*. In *Encyclopedia of Computational Chemistry*. Schleyer, P.R., Allinger, N.L., Clark, T., Gasteiger, J., Kollman, P.A., Schaefer III, H.F. and Screiner, P.R. (eds.). John Wiley and Sons Ltd., Chichester, 1994, Vol. 3, 1693-1697.
31. Carbó, R., Leyda, L., and Arnau, M. *How similar is a molecule to another? An electron density measure of the similarity between two molecular structures*. Int. J. Quant. Chem., 17, 1980, pp 1185-1189.
32. Constans, P., Amat, L., and Carbó-Dorca, R. *Towards a global maximization of the molecular similarity function: the superposition of two molecules*. J. Comp. Chem., 18, 1996, pp 826-846.
33. Mestres, J., Rohrer, D.C., and Maggiora, J.M. *MIMIC: a molecular-filed matching program. Exploiting applicability of molecular similarity approaches*. J. Comput. Chem., 18, 1997, pp 934-954.
34. Grant, J.A., Gallardo, M.A., and Pickup, B.T. *A Fast Method of Molecular Shape Comparison A simple application of a gaussian description of molecular shape*. J. Comput. Chem., 17, 1996, pp 1653-1666.
35. Perkins, T.D.J., Mills, J.E.J., and Dean, P.M. *Molecular surface-volume and property matching to superpose flexible dissimilar molecules..* J. Comput.-Aided Molec. Des., 9, 1995, pp 479-490.
36. Hermann, R.B., and Herron, D.K. *OVID and SUPER: Two overlap programs for drug design*. J. Comput.-Aided Molec. Des., 5, 1991, pp 511-524.
37. Nissink, J.W.M., Verdonk, M.L., Kroon, J., Mietzner, T., and Klebe, G. *Superposition of molecules: electron density fitting by application of Fourier transforms*. J. Comput. Chem., 18, 1997, pp 638-645.
38. Bladon, P. *A rapid method for comparing and matching the spherical parameter surfaces of molecules and other irregular objects*. J. Mol. Graphics, 7, 1989, pp 130-137.

39. Masek, B.B., Merchant, A., and Matthew, J.B. *Molecular skins: a new concept for quantitative shape matching of a protein with its small molecule mimics*. *Proteins*, 17, 1993, pp 193-202.
40. Kearsley, S.K., and Smith, G.M. *An alternative method for the alignment of molecular structures: Maximizing electrostatic and steric overlap*. *Tetrahedron Computer Methodology*, 3, 1992, pp 615-633.
41. Carbó, R., and Calabuig, B. *MOLSIMIL-88: molecular similarity calculations using CNDO-like approximation*. *Comp. Phys. Commun.*, 55, 1989, pp 117-126.
42. Gironés, X., Robert, D., and Carbó-Dorca, R. *TGSA: A Molecular Superposition Program Based on Topo-Geometrical Considerations*. *J. Comp. Chem.*, 22, 2001, pp 255-263.
43. Ampac 6.01, Semichem, 7128 Summit, Shawnee, KS 66216DA, 1994.
44. Frisch, M.J., Trucks, G. W., Schlegel, H.B., Scuseria, G.E., Robb, M.A., Cheeseman, J.R., Zakrzewski, V.G., Montgomery, Jr., J.A., Stratmann, R.E., Burant, J.C., Dapprich, S., Millam, J.M., Daniels, D., Kudin, K.N., Strain, M.C., Farkas, O., Tomasi, J., Barone, V., Cossi, M., Cammi, R., Mennucci, B., Pomelli, C., Adamo, C., Clifford, S., Ochterski, J., Petersson, G.A., Ayala, P.Y., Cui, Q., Morokuma, K., Malick, D.K., Rabuck, A.D., Raghavachari, K., Foresman, J.B., Cioslowski, J., Ortiz, J.V., Stefanov, B.B., Liu, G., Liashenko, A., Piskorz, P., Komaromi, I., Gomperts, R., Martin, R.L., Fox, D.J., Keith, M. T., Al-Laham, A., Peng, C.Y., Nanayakkara, A., Gonzalez, C., Challacombe, M., Gill, P.M.W., Johnson, B., Chen, W., Wong, M.W., Andres, J.L., Gonzalez, C., Head-Gordon, M., Replogle, E.S., and Pople, J.A. *Gaussian 98*, Revision A.6, Gaussian, Inc., Pittsburgh PA, 1998.
45. Stahl, M., and Walters, P. *Babel 1.6*, Department of Chemistry, University of Arizona, Tucson, AZ, 1996.
46. *PC Spartan Pro*, version 1.4, 1997, Wavefunction Inc, Irvine.
47. Grigorov, M., Weber, J., Tronchet, J.M.J., Jefford, C.W., Milhous, W.K. and, Maric, D. *A QSAR Study of the Antimalarial Activity of Some Synthetic 1,2,4-Trioxanes*. *J. Chem. Inf. Comput. Sci.*, 37, 1997, 124-130.
48. Viswanadhan, V.N., Mueller, G.A., Basak, S.C., and Weinstein, J.N. *Comparison of a Neural Net-Based QSAR Algorithm (PCANN) with Hologram- and Multiple Linear Regression-Based QSAR Approaches: Application to 1,4-Dihydropyridine-Based Calcium Channel Antagonists*. *J. Chem. Inf. Comput. Sci.*, 41, 2001, pp 505-511
49. Ampac 6.01, Semichem, 7128 Summit, Shawnee, KS 66216DA, 1994.
50. Dewar, M. J. S., Zoebisch, E. G., Healy, E. F., and Stewart, J. J. P. *AM1: A New General Purpose Quantum Chemical Molecular Model*. *J. Am. Chem. Soc.*, 107, 1985, pp 3902-3909.

5.7 Article Adjunt

En aquesta secció s'inclou a mode d'annex la contribució llistada a continuació, la qual ha estat publicada.

- Gironés, X., Robert, D., and Carbó-Dorca, R. *TGSA: A Molecular Superposition Program Based on Topo-Geometrical Considerations*. J. Comp. Chem., 22, 2001, pp 255-263.

TGSA: A Molecular Superposition Program Based on Topo-Geometrical Considerations

XAVIER GIRONÉS, DAVID ROBERT, RAMON CARBÓ-DORCA

Institute of Computational Chemistry, University of Girona, Catalonia 17071, Spain

Received 22 May 2000; accepted 7 August 2000

ABSTRACT: In this article, a new molecular alignment procedure to provide general-purpose, fast, automatic, and user-intuitive three-dimensional molecular alignments is presented. This procedure, called Topo-Geometrical Superposition Approach (TGSA), is only based on comparisons of atom types and interatomic distances; hence, the procedure can handle large molecular sets within affordable computational costs. The method is able to accurately align 3D structures using the common molecular substructures, as inferred by the bonding pattern (atom correspondences), where present. The algorithm has been implemented into a program named TGSA99, and it has been tested over eight different molecular sets: flavilium salts, amino acids, indole derivatives, AZT, steroids, anilide derivatives, poly-aromatic-hydrocarbons, and inhibitors of thrombine. The TGSA algorithm performance is evaluated by means of computational time, number of superposed atoms, and index of fit between the compared structures. © 2000 John Wiley & Sons, Inc. *J Comput Chem* 22: 255–263, 2001

Keywords: molecular superposition; alignment; topo-geometrical approach; TGSA; computer program

Introduction

Molecular alignment has become a widely studied field, as it has been introduced and applied to several disciplines. The most common uses of molecular superposition involve X-ray crystallography, protein structural research, quan-

tification of similarity between a given structure optimized with different methods, and determination of the distortion produced by a chemical substitution.^{1–3} In addition, it has recently been applied to the evaluation of physico-chemical properties, database search, and pattern recognition and construction of molecular descriptors to be used in 3D-QSAR models.

So far, optimal alignment constitutes a major problem in the aforementioned fields. For example, all techniques based on obtaining information from direct comparison of two or more 3D molecular structures come to the conclusion that the results

Correspondence to: R. Carbó-Dorca; e-mail: quantum@stark.udg.es

Contract/grant sponsor: Fundació Maria Francisca de Roviralta; contract/grant number: ENV4-CT97-0508

depend on the way the studied objects are located in space. Some 3D-QSAR models (see, e.g., Ref. 4) are based on calculation of force fields at each point of a grid surrounding the molecules. Small variations on the relative spatial position of molecules may modify the values of the evaluated function at these grid points, and therefore, modifying the quality of the QSAR model obtained. Time efficiency is another important factor present in alignment processes. Searching for homologous compounds or for common chemical groups in large molecular databases⁵ may become a tedious task, if the method used to match or discard structures is slow.

In this way, several methods have been proposed to establish a criterion on how molecules might be superposed. For example, matching different molecular fields,⁶ such as steric, hydrophobic, shape, or electrostatic, allows to align the involved molecules according to common physico-chemical features. Other methodologies align the molecules according to the minimization or maximization of a function that depends on the relative position of both molecules in space. For example, minimization of a weighted sum of distances between corresponding atoms to different molecules,⁷ or related procedures align the molecules in such a way that a molecular similarity measure is maximized.⁸ Other approaches are related to the graph theory,⁹ which aligns the molecules according to subgraphs coincidences, or usage of reduced dimensional spaces that best fit the studied molecules.

The proposal of this article consists of defining a new, easy understandable, molecular alignment method based only on topological and geometrical features. The major goal is to obtain, with low computational costs, a general superposition procedure able to calculate pairwise alignments restricted to common skeleton recognition. This procedure, called Topo-Geometrical Superposition Algorithm (TGSA), has been designed to operate with homogeneous sets of molecules, and due to the basis of the method lies on common substructure correspondences, improper alignments could be obtained where mutual structural features are not present. TGSA will be explained in the next sections, and has been implemented in a computer program named TGSA99.

Theory of Operation

The main purpose of TGSA consists on finding the largest molecular substructure common to a series of molecules, and orienting the molecular



FIGURE 1. Molecular structures of glutamine and histidine. [Color figure can be viewed in the online issue, which is available at www.interscience.com.]

superposition towards the matching of such substructure. To provide a better comprehension of the TGSA, some of the procedure concepts that will be discussed along this section have been graphically illustrated with an application example. The chosen molecular structures to be superposed, presented in Figure 1, are two essential amino acids, glutamine and histidine, that present an obvious common skeleton, the $-\text{CHNH}_2\text{COOH}$ group, which is the desired target to be detected and superposed. The selected molecular coordinates have been obtained from the Spartan program database.¹⁰

DIAD DEFINITION

TGSA algorithm only requires the atomic numbers and coordinates of the molecules to be superposed. If real chemical problems are studied, these coordinates should represent molecular bioactive conformations, as TGSA considers the molecules as rigid bodies and no conformational search is done. Once this information has been supplied, the first step involves the definition of atomic pairs, called *diads*. Diads are defined only if the involved atom pairs are bonded. This criterion has been adopted by assuming accepted the definition of common skeleton: coincident sets of atoms in space within a group of molecules. The way bonds are defined will be discussed in the next section, devoted to the TGSA practical implementation. The whole number of existing bonds in a studied molecule determines the number of diads that will be taken into account during the alignment process. At this point, it must be stressed that hydrogen atoms are not included in any TGSA process in order to spare computational time.

DIAD COMPARISON

When all diads have been defined for each one of the molecules, every diad of one molecule is compared to all diads of the other by means of their interatomic distances, within a given length threshold. This threshold is included for taking into account the present fluctuations in backbone atom



FIGURE 2. Graphical example of a pair of similar diads. The highlighted bond is located between the carboxylic and the chiral carbons. [Color figure can be viewed in the online issue, which is available at www.interscience.com.]

conformations produced by the presence of different substitutions within a molecule. This simple procedure allows to discard a considerable amount of bonds, which do not belong to the common skeleton. Figure 2 presents an example of one of the selected diads, corresponding to a pair of carbon bonded atoms in the two molecular structures.

In both molecules in the example, one of the carbon atoms of the diads selected in Figure 2 corresponds to the central atom of the acidic group, whereas the other one is the chiral carbon present in all amino acids, except glycine.

TRIAD DEFINITION

Once all diads have been compared, the next step of the algorithm creates atomic *triads* by adding a third atom to the selected diads. Similarly to the previous step, the atom added must be bonded to at least one atom of the considered diad. Regarding the atomic positions, this step can be thought, in geometrical terms, as the generation of a triangle. Atoms occupy the vertices of this triangle, and two or three sides of a given triad correspond to effective chemical bonds.

TRIAD COMPARISON

All triangles obtained for one molecule are compared to all triangles obtained by the second molecule by means of their interatomic distances, within a given length threshold. The value of the distance threshold for triads is the same as the adopted in the diads comparison step. If the three distances of both compared triangles are similar, both triads also are considered similar and stored. The triads that do not fulfill this criterion are automatically discarded. Figure 3 presents a pair of selected triads for the amino acids analyzed.

One of the selected triads, presented in Figure 3, is made up of the carbon-carbon diad, previously selected, together with the carboxylic oxygen. In this case, the triads are formed by two effective bonds; however, nothing opposes the participation of the



FIGURE 3. Graphical example of a pair of similar triads (highlighted). In this case, the third selected atom is the carboxylic oxygen. [Color figure can be viewed in the online issue, which is available at www.interscience.com.]

three bonds, as it would happen if, for instance, epoxy groups were selected.

The reason for creating triads is such that when an object is to be placed in a given position in space, three points of reference are needed, three atoms in our case. Upon completing the triad comparison step, those selected are superposed, and the resulting molecular alignment is univocally determined. The case presented in Figure 3 satisfies the distances condition, and would be superposed.

SUPERPOSITION PROCEDURE

At this point, two opposite situations can occur: either no pair of triads or several pairs are selected. These situations mainly correspond to heterogeneous and homogeneous sets of molecules, respectively. The first case can be easily solved by means of increasing the distance threshold and then repeating the whole process as many times as needed, until some triad is selected. However, it should be pointed out that TGSA is designed to find common molecular substructures; thus, homogeneity of the molecular set is supposedly present. The second situation requires the definition of a method able to select between the set of similar triads. Here, the adopted criterion consists of choosing the pair of triads that, when superposed, leads both molecular structures to superpose by sharing the maximal number of identical atoms, measured with a second given threshold distance. In addition, if different triad superpositions match the same number of atoms, the selected triad couple will be the one that minimizes the *index of fit*, defined in the practical implementation section, associated to intermolecular distances [see eq. (2)]. Figure 4 shows the solution found for the example molecules.

As it can be observed in Figure 4, the final molecular alignment correctly superposes the common amino acid group present in both molecules. As commented above, the most similar pair of triads corresponds to the C—C=O triangle, and matching these molecular substructures results in an alignment that also overlays the hydroxilic oxygen, the



FIGURE 4. Final molecular alignment. Superposed atoms are highlighted for clarity. [Color figure can be viewed in the online issue, which is available at www.interscience.com.]

amino group, and a carbon atom of the side chain. As a whole, six atoms are superposed, excluding the hydrogens.

Practical Implementation

Having outlined the TGSA procedure, in this section a detailed discussion about the actual operation flow chart of TGSA99 is performed. TGSA99 has been coded in FORTRAN 90 language and tested in a Silicon Graphics Origin2000 with an R10000/180 MHz processor and 600 Mb RAM.

When the TGSA program is executed, the coordinates of the first molecule, considered to be fixed in space, are read and the number of nonhydrogen atoms determined. Then, all interatomic distances are calculated and stored in a matrix that will be later used to retrieve these distance values, thus avoiding having to recompute them at later stages. Diads are assigned by means of the effective bonds present in the molecule. If no bond information is present in the input file, the following approach is applied: two atoms are considered to be bonded if their interatomic distance is below 1.1 times the sum of their respective covalent single bond radii.¹¹ Bond information is then stored in a logical matrix, which allows a fast bond determination. A second molecule, considered the mobile one, which will be superposed to the previous by means of translations and rotations, is subjected to the same procedure as described above.

The next steps are carried out involving pairs of molecules. The first search compares the diads of both molecules: that is, each diad of one molecule is compared to all diads of the other. Diads are considered to be similar if they differ within a given distance threshold. The user may change the threshold value, but practical experience demonstrates that a value higher than 0.10 atomic units (au), set as default in the program, does not improve the results in homogeneous series while increasing the computation time. Values below 0.10 au and up to 0.05 au also provide good results, but a further

decrement of this parameter may result in incoherent superpositions. A logical matrix is built up to keep information on comparable diads.

When all diads have been formed and the selected ones stored, the triad creation procedure is carried out. To all selected diads an additional atom is combined, which produces the aforementioned atomic triangles. The new added atoms must be bonded to at least one atom of the selected diads. Each triad created in the fixed molecule is compared to all possible triads in the mobile one by means of the new two interatomic distances appearing because of the third atom addition, using the same given threshold as in the diad comparison step. A pair of triads is considered to be similar if both created distances are comparable. If the triads are similar, they are tested for superposition and overlaid one to another.

All superposition tests are evaluated by the number of identical atoms matched and the index of fit. The optimal superposition will coincide with the one that matches the greatest number of equal atoms and, in the case that more than one pair of triads is found with this characteristic equal, it is chosen the one, which minimizes the value of the index of fit. An atom is considered to be matched to another only if they have the same atomic charge and their distance lies below 0.30 au. Even though this last threshold may also be modified, several tests have been carried out, and the experience has proven this quantity as the one yielding an acceptable algorithmic behavior. The index of fit between two molecules *A* and *B* (C_{AB}) is defined as a ratio between the product of the sum of interatomic distances in *A* and *B*, and the sum of intermolecular distances between both molecules. C_{AB} is mathematically expressed by means of the sum of squared difference modules:

$$d_{AB} = \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} |\mathbf{x}_{i,A} - \mathbf{x}_{j,B}|^2, \quad (1)$$

where *n* is the number of atoms and **x** are the Cartesian coordinates (in au) of molecules *A* and *B*. Then the index of fit can be constructed as:

$$C_{AB} = \left(\frac{(d_{AA}d_{BB})^{1/2}}{d_{AB}} \right)^{1/2}. \quad (2)$$

In this way, C_{AB} assesses the goodness of superposition by means of an index comprised between the interval [0, 1]. The closest C_{AB} is to unity, the better the superposition is achieved. The ideal case, $C_{AB} = 1$, would point out that the compared molecules are the same. On the other hand, as C_{AB} approaches zero, the superposition obtained

is less accurate. Thus, C_{AB} provides a measure of the superposition by quoting the distances between nonmatched molecular substructures and molecular distortions.

If no pair of triads is found, the process is repeated, starting from the diad comparison step after doubling the distance threshold, until at least one pair of comparable triads is found. This situation is not likely to happen if the studied molecular set is built up of reasonably similar molecules.

Finally, the selected pair of triads is overlaid and the optimal superposition, within this procedure, is achieved. At this stage, the superposition process terminates, and any kind of computation can be performed afterwards. From this stage, the whole procedure is repeated until all molecules in a given set have been aligned.

Options and Built-in Calculations

In this section, the present capabilities of TGSA99 are described. They are related to how the molecules are aligned and the diverse calculations, which can be performed when the optimal superposition is achieved.

TGSA99 can perform the molecular superposition within a molecular set in two ways: (1) calculating all pairwise alignments; and (2) choosing a molecule as a template. The first option implies that the superposition procedure will compute the $N(N - 1)/2$ alignments, N being the number of molecules to be analyzed. The second option, using a given molecule as a template, will place this molecule in space, and the rest of the molecular set, considered as mobile molecules, will be superposed to the previously chosen structure. Within this framework, $N - 1$ alignment computations are executed, and all molecules are placed in space with respect to the chosen template. All molecular coordinates, conveniently oriented to correctly superpose the template, are stored so as to execute the desired computations afterwards.

Even if distance threshold values have already been fixed so as to perform fast and accurate superpositions in homogeneous series, they can be conveniently modified for user convenience. This situation may happen when the studied molecular set is made of highly heterogeneous molecules, and default values may not be able to provide a good superposition.

For visualization purposes, TGSA99 include a set of routines that provide output MOL formatted files that encode the molecules themselves, pairwise superpositions, and the entire molecular set alignment

based on a template molecular structure. These files can be visualized with the available software (see, e.g., ref. 12).

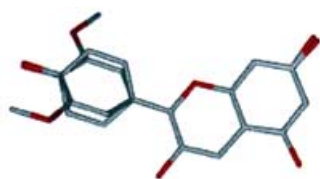
Finally, as a significant amount of research in our laboratory is related to molecular quantum similarity measures (MQSM),¹³ several routines dealing with these measures have been included in TGSA99. These measures include overlap-like,¹³ Coulomb-like,¹⁴ and the recently developed overlap-like kinetic energy¹⁵ quantum similarity measures. The values of these measures depend on the relative orientation of the molecules involved, so the use of different alignments allows obtaining different MQSM values. Within TGSA99, the MQSM evaluations that can be obtained consist of the pairwise alignments producing a unique similarity matrix and the ones based on templates. As template-based superposition unequivocally docks the entire molecular set, a different similarity matrix can be obtained for each template.

Application Examples

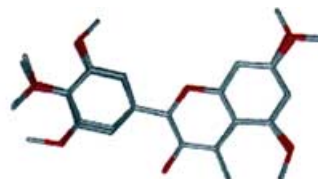
To demonstrate the applicability of the proposed TGSA method, several examples will be presented. The alignment procedure will be applied to eight different sets of compounds, which constitute sufficient a set of diverse situations:

1. A set of 16 flavilium salts, which produce an inhibitory effect in Xanthine Oxidase enzyme.¹⁶ This molecular set is built up from two fused and a bonded six-member rings.
2. The 20 essential amino acid set.¹⁰ This molecular group contains the aforementioned $-\text{CHNH}_2\text{COOH}$ group, which is desired to be detected and superposed.
3. A set of 23 indole derivatives, which produce a displacement of [^3H]-flunitrazepam from binding to bovine brain membrane.¹⁷ These structures contain the indole group, formed by two fused six- and five-membered rings.
4. A molecular set composed by 27 AZT derivatives,¹⁸ which possess a high activity against HIV. These compounds consist of the substitution of the original AZT molecule, which is formed by a six-membered ring and a five-membered one.
5. The 31 Cramer steroid set [4a] is analyzed. This molecular set has been widely used as a benchmark in novel QSAR methods.

TABLE I.
Different Superposition Examples.



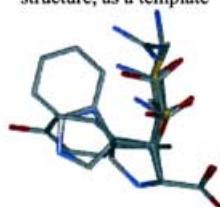
Alignment of two flavilium salts



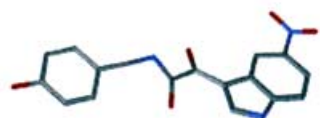
Global molecular alignment using a single structure, as a template



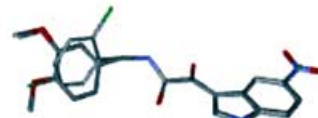
Alignment of proline and cysteine



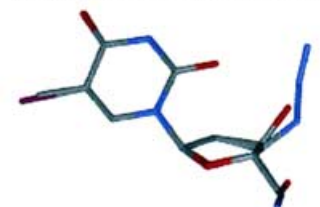
Global molecular alignment using proline as a template



Alignment of two indole derivatives



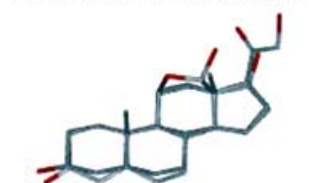
Global molecular alignment using a single structure as a template



Alignment of two AZT derivatives



Global molecular alignment using AZT as a template



Alignment of aldosterone and androstanediol.



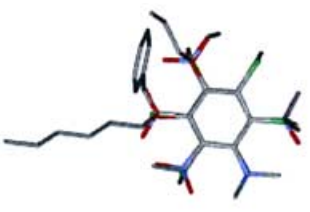
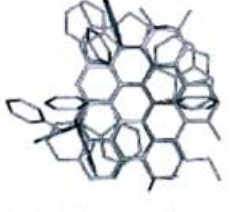
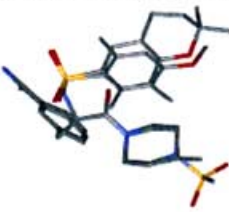
Global molecular alignment using aldosterone as a template

These compounds present a common structure of three six-membered rings and a five-membered ring.

6. A set of 33 anilide derivatives, which were submitted to a toxicity test related to the fish *Poecilia reticulata*. The molecular structures contain the common anilide group.¹⁹

7. A set of 60 poly-aromatic-hydrocarbons (PAH), which are interesting due to its toxicity.²⁰ These molecular structures, composed only by carbon and hydrogen, contain many comparable diads and triads, and thus allow testing the triad selection procedure efficiency.

TABLE I.
(Continued)

 Alignment of 2,4-dichloroanilide and 2-chloro,4-nitroanilide	 Global molecular alignment using anilide as a template
 Alignment of coronene and dibenzo[a,l]pyrene	 Global molecular alignment using coronene, as a template^a
 Alignment of two inhibitors	 Global molecular alignment using a single structure, as a template^{a,b}

The first column contains pairwise superpositions, and the second column presents alignments based on a template molecule. Hydrogen atoms have been removed for clarity.

^a For visualization purposes, only 40 molecules are shown.

^b For visualization purposes, only 20 molecules are shown.

8. A set of 88 benzamidine-type inhibitors of thrombin, trypsin and factor XA.²¹ This molecular set is built up from big size molecules with a large common skeleton composed by aromatic rings connected with substituted chains.

All these molecular sets were optimized with the AM1 Hamiltonian using an Ampac 6.01.²² The optimized geometries were fed into TGSA99 to calculate the full set of pairwise superpositions and an alignment based on a chosen template molecule. The obtained results are presented in Table I. The first column of the table presents a single example of pairwise alignment for each molecular set, and the second contains the alignment for all the compounds obtained when using a molecule as a template.

As can be seen in Table I, in both the pairwise and the template-based alignments the largest common substructures are detected and the molecules

superposed accordingly. To provide more information about the pairwise superposition, several data regarding the number of pairwise alignments, comparable diads and triads, C_{AB} , the mean time per superposition and the mean number of matched atoms per superposition are listed in Table II.

From data in Table II it can be inferred a dependence of the time needed to compute the pairwise alignment with the number of comparable diads and triads. The number of comparable diads and triads is related to the number of backbone atoms; thus, the duration of the superposition procedure obviously depends on the molecular size. It can also be seen that the quality of the superposition, evaluated by C_{AB} , is in all cases quite high. This last fact, together with the mean number of superposed atoms, states that, in all cases, the common skeleton has been detected and used to superpose the studied molecules. When a template molecule is used to generate the superposition, similar results

TABLE II. Number of Superpositions per Molecular Set and Mean Number of Comparable Diads, Triads, C_{AB} Value, Time and Matched Atoms per Pairwise Superposition.

Molecular Set	Flavilium Salts	Amino Acids	Indole Derivatives	AZT	Steroids	Aniline Derivatives	PAH	IT
Number of pairwise superpositions	120	190	253	351	465	528	1770	3828
Mean number of comparable diads ^a	475	20	396	122	456	60	346	519
Mean number of comparable triads ^a	325	14	171	88	365	54	274	312
Mean value of C_{AB}	0.944	0.896	0.993	0.991	0.990	0.948	0.951	0.983
Mean time (in seconds) per pairwise superposition ^b	0.179	0.005	0.246	0.073	0.303	0.009	0.139	1.241
Mean number of matched atoms ^a	17	6	22	14	18	7	12	23

^a Rounded to the first significant integer.

^b Average times obtained with the test machine (Silicon Graphics Origin2000/R10000/180 MHz/600 Mb RAM).

by means of comparable diads, error, times, and matched atoms are obtained. However, as this procedure requires only $N - 1$ superpositions, instead of the $N \cdot (N - 1)/2$ needed in the pairwise algorithm, the whole superposition process is $N/2$ times faster.

Conclusions

The development of TGSA has been presented, including the description of the theoretical methodological concepts and its computational application allowing a fast and user-transparent pairwise superposition of molecules. One of the main advantages of this method consists of its comprehensive simplicity, due to the fact that it is based in simple geometrical and topological concepts. Hence, its practical computational implementation is available to scientists not devoted to programming. In addition, the absence of any kind of quantum mechanical or semiempirical calculations results in a fast algorithm able to perform large amounts of alignments within reasonable time limits, as proven by the provided examples.

The practical applications, which are at present included in TGSA, have been devoted to MQSM, but the program has been structured in such a way that many other computations may be included. These applications could consist in force-field grids, volume or surface comparisons, pattern recognition, or database search, comparing a set of molecules to a given reference template.

The program can be executed within PC hardware²³ under the DOS/Windows/Linux environments, as well as in Unix Workstations. Future developments involve a graphical interface, to provide a user-friendly interface, and due to the algorithm characteristics, the parallelization of TGSA seems feasible, as no data dependencies are present within the execution. This feature would provide an even faster execution where multiprocessor architecture is available.

Acknowledgments

One of us, X. Gironés, wishes to acknowledge the University of Girona for a predoctoral fellowship.

References

1. Haneef, B. I.; Moss, D. S.; Standford, M. J.; Borkakoti, N. *Acta Crystallogr* 1985, A41, 426.
2. Kilean, R. C. G. *Acta Crystallogr* 1972, A28, 657.
3. Sippl, M. J.; Stegbuchner, H. *Comput Chem* 1991, 15, 73.
4. References devoted to 3D-QSAR models: (a) Cramer, R. D., III; Paterson, D. E.; Bunce, J. D. *J Am Chem Soc* 1988, 110, 5959; (b) Bravi, G.; Gancia, E.; Mascagni, P.; Pegna, M.; Todeschini, R.; Zalianni, A. *J Comput Aided Mol Design* 1997, 11, 79; (c) Kellogg, G. E.; Kier, L. B.; Gaillard, P.; Hall, L. H. *J Comput Aided Mol Design* 1996, 10, 513.
5. References devoted to database search: (a) Allen, F. H.; Bellard, S.; Brice, M. D.; Cartwright, B. A.; Doubleday, A.; Higgs, H.; Hummelink, T.; Humelink-Peters, B. G.; Kennard, O.; Motherwell, W. D. S.; Rodgers, J. R.; Watson, D. G. *Acta Crystallogr* 1979, B35, 2331; (b) Artymiuk, P. J.; Bath, P. A.; Grindley, H. M.; Pepperrell, C. A.; Poirrette, A. R.;

- Thorner, D. A.; Wild, D. J.; Willet, P.; Allen, F. H.; Taylor, R. *J Chem Inf Comput Sci* 1992, 32, 617; (c) Attias, R. *J Chem Inf Comput Sci* 1983, 23, 102; (d) Brint, A. T.; Willet, P. *J Chem Inf Comput Sci* 1987, 27, 152; (e) Crandell, C. W.; Smith, D. H. *J Chem Inf Comput Sci* 1983, 23, 186; (f) Cringean, J. K.; Lynch, M. F. *J Inf Sci* 1989, 15, 211; (g) Gushurst, A. J.; Nourse, J. G.; Houndshell, W. D.; Leland, B. A.; Raich, D. G. *J Chem Inf Comput Sci* 1991, 31, 447; (h) Hagadone, T. R. *J Chem Inf Comput Sci* 1992, 32, 515; (i) Rasmussen, E. M.; Downs, G. M.; Willett, P. *J Comput Chem* 1988, 9, 378; (j) Willet, P. *J Chem Inf Comput Sci* 1985, 25, 114; (k) Willet, P. *J Inf Sci* 1989, 15, 223; (l) Willet, P.; Winterman, V.; Bawden, D. *J Chem Inf Comput Sci* 1983, 26, 36; (m) Willet, P.; Winterman, V.; Bawden, D. *J Chem Inf Comput Sci* 1983, 26, 109; (n) Zhang, H.; Minoh, M.; Ikeda, K. *J Inf Proc* 1992, 15, 108.
6. Grid based superpositions references: (a) Mestres, J.; Roher, D. C.; Maggiora, J. M. *J Comput Chem* 1997, 18, 934; (b) Grant, J. A.; Gallardo, M. A.; Pickup, B. T. *J Comput Chem* 1996, 17, 1653; (c) Perkins, T. D. J.; Mills, J. E. J.; Dean, P. M. *J Comput Aided Mol Design* 1995, 9, 479; (d) Hermann, R. B.; Herron, D. K. *J Comput Aided Mol Design* 1991, 5, 511.
7. Barakat, M. T.; Dean, P. M. *J Comput Aided Mol Design* 1990, 4, 295.
8. References devoted to molecular similarity measures: (a) Constants, P.; Amat, L.; Carbó-Dorca, R. *J Comput Chem* 1997, 10, 826; (b) Parretti, M. F.; Kroemer, R. T.; Rotham, J. H.; Richards, W. G. *J Comput Chem* 1997, 18, 1344; (c) McHahon, A. J.; King, P. M. *J Comput Chem* 1997, 18, 151; (d) Manaut, M.; Sanz, F.; Jose, J.; Milesi, M. *J Comput Aided Mol Design* 1991, 5, 371; (e) Kearsley, S. K.; Smith, G. M. *Tetrahedron Comput Method* 1990, 3, 615.
9. References devoted to Graph theory: (a) Bayada, D. M.; Simpson, R. W.; Johnson, A. P.; Laurencio, C. *J Chem Inf Comput Sci* 1992, 32, 680; (b) Brown, R. D.; Jones, G.; Willett, P. *J Chem Inf Comput Sci* 1994, 34, 63; (c) Artymiuk, P. J.; Grindley, H. M.; Poirrette, A. R.; Rice, D. W.; Ujah, E. C.; Willett, P. *J Chem Inf Comput Sci* 1994, 34, 54; (d) Rarey, M.; Dixon, J. *J Comput Aided Mol Design* 1998, 12, 471.
10. PC Spartan Plus; Wavefunction Inc.: Irvine, CA.
11. Lof, P. Elsevier's Periodic Table of Elements; Elsevier Science Publishers B.V.: Amsterdam, The Netherlands, 1987.
12. See, for example: WebLab View Pro; Molecular Simulations Inc., 1998. (a trial version can be downloaded from <http://www.msi.com>.)
13. Carbó, R.; Leyda, L.; Arnau, M. *Int J Quantum Chem* 1980, 10, 1185.
14. Domingo, L.; Carbó, R. *Int J Quantum Chem* 1987, 32, 517.
15. Carbó-Dorca, R.; Besalú, E.; Gironés, X. *Adv Quantum Chem*, vol. 38, to appear.
16. Costantino, L.; Rastelli, G.; Albasini, A. *Pharmazie* 1995, 50, 573.
17. Hadjipavlou-Litina, D.; Hansch, C. *Chem Rev* 1994, 94, 1483.
18. Chu, C. K.; Schinazi, R. F.; Ahn, M. K.; Ullas, G. V.; Gu, Z. P. *J Med Chem* 1989, 32, 612.
19. Urrestarazu, E.; Vaes, W. H. J.; Verhaar, H. J. M.; Hermens, J. L. M. *J Chem Inf Comput Sci* 1998, 38, 845.
20. Roy, T. A.; Krueger, A. J.; MacKerler, C. R.; Neil, W.; Arroyo, A. M. *SAR QSAR Environ Res* 1998, 9, 171.
21. Böhm, M.; Stürzebecher, J.; Klebe, G. *J Med Chem* 1999, 42, 458.
22. Ampac 6.01, Semichem, 7128 Summit, Shawnee, KS, 1994.
23. An operative version of TGSA99 is available upon request at quantum@iqc.udg.es.

6. La Semblança Molecular Quàntica

- **Resum:** En aquest capítol es tractarà a fons la teoria de la Semblança Quàntica, així com es presentaran diverses aplicacions en el camp del QSAR/QSPR. La discussió s'inicia per un repàs històric de l'evolució de la semblança molecular quàntica fins al dia d'avui, incloent formulació matemàtica, tipus de mesures i àmbits d'aplicabilitat. Finalment es tracta l'aplicació de les mesures com a font generatriu de descriptors moleculars i la seva aplicació, conjuntament amb diversos exemples, en el camp del QSAR/QSPR.
- **Paraules Clau:** Semblança Molecular Quàntica, Mesures de Semblança, Operadors, Energia de Repulsió interelectrònica, Aplicacions, QSAR, QSPR.

6.1. *El concepte de semblança*

Dins de la consciència humana, el concepte de la semblança entre dos o varis objectes està molt arrelat. Dins de la ment existeixen múltiples mecanismes associatius, els quals ens permeten trobar similituds i característiques comunes entre diferents cossos. De ben segur, tothom alguna vegada ha fet servir frases de l'estil 'tens una retirada a ...', 'això em recorda aquell dia que ...' o 'tots els ... són iguals' per a referir-se a unes característiques físiques coincidents entre dues persones, dues situacions que transcorren de mateixa anàloga o comportaments compartits per un col·lectiu, respectivament.

En un marc més elaborat com és la geometria, la semblança entre objectes es defineix en base a un factor d'escala en la mida; així per exemple, dos cercles, de radi 2 i 10 cm cadascun, no es pot dir que siguin iguals, ja que un té 25 vegades més àrea que l'altre, però tampoc es pot dir que siguin totalment diferents, ja que en essència són la mateixa figura. Llavors, se'ls anomena semblants, ja que només difereixen en la mida.

En el món científic, la semblança n'ha format part de diferents aspectes des del bon principi. En la filosofia clàssica de l'antiga Grècia es tractava d'agrupació de diferents substàncies amb característiques similars per a, partint de les seves semblances, poder elucidar de quina proporció dels quatre 'elements fonamentals' estaven formats. En el món de la química moderna, la ben coneguda Taula Periòdica dels Elements de Mendeleiev fou construïda a partir de l'estudi i la observació minuciosa de comportaments i reactivitat química similar al llarg dels elements coneguts a l'època, fins al punt de poder deixar espais lliures a nous elements abans de ser descoberts i de fer estimacions, en molts casos molt precises, de diferents propietats fisicoquímiques.

Tot i que conceptualment la semblança entre dos objectes és fàcil d'exposar, la seva quantificació és bastant més difícil, o en alguns casos impossible. Reprenent els primers exemples donats, no s'empren frases del tipus 'tens una semblança del 73 per cent amb ...' o 'el 58 per cent d'avui ha estat com ahir'. Tanmateix, en algunes disciplines científiques, la facilitat amb que diferents aspectes són tractats numèricament, permet poder arribar a la possible quantificació, no sempre exempta de dificultat, de semblances entre diversos objectes.

6.2. Semblança Molecular Quàntica

En el camp concret de la química, és més que admissible que en algun moment es plantegés la possibilitat de determinar quant semblants són dues molècules. Diverses solucions s'han proposat, sovint lligades a diferents aspectes sobre els que es desitja enfocar la similitud molecular, tals com grups funcionals o subestructures comunes. Tanmateix, els àtoms i les molècules són objectes que es regeixen per la Mecànica Quàntica, són pròpiament *objectes quàntics*, i la inclusió d'aquest aspecte dins de la definició de la semblança entre molècules hauria de ser considerada.

Amb aquest raonament, la primera mesura quantitativa de la semblança entre dues molècules, considerades com objectes quàntics, fou proposada per Carbó l'any 1980 [1]. En aquest treball, es proposà una mesura numèrica quantitativa de la semblança entre dues molècules a partir del recobriment de les seves respectives distribucions electròniques. A partir de la primera definició seguí un desenvolupament posterior en el mateix grup de recerca [2-8], recopilant-se els desenvolupaments en diferents treballs de revisió [9-12], els quals rendeixen comptes de la natura de la definició, noves mesures i avenços computacionals del càlcul, i aplicacions pràctiques. Altres grups de recerca van adoptar la semblança quàntica, enriquint-la amb els seus treballs i ampliant-ne

l'horitzó de possibles aplicacions. Alguns exemples destacables són les contribucions de Cioslowsky [13-16], Allan i Cooper [17-24] i Richards [25-36], així com d'altres publicacions puntuals de diversos autors [37-46].

Des d'un altre punt de vista, i substituint la mecànica quàntica per topologia, Herndon [47-50] va abordar també el problema de la mesura quantitativa de similitud fent una síntesi de teoria de graphs i semblança molecular. Aquest enfocament va tenir ressò en els treballs de diversos autors, com per exemple Mezey [51-61], Ponec [62-68] i d'altres [69-72].

L'impacte que ha tingut la semblança molecular ha estat notori i ha portat a la publicació de diverses monografies [73-80] i a la celebració bianual del *Girona Seminar on Molecular Similarity*, el qual va celebrar la seva cinquena edició l'any 2001 [81], per a poder compartir i rebre de primera mà els últims avenços en aquest camp.

6.3. Formulació matemàtica de la semblança molecular quàntica

En aquest apartat es descriu el gruix teòric de la semblança quàntica. Es revisaran els fonaments bàsics de la teoria: les funcions densitat i la seva inclusió en la definició general de la teoria. Posteriorment es passarà a un llistat extens de la gran diversitat de mesures de semblança que existeixen en l'actualitat, així com la seva normalització per a l'obtenció de magnituds intuïtives i numèricament tractables. Finalment, es farà un incís en l'efecte de la sobreposició en la semblança entre dos objectes.

6.3.1. Funcions densitat

El conjunt d'interaccions a que està sotmès un objecte en el món mecanoquàntic ve representat per l'operador Hamiltonià, $\hat{H}$, el qual un cop construït permet la formulació de l'equació d'Schrödinger independent del temps

$$\hat{H}\Psi = \varepsilon\Psi, \quad (1)$$

la qual permet l'obtenció dels diferents estats estacionaris que descriuen l'objecte, matemàticament representats per la funció d'ona Ψ , i les energies, ε , associades a aquests estats [82-84]. Tanmateix, la resolució d'aquesta equació planteja grans dificultats segons la formulació de $\hat{H}$ i del nombre de partícules que formin l'objecte que s'estudia, ja que Ψ depèn de les coordenades espacials i d'spin de les partícules associades.

A partir de la funció d'ona, es pot descriure la *funció densitat* ρ del sistema a partir del mòdul al quadrat de Ψ . Matemàticament:

$$\begin{aligned}\rho &= \Psi^* \Psi = |\Psi|^2 \\ &\downarrow \\ \rho(\mathbf{r}_1, \dots, \mathbf{r}_N, s_1, \dots, s_N) &= \Psi^*(\mathbf{r}_1, \dots, \mathbf{r}_N, s_1, \dots, s_N) \Psi(\mathbf{r}_1, \dots, \mathbf{r}_N, s_1, \dots, s_N) = |\Psi(\mathbf{r}_1, \dots, \mathbf{r}_N, s_1, \dots, s_N)|^2\end{aligned}\quad (2)$$

on N és el nombre de partícules composant l'objecte.

La funció densitat també depèn de les N coordenades espaials i les N d'spin; tanmateix, la seva dimensionalitat és reduïda integrant sobre diverses coordenades. Una expressió reduïda de la densitat és la densitat de primer ordre, àmpliament representada gràficament en el capítol 3, que correspon a la integració de totes de coordenades d'spin i de totes les coordenades espaials menys la d'una partícula, un electró en el cas d'una molècula. Matemàticament:

$$\rho(\mathbf{r}_1) = \int \dots \int \rho(\mathbf{r}_1, \dots, \mathbf{r}_N, s_1, \dots, s_N) d\mathbf{r}_2 \dots d\mathbf{r}_N ds_1 \dots ds_N, \quad (3)$$

que correspon al concepte general de quan hom es refereix a funció densitat¹. Aquesta densitat de primer ordre, degudament representada en forma numèrica matricial, és la que s'obté de manera estàndard en els programes de càlcul com Gaussian [85] o Gamess [86].

La funció densitat s'interpreta com a una distribució de probabilitat de trobar la partícula en alguna regió de l'espai. Prenent la definició estadística de Von Neumann [82], la densitat es normalitza a la unitat, tal que:

$$\int \rho(\mathbf{r}_1) d\mathbf{r}_1 = 1, \quad (4)$$

que dit d'altra manera, la probabilitat de trobar l'electró en tot l'espai és la unitat (100% de probabilitat de que l'electró *existeixi* en l'espai). D'altra banda, es pot prendre la definició de distribució de càrrega, normalitzant la densitat al nombre de partícules, o electrons N_e , tal que:

$$\int \rho(\mathbf{r}_1) d\mathbf{r}_1 = N_e, \quad (5)$$

donant compte de la concentració de la càrrega en el sistema. Anàlogament a l'anterior equació (4), la integració sobre tot l'espai dóna el nombre de partícules, altrament dit, en l'espai existeixen les

¹ Al llarg d'aquest capítol, qualsevol referència a funció densitat, o simplement ρ , es referirà a la funció densitat de primer ordre.

partícules components del sistema. Aquesta última normalització és l'emprada al llarg d'aquest capítol en els diferents exemples d'aplicació que es mostraran.

6.3.2. Definició general

Definits els elements sobre els quals es fonamenta la semblança quàntica, les funcions densitat, es pot donar una definició de la mateixa. La semblança quàntica es tradueix matemàticament al recobriment entre dues funcions densitat, corresponents als dos objectes a estudiar, ponderades per un operador no diferencial definit positiu:

$$Z_{AB}(\Omega) = \iint \rho_A(\mathbf{r}_1) \Omega(\mathbf{r}_1, \mathbf{r}_2) \rho_B(\mathbf{r}_2) d\mathbf{r}_1 d\mathbf{r}_2, \quad (6)$$

on A i B són els objectes quàntics a estudiar, $\rho_A(\mathbf{r}_1)$ i $\rho_B(\mathbf{r}_2)$ les respectives funcions densitat, $\Omega(\mathbf{r}_1, \mathbf{r}_2)$ l'operador sota el qual es porta a terme la mesura, i $Z_{AB}(\Omega)$. Per definició, les mesures de semblança són reals i definides positives, $Z_{AB}(\Omega) \in \mathbf{R}^+$.

Donat un conjunt d' n objectes es pot definir la *matriu de semblança* $\mathbf{Z}$ ($n \times n$) com el conjunt de les mesures de semblança molecular quàntica entre parelles d'objectes, $\{Z_{ij}\}, i, j \in [1, n]$. Aquesta matriu, tal i com s'exposarà en un apartat posterior, serveix d'entrada de descriptors moleculars als mètodes estadístics per a la creació de models QSAR.

Segons l'expressió que adopti l'operador de (6), les mesures de semblança poden ser de diferents tipus, tal i com s'exposa en el següent apartat.

Val a dir que el càlcul de la semblança quàntica, i sobretot en funció de l'operador escollit, pot arribar a ser molt costós computacionalment. Segons el problema que es vulgui tractar (veure el següent apartat sobre àmbits d'aplicació) es podrà optar per l'ús de densitats ab initio, amb el qual els càlculs poden ser computacionalment molt cars, o densitats ajustades [87-89], les quals simplifiquen notablement el procés.

6.3.3. Tipus de mesures de semblança molecular quàntica

Mesura de tipus recobriment (*overlap-like*) [1]. L'operador adopta la forma d'una distribució de delta de Dirac:

$$\begin{aligned}
Z_{AB} &= \iint \rho_A(\mathbf{r}_1) \delta(\mathbf{r}_1 - \mathbf{r}_2) \rho_B(\mathbf{r}_2) d\mathbf{r}_1 d\mathbf{r}_2 \\
&\quad \downarrow \\
Z_{AB} &= \int \rho_A(\mathbf{r}) \rho_B(\mathbf{r}) d\mathbf{r}
\end{aligned} \tag{7}$$

Aquesta mesura aporta informació sobre la densitat compartida, o en termes menys precisos però més gràfics, el *volum compartit* entre ambdues funcions densitat.

Mesura de tipus Coulomb (*Coulomb-like*) [11]. En aquest cas es pren la forma de l'operador de repulsió electrònica:

$$Z_{AB} = \iint \rho_A(\mathbf{r}_1) |\mathbf{r}_1 - \mathbf{r}_2|^{-1} \rho_B(\mathbf{r}_2) d\mathbf{r}_1 d\mathbf{r}_2. \tag{8}$$

Aquesta mesura dóna un valor relacionat amb la repulsió coulombiana entre les dues distribucions electròniques. D'aquesta manera, està associada a interaccions electrostàtiques.

Mesura de triple densitat (*triple density-like*) [7]. En aquesta mesura, l'operador consisteix en una tercera funció densitat:

$$Z_{AB,C} = \iiint \rho_A(\mathbf{r}_1) \rho_C(\mathbf{r}_3) \rho_B(\mathbf{r}_2) d\mathbf{r}_1 d\mathbf{r}_2 d\mathbf{r}_3. \tag{9}$$

Aquesta mesura relaciona el volum compartit entre tres funcions densitat, permetent així mesurar la semblança entre dos objectes (A i B), considerant un tercer objecte (C) que actua de referència. Així, donat un conjunt de n objectes, es poden construir n matrius de semblança de triple densitat, on cada membre del conjunt actua d'operador.

Mesures de semblança sintonitzades (*tuned*). La particular estructura de les mesures de semblança permet la combinació de diferents operadors simultanis en el càlcul, ja no de la mesura de semblança puntual entre dos objectes, sinó de la matriu de semblança final. D'aquesta manera es pot obtenir una nova matriu de semblança [90,91]:

$$\mathbf{Z} = \sum_{i=1}^m c_i \mathbf{Z}(\Omega_i), \tag{10}$$

on les $\mathbf{Z}(\Omega_i)$ són matrius de semblança obtingudes a partir de definicions anteriors, els c_i són els coeficients de l'expansió, i m és el nombre de matrius emprades. Per a mantenir les propietats de les matrius de semblança i per a una interpretació més clara dels resultats s'imposen les següents condicions sobre els coeficients:

$$c_i > 0, \forall i \wedge \sum_{i=1}^m c_i = 1. \tag{11}$$

Aquestes *condicions de convexitat* permeten associar directament un percentatge de contribució a un valor c_i , fent-se més fàcil la interpretació de la matriu resultant. Teòricament, i donades m matrius de semblança inicials, poden existir infinites combinacions de coeficients per a obtenir infinites noves matrius de semblança; tanmateix, el càlcul dels coeficients en l'aplicació pràctica de les matrius de semblança sintonitzades en QSAR està orientat a la maximització de la capacitat de predicció del model.

Mesures de semblança d'energia cinètica (*kinetic-like*) [92]. No només les funcions densitat electròniques poden incloure's en el marc de la semblança quàntica definit a (6), sinó de manera general, qualsevol funció definida positiva. La funció densitat de tipus cinètic, $K(\mathbf{r})$, es pot deduir a partir de:

$$K(\mathbf{r}) = \sum_i w_i |\nabla \varphi_i(\mathbf{r})|^2, \quad (12)$$

i igualment ser emprades en mesures de semblança, emprant:

$$Z_{AB} = \iint K_A(\mathbf{r}_1) \Omega(\mathbf{r}_1, \mathbf{r}_2) K_B(\mathbf{r}_2) d\mathbf{r}_1 d\mathbf{r}_2, \quad (13)$$

utilitzant qualsevol dels operadors anteriorment exposats, i incloent-se, si escau, dins de mesures sintonitzades.

Aquesta mesura de semblança, emprant l'operador de recobriment, també aporta informació sobre el *volum compartit* entre ambdues distribucions. Tot i que ha sigut emprada amb èxit relatiu en QSAR [93,94], la seva aplicació encara es troba en fase de desenvolupament.

Mesures de semblança amb densitats de segon ordre [95]. No només s'han emprat densitats de primer ordre en els càlculs de semblança, sinó que també s'han utilitzat densitats intraculars i extraculars.

Mesures d'autosemblança molecular (*self-similarity*). Aquest tipus particular de mesures resulten, de manera independent de la forma de l'operador, d'emprar a banda i banda de la integral de semblança la mateixa funció, comparant-se l'objecte amb sí mateix. De manera general, s'expressen com:

$$Z_{AA}(\Omega) = \iint \rho_A(\mathbf{r}_1) \Omega(\mathbf{r}_1, \mathbf{r}_2) \rho_A(\mathbf{r}_2) d\mathbf{r}_1 d\mathbf{r}_2, \quad (14)$$

formant la diagonal de la matriu de semblança $\mathbf{Z}(\Omega)$. Les mesures d'autosemblança proporcionen informació sobre la concentració de la càrrega dins de l'objecte emprat. En el cas concret de l'ús de l'operador de recobriment, aquesta magnitud és mesurable experimentalment [96,97].

Mesures d'autosemblança de fragments [98-100]. Aquestes mesures prenen la mateixa forma que el cas anterior, llevat que no es mesura l'autosemblança per a tota la molècula, sinó només de determinades subestructures d'interès, p.e. el grup carboxílic d'un àcid orgànic. A la pràctica, en les mesures de fragments es pren la part de la densitat que pertoca al fragment d'interès com a densitat d'entrada en el càlcul.

Mesures d'autosemblança sobre densitats de Forats de Fermi. Tot i que més ampliada a l'apartat 6.8, aquest nou tipus de mesura d'autosemblança cobreix la plaça vacant entre les dues mesures anteriors. El càlcul directe de l'autosemblança engloba la totalitat de la molècula, donant magnituds molt elevades en presència d'àtoms pesats. D'altra banda, les mesures restringides a fragments, eliminen aquesta inconveniència al separar el fragment d'interès del total de la densitat molecular. Tanmateix, aquesta separació és *artificial* en el sentit en que, tot i que la resta de la molècula ja ha influït sobre la densitat del fragment a estudiar, l'esquelet molecular resta ja descartat. Les mesures de d'autosemblança sobre densitats de Forats de Fermi, els qual es promitgen al voltant de la regió d'interès, incrementen el paper del fragment rellevant, mantenint encara la molècula sencera. Com s'ha comentat, aquesta mesura, així com un exemple d'aplicació, es troba ampliat en l'apartat 6.8.

En la literatura existeixen diverses comparacions entre valors de mesures de semblança generades a partir de l'ús de diversos operadors [101,102].

6.3.4. Normalització de les mesures de semblança

El valor de la mesura de semblança entre dos objectes per si sol poca informació conceptual aporta sobre la similitud dels dos objectes comparats. A fi de fer més intuïtiva, i numèricament més tractable, la mesura de semblança, s'han definit una sèrie, val a dir que bastant extensa, de transformacions matemàtiques, també anomenades *índexs de semblança quàntica*. D'entre aquests, els més usats són l'índex de Carbó i el de distància euclidiana.

L'índex de Carbó, definit com a:

$$C_{AB} = \frac{Z_{AB}}{(Z_{AA}Z_{BB})^{1/2}}, \quad (15)$$

és l'índex més emprat en la transformació de mesures de semblança. L'índex està definit en l'interval existent entre zero i la unitat, donant el tant per 1 de semblança entre els dos objectes estudiats, sent més semblants els dos objectes quan més proper és el valor de C_{AB} a 1. És trivial veure com si s'empra una mesura d'autosemblança, com la definida a l'equació (14), el valor que pren $C_{AB} = 1$, o en altres paraules, un objecte és idèntic a sí mateix.

Geomètricament, l'índex de Carbó s'interpreta com el cosinus de l'angle que formen les dues funcions densitats dels objectes A i B en l'espai n -dimensional. Matemàticament:

$$C_{AB} = \frac{\langle \rho_A | \rho_B \rangle}{\|\rho_A\| \|\rho_B\|} = \cos \alpha_{AB}, \quad (16)$$

d'aquí la seva denominació d'índex de classe C .

Donada una matriu de semblances $\mathbf{Z}$, aquesta es pot transformar a matriu d'índexs Carbó, $\mathbf{C}$, aplicant la transformació (15) als elements de $\mathbf{Z}$.

L'índex de distància euclidiana, expressat com a:

$$D_{AB} = (Z_{AA} + Z_{BB} - 2Z_{AB})^{1/2}, \quad (17)$$

també és força emprat en transformades de semblança. Aquest índex està definit en l'interval $[0, \infty)$, sent més semblants els objectes quan més proper estigui l'índex a zero. Anàlogament a l'índex anterior, si els objectes comparats són idèntics, el valor de l'índex serà exactament zero.

Geomètricament, aquest índex s'interpreta com a la norma de la diferència entre les dues funcions densitat dels objectes comparats:

$$\begin{aligned} D_{AB}^2 &= \|\rho_A\|^2 + \|\rho_B\|^2 - 2\langle \rho_A | \rho_B \rangle = \|\rho_A - \rho_B\|^2 \\ &\quad \downarrow \\ D_{AB} &= \|\rho_A - \rho_B\| \end{aligned} \quad (18)$$

d'aquí el nom d'índex de distància² per la seva analogia amb la norma d'una distància euclidiana.

Existeixen altres definicions d'índexs a la literatura [28,41,103], tot i que no seran emprats en aquest treball.

² A títol de curiositat, el títol original del treball seminal de la semblança [1] era 'How distant is a molecule to another?', tanmateix els referees van jutjar més apropiada la seva forma actual.

6.3.5. Dependència de la sobreposició molecular en les mesures de semblança

Tot i que s'ha aprofundit en el concepte i metodologies de sobreposició en el capítol anterior, només fer un incís en la forta dependència que presenta el valor de la mesura de semblança amb la orientació relativa dels objectes comparats [104,105].

Com a solució a la problemàtica que pot comportar la manca de criteris per a dur a terme la sobreposició, a l'Institut de Química Computacional s'han desenvolupat dos mètodes d'alineament molecular:

- L'algorisme de sobreposició per màxima semblança.
- L'algorisme topo-geomètric.

L'algorisme de sobreposició per màxima semblança [105] considera com a òptim aquell alineament que proporciona un valor màxim de la mesura de semblança per a un operador concret. Així, donats dos objectes i un operador, la mesura de semblança resultant serà:

$$Z_{AB}(\Omega; \Theta) = \max_{\Theta} \iint \rho_A(\mathbf{r}_1) \Omega(\mathbf{r}_1, \mathbf{r}_2) \rho_B(\mathbf{r}_2; \Theta) d\mathbf{r}_1 d\mathbf{r}_2, \quad (19)$$

sent Θ el conjunt de translació-rotacions necessari per a sobreposar de manera òptima l'objecte B sobre A , considerat com a fix en l'espai.

A la pràctica, la búsqueda del màxim de semblança es fa realitzant una recerca intensiva, la qual involucra multitud de sobreposicions intermitjtes, en les quals s'avalua la mesura de semblança. Aquella sobreposició que proporciona la màxima semblança es pren com l'òptima.

Una de les conseqüències d'aquest mètode és la seva tendència a sobreposar àtoms pesats, aquells enllà del tercer període, degut a que aquests concentren bona part de la densitat molecular total. Aquesta sobreposició porta a alineaments poc intuïtius, ja que l'estructura comuna entre les molècules no suposa prou densitat comparada amb les corresponents als àtoms grans, i com a resultat no resulta rentable la seva sobreposició. A fi d'evitar aquestes sobreposicions amb poc *sentit químic*, es va desenvolupar el mètode topo-geomètric, discutit al capítol anterior i breument presentat a continuació.

De manera radicalment oposada, l'algorisme *topo-geomètric* [106] considera que la sobreposició òptima és aquella que alinea les molècules en base a un esquelet comú, considerant tan sols els tipus atòmics i les distàncies d'enllaç. Donades dues molècules, la sobreposició segons aquest criteri és única i no depenent de la mesura a realitzar posteriorment.

Apart dels mètodes desenvolupats a l'Institut i relacionats amb les mesures de semblança, existeixen altres treballs relacionats amb altres problemes. Per exemple, en determinació de farmacòfors [107], la comparació quantitativa de l'estereoquímica molecular o les mesures de distorsió en cristalls [108], o el reconeixement de patrons en bases de dades estructurals tridimensionals [109].

Els primers procediments d'alineament molecular s'inicien als anys 70, amb les contribucions de Gabuzzo [110] i McLachan [111], continuat després per Gerber i Müller [112] i Redington [113]. Aquests treballs estan basats en la minimització de mesures d'error com a distàncies entre posicions atòmiques de les molècules a alinear. Altres mètodes prenen els valors de propietats fisicoquímiques com a base de les sobreposicions, tals com simulacions de camps estèrics o electroestàtics calculats en xarxes tridimensionals [114-116] o ajustats analíticament [117,118]. També s'han emprat índexs de semblança com a base de sobreposició, utilitzant mètodes de Monte Carlo [119] o de gradient [120]. Comentar també que s'ha considerat la flexibilitat molecular [121], així com la sobreposició de molècules dissemblants [122].

6.4. Àmbits d'aplicabilitat

La definició d'una mesura quantitativa de semblances entre molècules, o en general objectes quàntics pels que es pot definir una funció densitat, obre un ampli ventall d'aplicacions en diferents disciplines computacionals. D'entre elles, les més destacades són:

- *Generació de descriptors moleculars pel seu ús en QSAR*, tal i com es descriurà amb profunditat en un apartat posterior.
- *Anàlisi de distribucions de càrrega*. La semblança quàntica es capaç de reflectir la concentració de càrrega en diferents espècies, o en la mateixa calculada a diferents nivells, a partir de la comparació de llurs densitats electròniques. Aquestes anàlisis permeten la classificació de funcionals DFT [123], quantificar paràmetres DFT [124] o determinar l'efecte de la solvatació [125], entre d'altres aplicacions.
- *Determinació d'un mètode òptim d'optimització de geometria*. Donada una estructura molecular determinada experimentalment, es poden optimitzar models d'aquesta estructura a diferents nivells teòrics (ab initio, DFT, AM1, ...) amb diferents bases, i posteriorment calcular la semblança entre les densitats obtingudes pels models i les obtingudes de càlculs puntuals (o densitats ajustades) sobre l'estructura original.

Aquells models que més se semblin a l'original seran bons candidats³ per a modelar estructures derivades de l'original.

- *Avaluació de propietats atòmiques dins les molècules* [126-128], a partir de càlculs d'autosemblança entre els diferents àtoms constituents d'una molècula.
- *Priorització d'orbitals per a càlculs d'interacció de configuracions truncats* [129,130], a partir de càlculs de semblança entre orbitals atòmics i moleculars.
- *Determinació de tipus d'estats de transició* [131,132] a partir de mesures de semblança dels reactius i dels productes d'una reacció amb el respectiu estat de transició. Aquests càlculs permeten elucidar quant propera està una barrera de reacció als productes (product-like) o als reactius (reactant-like), permetent-se una quantificació del postulat de Hammond [133], així com detectar-ne excepcions (comportament anti-Hammond).
- *Quantificació dels efectes del substituent en la reactivitat química* [98-100] a partir de mesures d'autosemblança del grup funcional implicat en el procés.
- *Modelització de propietats en nuclis atòmics* [134-136], a partir de mesures de semblança entre funcions densitat nuclears [137,138]. Tot i que en la química quotidiana es consideren els nuclis com a càrregues puntuals, aquests realment són objectes volumètrics als quals se'ls pot associar una funció densitat.

6.5. Semblança Molecular Quàntica en QSAR

Un cop establertes les bases de la semblança molecular quàntica, aquesta fou considerada una eina vàlida per a la generació de descriptors molecular per al seu ús en QSAR. El primer treball on s'apliquen mesures de semblança es remunta a l'any 1983 amb el treball de Martín [139], on es va fer un estudi qualitatiu de modelització de l'activitat molecular emprant potencials electroestàtics, amb els quals, i seguint en la mateixa línia de treball, es van aportar diferents contribucions en anys posteriors [140-144]. De manera similar, Richards [145] utilitzà valors de l'índex Carbó obtinguts per comparació de potencials electroestàtics com a descriptors moleculars per a la construcció de models de predicció. De manera similar, Richard [146] va fer servir potencials electroestàtics com a base de càlcul de semblances pel posterior ús en QSAR.

Incloent conceptes de la teoria de graphs, Rum i Herndon [147,148] construïren una matriu d'índexs de semblança molecular, els quals estaven compresos entre 0 i 1, molècules totalment

³ No s'especifica aquell mètode que dona la màxima semblança com a òptim, ja que cal considerar altres factors com viabilitat del càlcul o grandària del conjunt de molècules a modelar.

diferents o idèntiques respectivament, les columnes de les quals s'empraren com a descriptors moleculars en un model de regressió multilíneal.

Good va descriure un protocol d'aplicació de les mesures de semblança en QSAR molt similar al emprat actualment [149-151]. Els valors continguts a les matrius provenien descriptors de la forma molecular i de comparacions de potencials electroestàtics, els quals es calculaven aproximant-los a funcions gaussianes [117,152], o bé a partir de discretitzar-los en forma de xarxa de punts rectangular tridimensional. Els resultats del càlculs es submetien a una adaptació discreta de l'índex Carbó. En aquests treballs s'aplicaren sistemes estadístics de reducció de dimensions, via PLS o xarxes neurals, i validació estadística dels resultats del models de predicció.

Inicialment, la semblança molecular només s'emprava com a eina qualitativa en els estudis QSAR, interpretant les diferents agrupacions de punts-molècula en funció d'alguna activitat molecular d'interès [153,154]. Posteriorment es postulà l'existència d'un operador que relacionava les mesures de semblança amb la manifestació de l'activitat molecular [155]. La implementació d'aquests conceptes i principis es traduí en un protocol teòric de l'aplicació de les mesures de semblança en QSAR [156].

En l'actualitat, apart de l'ús de matrius de semblança, també es treballa amb altres formes de semblança, enumerades en l'apartat d'aplicabilitat. Algunes d'aquestes línies de recerca contemplen autosemblances moleculars [98,157], de fragments [99,100], l'energia de repulsió electró-electró [158-160], la qual s'exposarà en un apartat posterior, i els índexs topològics de semblança quàntica [161,162], els quals es troben encara en període de desenvolupament. Dins del nostre grup de recerca, la construcció de models QSAR s'ha dut a terme satisfactòriament emprant la semblança molecular quàntica dins dels àmbits farmacològic [94,155,156,161,163-166] i d'avaluació de la toxicitat [158-160,167].

6.6. Exemples d'aplicació I: matrius de semblança

En aquest apartat es descriurà l'ús de matrius de semblança molecular quàntica com a descriptors en models QSAR presentant tres exemples, els quals estan relacionats amb propietats físiques, activitats biològiques, i, com a cas particular d'activitat biològica d'especial interès, toxicitat. L'estructura d'aquest apartat comença amb el modelatge molecular, seguidament s'exposa el procés de càlcul de les MQSM, i finalment els exemples d'aplicació.

6.6.1. Modelatge Molecular

Totes les molècules presentades en aquesta secció han estat construïdes amb assistents gràfics per a l'edició i poliment inicial (de l'anglès *cleaning*) de les estructures, tipus HyperChem [168], AGUI (Ampac Graphical User Interface) [169] o WebLab Pro [170]. Un cop definides les estructures, aquestes s'importen al programa Ampac per a la seva optimització a nivell AM1 [171], o bé dins del programa PC-Spartan amb el camp de forces de mecànica molecular Sybil [172].

A fi de d'obtenir la funció densitat per a cada molècula del conjunt, s'utilitza l'aproximació promolecular, descrita àmpliament al Capítol 3, sobre les coordenades optimitzades de les molècules amb ajustos a la base 3-21G per a simplificar els càlculs.

6.6.2. Càlculs de MQSM i Exemples

Tal i com s'ha descrit a la definició general de les MQSM, a fi de calcular les matrius de semblança **Z**, cal portar a terme el procés de càlcul de la mesura de semblança per parelles. Com que **Z** és simètrica, donat un conjunt d' n molècules, caldrà realitzar $(n+1) \cdot n/2$ càlculs de semblança per a obtenir la matriu sencera. Per a cada molècula, i degut a la dependència ja descrita de la mesura amb la posició relativa de les molècules en l'espai, cal un procés de sobreposició. En els exemples que es presentaran s'ha emprat l'Algorisme de Sobreposició Topo-Geometric (TGSA) [106], breument descrit anteriorment i àmpliament detallat al Capítol 5.

Dels diferents tipus d'operadors que poden ponderar la mesura, dels presents a la definició general, s'ha emprat tan sols el de *Coulomb* (Equació (8)).

Finalment, les mesures obtingudes es transformen a Índex Carbó, segons l'equació (15).

Exemple d'aplicació I: Correlació del punt d'ebullició d'un conjunt d'alcanes

El primer exemple d'aplicació de matrius de semblança en QSAR correspon a un conjunt de 529 alcanes [173], sobre el qual s'ha avaluat el seu punt d'ebullició (b.p., de l'anglès *boiling point*). El punt d'ebullició d'alcanes, a pressió normal, ha esdevingut una propietat estàndard (de l'anglès *benchmark*) per a l'avaluació de nous models QSAR, pel que resulta interessant veure la capacitat de les MQSM a l'hora de predir-lo.

El conjunt estudiat, presentat conjuntament amb els valors dels b.p. al final d'aquest capítol degut a la seva extensió, s'ha correlacionat amb PLS fins a un total de 15 factors, derivats de realitzar les MQSM amb l'operador de Coulomb. A fi d'alleugerir la càrrega computacional de manejar una matriu de semblança de 529×529 , aquesta s'ha reduït amb dos senzills procediments: eliminació d'aquelles columnes que presenten un coeficient de variació del 2% o inferior, i eliminació de les columnes redundants, o aquelles que donen correlació amb altres columnes de més de $r^2 = 0.85$. En aquest cas, aquest procés ha reduït lleugerament la dimensionalitat de la matriu a 529×514 . L'evolució de les magnituds estadístiques es mostren gràficament a continuació:

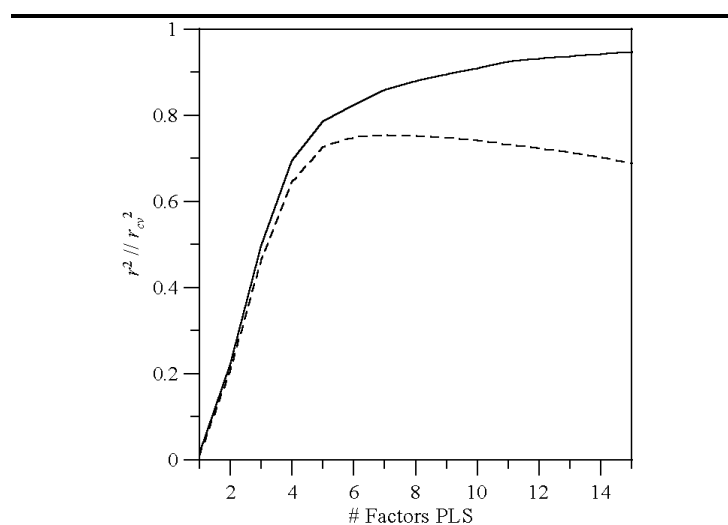


Figura 1.- Evolució de la qualitat d'ajust (r^2 , en línia contínua) i de predicció (r_{cv}^2 , en línia discontinua) amb l'increment del nombre de factors PLS utilitzats.

Tal i com es pot veure a l'anterior Figura 1, la gràfica presenta un ràpid augment de l'ajust i capacitat de predicció amb l'addició de nous factors fins al cinquè, augmenta lentament fins al setè, a partir del qual el creixement s'inverteix lentament. Així, com a compromís entre nombre de descriptors emprats i capacitat de predicció, es tria 5 factors com a balanç òptim, portant al següent model de predicció:

$$\text{b.p.} = 6.966\mathbf{f}_1 + 37.19\mathbf{f}_2 + 76.40\mathbf{f}_3 + 99.81\mathbf{f}_4 + 75.15\mathbf{f}_5 \quad (20)$$

$$n = 529$$

$$r^2 = 0.785$$

$$\sigma = 20.20$$

$$r_{cv}^2 = 0.727$$

Probabilitat de correlació aleatòria: inferior a 10^{-100}

Random Test, 100 permutacions

$$r_{mitjà}^2 = 0.302$$

$$r_{cv, mitjà}^2 = 0.024$$

$$r_{màx}^2 = 0.403$$

$$r_{cv, màx}^2 = 0.176$$

Els resultats del model de predicció, així com els del Random Test, es poden veure gràficament a les Figures 2 i 3 respectivament:

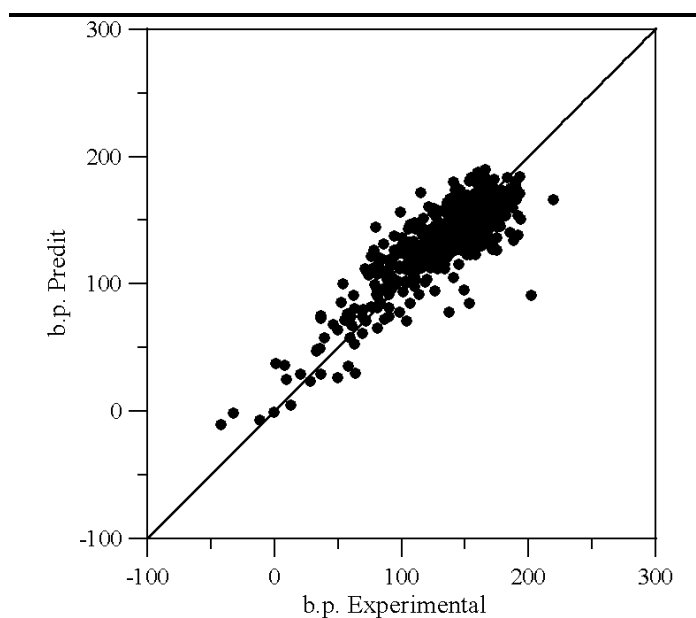


Figura 2.- Gràfic de valors de b.p. experimentals contra predits per validació creuada per a un sistema de 529 alcans [173].

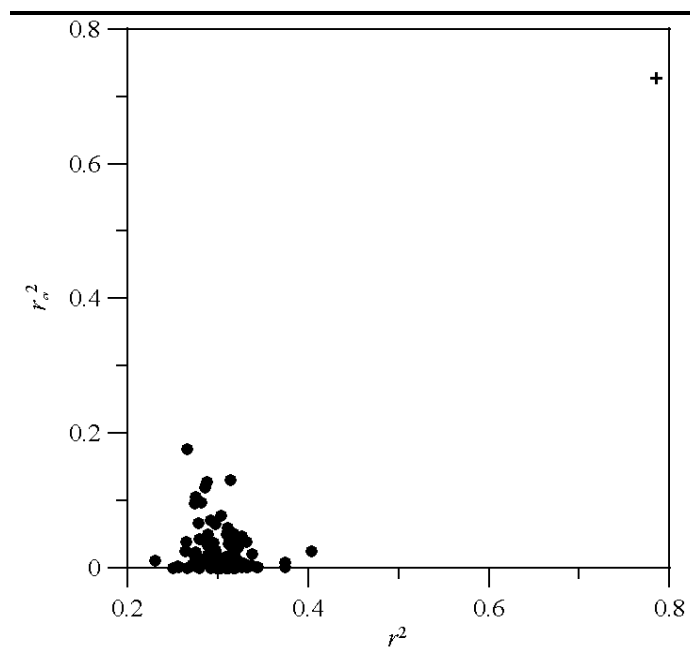


Figura 3.- Random Test per a un sistema de 529 alcans [173]. (+) Valor original, (●) Valors Permutats.

Tal i com es pot veure a la Figura 2, s'aconsegueix un model amb una bona capacitat de predicció. Els valors predits per validació creuada es troben, excepte per uns pocs compostos infraestimats, raonablement propers als experimentals. La Figura 3 mostra gràficament els resultats del Random Test, el qual no deixa lloc a dubtes sobre la total absència de correlació fortuïta en el model proposat.

Exemple d'aplicació II: Correlació de l'activitat antimalarial d'un conjunt de peroxy-cetals cíclics

En aquest segon exemple es presenta l'estudi de correlació d'un conjunt de 20 peroxy-cetals cíclics amb la seva respectiva activitat antimalarial [174], expressada com a aquella concentració nanomolar de compost (en escala logarítmica) capaç d'inhibir el 50% de la síntesi i reducció de la hidrofolat reductasa ($\log IC_{50}$) en el paràsit *Plasmodium Falciparum* [175].

En aquest cas, la tècnica estadística emprada per a tractar la matriu de semblances, calculada amb l'operador de Coulomb, ha estat el PCA, construint-se models fins a 7 components principals (PCs). La tria de 7 PCs respon a emprar un filtre a la variança associada a cada component, el qual ha estat fixat al 3%, reduïnt-se la dimensionalitat de la matriu de 20×20 a 20×7 . L'evolució de la correlació es mostra a la següent Taula 1.

Taula 1.- Evolució dels paràmetres estadístics amb l'addició de nous PCs en el model de correlació. (en cursiva el model triat)			
PCs emprats	r^2	r_{cv}^2	% Variança acumulada
2	0.546	0.521	≈17
1,2	0.592	0.544	≈66
1,2,6	0.738	0.689	≈69
<i>1,2,3,6</i>	<i>0.778</i>	<i>0.718</i>	<i>≈77</i>
1,2,3,4,6	0.795	0.722	≈81
1,2,3,4,6,7	0.809	0.721	≈84
1,2,3,4,5,6,7	0.809	0.698	≈88

Tal i com s'infereix de l'anterior taula, l'ordre de tria dels PCs no coincideix amb l'ordre de variança explicada per cada PC. Això és degut a l'ús de la tècnica *Most Predictive Variables Method* (veure Capítol 4), la qual ordena els descriptors segons la seva correlació individual amb la propietat estudiada. Amb aquests resultats i de manera anàloga a l'exemple anterior, s'escull el model de 4 PCs com a balanç òptim entre capacitat de predicció i nombre de descriptors emprats. El model de predicció proposat, així com els paràmetres estadístics associats es mostren a la següent equació:

$$\log IC_{50} = 1.739 \cdot PC_1 - 10.18 \cdot PC_2 - 4.042 \cdot PC_3 - 11.37 \cdot PC_6 + 2.107 \quad (21)$$

$$n = 20$$

$$r^2 = 0.778$$

$$\sigma = 0.158$$

$$r_{cv}^2 = 0.718$$

Probabilitat de correlació aleatòria: $8.58 \cdot 10^{-3}$

Random Test, 100 permutacions

$$r_{mitjà}^2 = 0.197$$

$$q_{mitjà}^2 = -0.495$$

$$r_{m\grave{a}x}^2 = 0.530$$

$$q_{m\grave{a}x}^2 = 0.156$$

I els resultats gràfics de la validació creuada i del Random Test a les següents Figures 4 i 5:

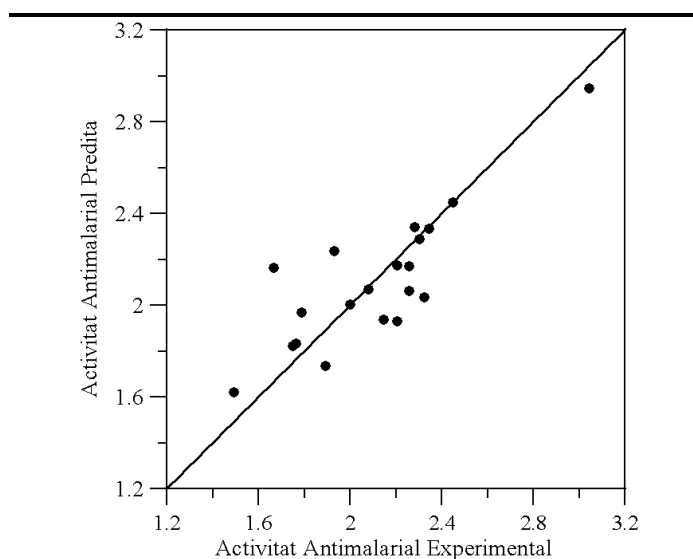


Figura 4.- Gràfic de valors de $\log IC_{50}$ experimentals contra predits per validació creuada per a un sistema de 20 peroxy cetals cíclics [174].

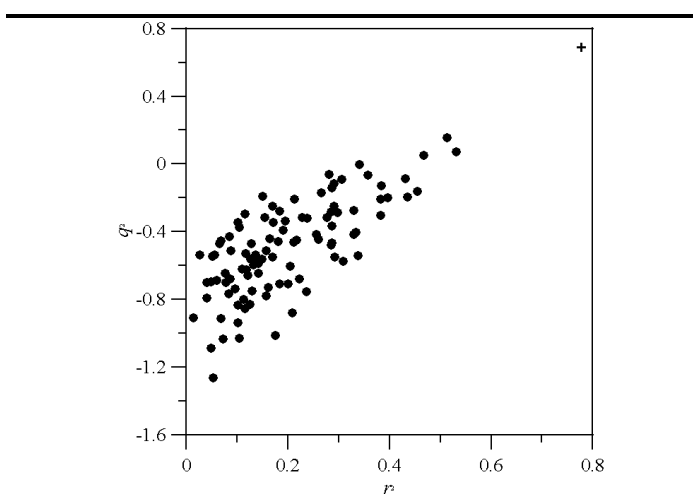


Figura 5.- Random Test per a un sistema de 20 peroxy cetals cíclics [174]. (+) Valor original, (•) Valors Permutats.

Tal i com s'observa en la representació de la validació creuada, l'activitat antimalarial és predita dins d'uns intervals força estrets, llevat per un parell de compostos, els quals són lleugerament sobreestimats. La figura del Random Test verifica la validesa tan del model com de la metodologia en aquest sistema, no donant cap mena de correlació significativa en els models aleatoris.

Exemple d'aplicació III: Correlació de la toxicitat aquàtica d'un conjunt de derivats del benzè

En aquest últim exemple es presenta l'aplicació de les mesures de semblança en el camp de la descripció de la toxicitat molecular. L'estudi de toxicitats, com a casos particulars d'activitat biològica, ha esdevingut recentment un camp de gran interès degut fonamentalment a la implantació de noves directrius medioambientals i a una creixent conscienciació amb l'entorn. La síntesi de nous productes d'utilitat industrial o comercial sovint ha portat el problema de la seva interacció amb el medi un cop han acomplert la seva funció. Els estudis teòrics de modelatge molecular poden ser d'ajuda en aquest camp des de dos punts de vista complementaris: ajudar a maximitzar la propietat o activitat desitjada del compost i intentar minimitzar l'impacte amb l'entorn.

L'exemple que es mostra consta d'un conjunt de 92 derivats benzènics on s'ha avaluat la seva toxicitat aquàtica en termes de $-\log EC_{50}$ [176], o concentració nanomolar de compost necessària per a reduir una població de *Poetilia Reticulata*, un petit peix molt sensible als canvis de les condicions aquàtiques, al 50% en 96 hores. De manera similar al primer exemple, s'han calculat les mesures de semblança Coulomb i la tècnica de PLS per a construir models de predicció fins a l'ús de 8 factors. L'eliminació de variables redundants, d'igual manera que en el primer exemple, ha portat a la reducció de la matriu de descriptors de 92×92 a 92×71 . L'evolució de la correlació es mostra a la següent Taula 2:

Taula 2.- Evolució dels paràmetres estadístics amb nous factors PLS en la correlació. (en cursiva el model triat)

# factors	r^2	r_{cv}^2
1	0.076	0.103
2	0.345	0.309
3	0.656	0.507
4	<i>0.733</i>	<i>0.646</i>
5	0.800	0.678
6	0.830	0.691
7	0.850	0.687
8	0.871	0.654

I el model de predicció, així com paràmetres estadístics rellevants, compostat per 4 descriptors a la següent equació:

$$-\log LC_{50} = 0.501f_1 + 1.369f_2 + 2.507f_3 + 1.351f_4 \quad (22)$$

$$n = 92$$

$$r^2 = 0.733$$

$$\sigma = 0.386$$

$$r_{cv}^2 = 0.646$$

Probabilitat de correlació aleatòria: $0.488 \cdot 10^{-21}$

Random Test, 100 permutacions

$$r_{mitjà}^2 = 0.232$$

$$r_{cv, mitjà}^2 = 0.055$$

$$r_{màx}^2 = 0.422$$

$$r_{cv, max}^2 = 0.311$$

Els resultats del model proposat, es mostren gràficament a les següents Figures 6 i 7:

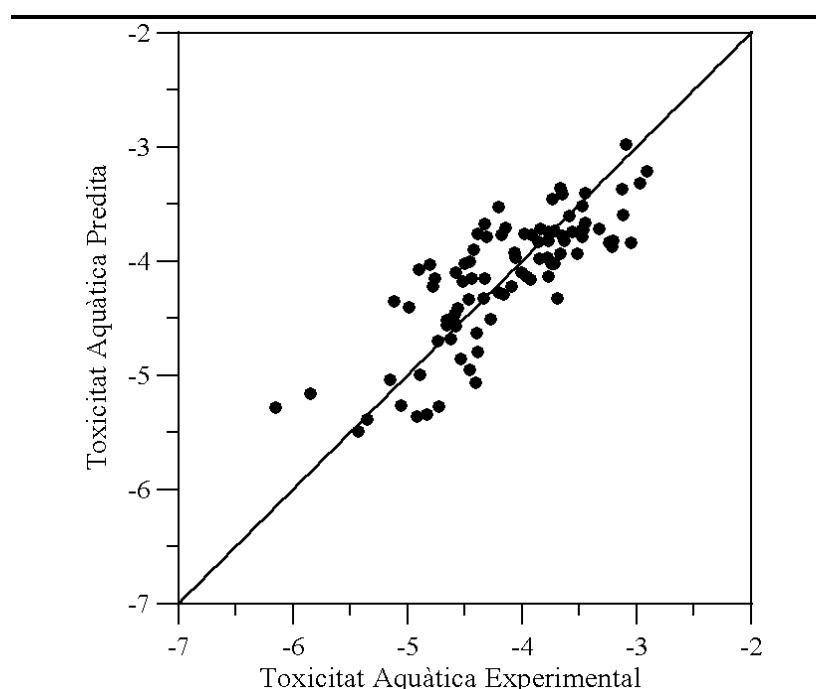


Figura 6.- Gràfic de valors de $-\log EC_{50}$ experimentals contra predits per validació creuada per a un sistema de 92 derivats del benzè [176].

Tal i com s'aprecia a la Figura 6, s'obté una notable correlació resultant de l'ús de quatre factors de PLS, sense evidenciar-se desviacions importants. En el cas concret de la toxicitat, més important que l'exacte predicció d'un valor, pot resultar més interessant poder realitzar classificacions dels compostos en base a llimars de toxicitat. D'aquesta manera, si es divideix aquest gràfic en tres regions, corresponents a baixa, mitjana i alta toxicitat, la major part dels compostos es troben correctament classificats, i aquells que es prediuen en una altra zona del gràfic,

ho fan tan sols per salts d'una regió. De manera similar als exemples anteriors, el Random Test es salva satisfactòriament, donant-se una clara separació entre els models permutats i el real.

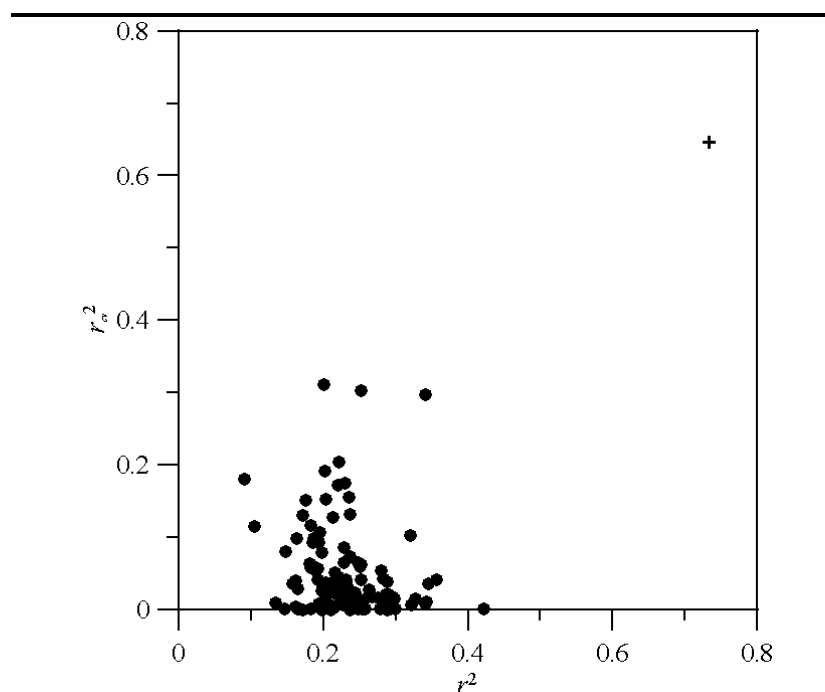


Figura 7.- Random Test per a un sistema de 92 derivats del benzè [176]. (+) Valor original, (•) Valors Permutats.

L'estudi previ [176], involucrava un conjunt inicial de 161 molècules i emprava 5 descriptors: log P, la càrrega parcial més positiva sobre un àtom d'hidrogen, la càrrega parcial més negativa sobre un àtom no-hidrogen i les energies dels orbitals frontera HOMO i LUMO. Els resultats de l'estudi eren molt satisfactoris, assolint $r^2=0.970$ i $q^2=0.930$. En aquest exemple que s'ha presentat, tan sols s'han pres les molècules que presentaven un anell benzènic comú. Un altre estudi sobre aquesta família va ser dut a terme en el nostre grup de recerca [167], en el qual s'emprava tota la matriu de semblança, utilitzant anàlisi de components principals com a tècnica estadística de correlació sobre una matriu de semblances calculada emprant l'algorisme de la màxima sobreposició. Els resultats foren força satisfactoris, arribant a $r^2=0.740$ i $q^2=0.716$ amb tres components i $r^2=0.892$ i $q^2=0.840$ amb cinc components.

6.7. Exemples d'aplicació II: V_{ee} com a descriptor molecular

En aquest apartat es descriurà l'ús de l'energia de repulsió interelectrònica com a descriptor en models QSAR. Es comença amb una breu introducció, passant al rerafons teòric, i presentant

diversos exemples, que comprenent models de correlació de propietats moleculars, activitat biològica i toxicitat, com a cas particular d'activitat biològica d'elevat interès, on s'ha aplicat satisfactòriament el mètode.

6.7.1. Introducció

L'ús de l'energia de repulsió electró-electró ($\langle V_{ee} \rangle$) pren el seu punt de sortida com a continuació de treballs previs d'Amat i Ponec [98,177], on es posava de manifest la possibilitat d'emprar mesures d'autosemblança com a alternativa al coeficient de partició octanol-aigua ($\log P$), el qual ha estat utilitzat en la creació de models de predicció de diversos conjunts moleculars (veure Capítol 4 i referències 178-181).

La intenció d'aquest nou descriptor és aportar una nova alternativa mecanoquàntica en la construcció de models QSAR clàssics (basats en el mètode de Hansch) senzills, usualment construïts a partir de paràmetres empírics. Així mateix, també es pretén assimilar $\langle V_{ee} \rangle$ dins del context teòric de la semblança molecular quàntica a partir de la seva consideració com a cas particular de mesura d'autosemblança.

6.7.2. $\langle V_{ee} \rangle$ i semblança molecular quàntica

L'expressió pel valor esperat de l'operador de repulsió interelectrònica per a un sistema de p -partícules es pot expressar com:

$$\Omega \equiv V_{ee} = \sum_{i=1}^{p-1} \sum_{j=i+1}^p \left(|\mathbf{r}_i - \mathbf{r}_j| \right)^{-1} . \quad (23)$$

En aquest cas, és àmpliament conegut que l'ús de l'operador de Coulomb proporciona un valor esperat, en el cas de trobar-se en el model monoconfiguracional de capa tancada, on les integrals de Coulomb $\{ \langle ii | jj \rangle \}$ i de Bescanvi $\{ \langle ij | ij \rangle \}$ juguen un paper predominant [182]:

$$\langle V_{ee} \rangle = \sum_i \sum_j (2 \langle ii | jj \rangle - \langle ij | ij \rangle) > 0 . \quad (24)$$

Tot i que les dues parts de l'equació (24) podrien ser utilitzades directament com a mesures d'autosemblança [2], la natura definida-positiva de l'energia de repulsió interelectrònica

mecanoquàntica, en la seva totalitat, pot també ser considerada com a una mesura de semblança, obrint la porta al possible ús de $\langle V_{ee} \rangle$ com a descriptor molecular.

L'avantatge principal de l'ús de $\langle V_{ee} \rangle$ consisteix en que el seva determinació es realitza de manera rutinària en el programari de càlcul mecanoquàntic disponible actualment, possibilitant de certa manera la *universalització* del concepte de semblança molecular quàntica a grups de recerca que no disposen de programes específics per al càlcul de mesures de semblança ordinàries.

6.7.3. Construcció de models lineals senzills amb $\langle V_{ee} \rangle$

En els exemples presentats a l'apartat anterior, s'emprava la matriu de semblances $\mathbf{Z}=\{Z_{IJ}\}$ ($n \times n$) com a font de descriptors per a la construcció dels models de predicció amb una propietat molecular, π_i , com:

$$\pi_i = \sum_j a_j Z_{ij}, \quad (25)$$

on els a_i són els coeficients de l'expansió lineal. La matriu $\mathbf{Z}$, construïda a partir de comparacions entre parelles de molècules, conté en la diagonal principal les mesures d'autosemblança, podent-se descomposar $\mathbf{Z}$ com a:

$$\pi_i = a_i Z_{ii} + \sum_{j \neq i} a_j Z_{ji}, \quad (26)$$

on s'ha aïllat les autosemblances de la resta de mesures.

Quan s'estudien sèries moleculars homòlogues, el sumatori de (26) esdevé aproximadament constant al llarg del conjunt [98,177]. Llavors, es pot intentar establir relacions lineals entre diferents propietats moleculars i les respectives mesures d'autosemblança:

$$\pi_i \cong a_i Z_{ii} + b. \quad (27)$$

Tenint en compte l'assimilació dins del marc de les autosemblances de $\langle V_{ee} \rangle$, es pot rescriure l'anterior equació (27) com a:

$$\pi_i \cong a_i \langle V_{ee} \rangle_i + b. \quad (28)$$

En els exemples que es presentaran, no es mostraran exemples gràfics de resultats del test aleatorietat, degut a que l'ús d'un sol descriptor difícilment pot portar a sobreparametrització, fet que aquest test posa en evidència. Tanmateix, els resultats d'aquest test es resumiran breument

juntament a l'equació dels models proposats en forma de mitjana dels valors de r^2 i q^2 , així com els valors màxims que s'obtenen per permutacions dels vectors de propietats estudiades.

6.7.4. Modelatge molecular

A fi de calcular la magnitud de $\langle V_{ee} \rangle$ cal construir les molècules, optimitzar-ne la geometria, i sobre l'estructura resultant, considerant fase gas, realitzar el càlcul ab initio necessari per a obtenir el valor de $\langle V_{ee} \rangle$. El programari utilitzat és:

- *Construcció de molècules i poliment inicial de l'estructura*: HyperChem [168], AGUI (Ampac Graphical User Interface) [169] o WebLabPro 3 [170].
- *Optimització de geometria*: Ampac 6 [169] amb el Hamiltonià semiempíric AM1 [171] o Gaussian 94 [183] a nivell HF/3-21G*.
- *Càlcul ab initio*: Gaussian 94 [183] a nivell HF/3-21G*.

El primer pas no és determinant del resultat, tanmateix el pas d'optimització de geometria, segons el mètode triat, pot portar a lleugeres variacions en el valor de $\langle V_{ee} \rangle$ per a la mateixa molècula. La tria d'un o altre procediment es va fer en base als recursos disponibles en el moment i en funció de la grandària de les estructures implicades i del conjunt molecular tractat (veure articles adjunts).

6.7.5. Interpretació de $\langle V_{ee} \rangle$

Tal i com es va demostrar en un treball previ [177], les mesures d'autosemblança es poden utilitzar com a una alternativa al descriptor empíric $\log P$ en estudis QSAR. De manera similar $\langle V_{ee} \rangle$ pot constituir-ne una alternativa viable per a la construcció de models de predicció senzills en sèries moleculars homòlogues, així com a un complement a d'altres descriptors de qualsevol natura en una anàlisi de Hansch.

$\langle V_{ee} \rangle$ es pot interpretar com a un descriptor fonamentalment dels efectes estèrics, mida molecular, ja que la seva magnitud està fortament lligada al nombre d'àtoms i el seu tipus. Així mateix, $\langle V_{ee} \rangle$, degut a la seva natura mecanoquàntica, és també capaç de reflexar efectes

electrònics, ja que les interaccions intramoleculars estan parcialment connectades a l'energia de repulsió interelectrònica.

No és la intenció d'aquest treball la creació d'un descriptor capaç de relacionar universalment qualsevol propietat molecular, sinó mostrar la possibilitat d'assimilació de diferents magnituds dins del marc de la semblança molecular quàntica. A més a més, el fet que el seu càlcul sigui rutinari dins del programari disponible, permet la construcció d'un primer model QSAR tan bon punt els càlculs sobre el conjunt molecular han conclòs.

Per acabar de perfilar la connexió de $\langle V_{ee} \rangle$ amb la semblança quàntica, es presenta un model de correlació entre $\langle V_{ee} \rangle$ i $\log P$ per a un conjunt d'amides. Les dades molecular, d'energia de repulsió i $\log P$ es donen a la següent Taula 3, i el model obtingut a l'equació (29). A fi de fer més llegibles els models subsegüents, els valors de $\langle V_{ee} \rangle$ es troben normalitzats i simbolitzats per $\langle V_{ee} \rangle^{(N)}$, o expressat matemàticament: $\langle V_{ee} \rangle_i^{(N)} = (\langle V_{ee} \rangle_i - \langle \bar{V}_{ee} \rangle) \sigma^{-1}$, on $\langle \bar{V}_{ee} \rangle$ és la mitjana dels valors de $\langle V_{ee} \rangle$ pel conjunt molecular, i σ n'és la desviació estàndard.

Taula 3.- Estructures, $\langle V_{ee} \rangle$ i valors de $\log P$ experimentals [177] i predits per a un conjunt d'amides.

Estructures	$\langle V_{ee} \rangle$	$\log P$	$\log P$ per validació creuada
Formamida	131.33	-1.51	-1.46
Acetamida	194.51	-1.26	-0.99
Propilamida	264.41	-0.66	-0.55
Butilamida	337.14	-0.21	-0.05
Pentilamida	413.94	0.33	0.49
Hexilamida	493.73	0.87	1.10
N-metilformamida	191.63	-0.97	-1.09
N-etilformamida	258.85	-0.43	-0.63
N-propilformamida	330.24	0.11	-0.14
N-butilformamida	405.91	0.65	0.34

$$\log P = 0.746 \langle V_{ee} \rangle^{(N)} - 0.308 \quad (29)$$

$$n = 10 \quad r^2 = 0.955 \quad \sigma = 0.197 \quad q^2 = 0.931$$

S'observa una bona correlació, $r^2=0.955$, entre els valors de $\langle V_{ee} \rangle^{(N)}$ i el $\log P$, confirmant un relació real entre ambdues magnituds. En l'anterior estudi [177] es va aconseguir també una

bona correlació, $r^2=0.954$, emprant autosemblances de recobriment octanol/aigua, la qual s'obté a partir de comparar densitats calculades en entorns octanol i aigua, respectivament, obtenint estructures, i consecuentment, distribucions electròniques diferents. L'afinitat entre ambdós models mostra que els dos paràmetres contenen la mateixa informació estructural.

Així mateix, aquest exemple també posa de manifest l'estreta relació entre la mida molecular i $\langle V_{ee} \rangle$, ja que, com es pot veure a la Taula 3, el valor de $\langle V_{ee} \rangle$ augmenta al augmentar la mida molecular. Així mateix, $\langle V_{ee} \rangle$ també capta diferències subtils en la connectivitat: la acetamida i la N-metilformamida tenen la mateixa fórmula empírica, però diferents tipus d'enllaços, i aquest fet es manifesta en el valor de $\langle V_{ee} \rangle$, el qual pren 194.51 i 191.63, respectivament.

6.7.6. Exemple: $\langle V_{ee} \rangle$ com a descriptor de la refractivitat molar

El primer exemple que es presenta d'aplicació de $\langle V_{ee} \rangle$ correspon a la descripció d'una propietat molecular, l'índex de refractivitat molar, MR (de l'anglès *Molar Refractivity*) per a un conjunt de 183 molècules orgàniques senzilles, les quals inclouen alcans lineals i cíclics, alguns i compostos oxo-substituïts [184]. La refractivitat molar és una mesura de la capacitat de les distribucions electròniques d'una molècula de desviar feixos lluminosos. Les dades estructurals, així com els valors de la MR , es reproduïxen al final del capítol.

El model obtingut per a aquest conjunt molecular, així com les dades estadístiques pertinents, s'exposen a la següent equació:

$$MR = 8.994\langle V_{ee} \rangle^{(N)} + 31.86 \quad (30)$$

$$n = 183$$

$$r^2 = 0.900$$

$$\sigma = 2.88$$

$$q^2 = 0.897$$

Probabilitat de correlació aleatòria: $0.825 \cdot 10^{-33}$

Random Test, 100 permutacions

$$r_{mitjà}^2 = 6.962 \cdot 10^{-3}$$

$$q_{mitjà}^2 = -16.10 \cdot 10^{-33}$$

$$r_{m\grave{a}x}^2 = 0.071$$

$$q_{m\grave{a}x}^2 = 0.047$$

Gràficament, es poden veure els resultats de la validació creuada a la següent Figura 8:

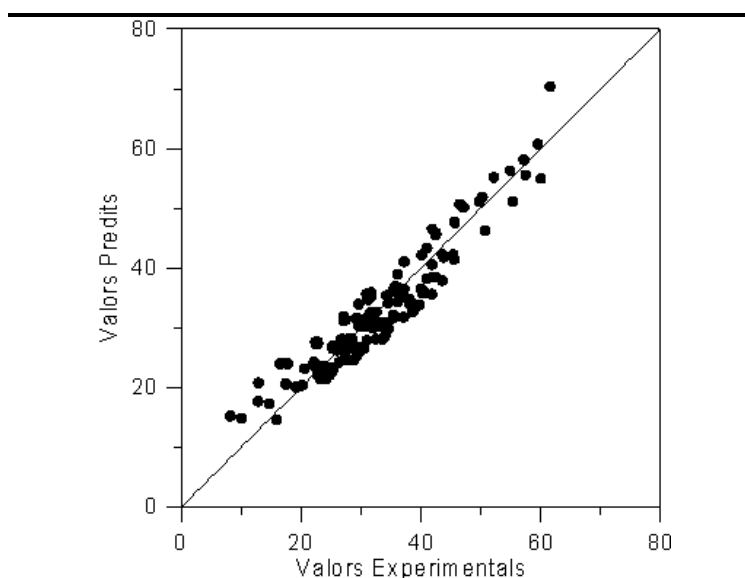


Figura 8.- Representació gràfica dels valors experimentals contra els predits per validació creuada.

En aquest exemple, $\langle V_{ee} \rangle$ proporciona una correlació acceptable amb la *MR*. La distribució principal de les dades està en funció de la mida molecular, ja que la *MR* creix a mesura que el volum augmenta, però dins de molècules de volum similar, les diferències en la connectivitat es posen de manifest en el descriptor.

L'estudi previ [184] comparava el significat fisicoquímic de diferents paràmetres topològics. Els resultats eren similars, $r^2=0.824$, emprant l'índex de primer ordre de connectivitat molecular [185-187], o considerablement pitjors quan s'utilitzava l'índex de connectivitat de segon ordre [185-187], $r^2=0.672$, o $\log P$, $r^2=0.640$, com a descriptors moleculars.

Aquest exemple, degut a la gran quantitat de molècules present, permet fer un assaig de la capacitat real de predicció del model. Per a realitzar el test, el conjunt total de molècules s'ha dividit en dos grups de 122 (conjunt d'entrenament, 2/3 del conjunt original) i 61 (conjunt de test, 1/3 del conjunt original) molècules respectivament, triades a l'atzar dins del conjunt original. El model de predicció es calcula per al conjunt d'entrenament, i a partir de l'equació que s'obté, es prediuen les *MR* per al conjunt de test. Al no intervenir per a res les molècules de test en la construcció del model de regressió, les prediccions obtingudes són totalment reals. El model que s'obté pel conjunt d'entrenament es mostra a la següent equació:

$$MR = 9.077 \langle V_{ee} \rangle^{(N)} + 31.50 \quad (31)$$

$$n = 122$$

$$r^2 = 0.896$$

$$\sigma = 3.02$$

$$q^2 = 0.890$$

Probabilitat de correlació aleatòria: $0.273 \cdot 10^{-27}$

Random Test, 100 permutacions

$$r_{mitjà}^2 = 9.078 \cdot 10^{-3}$$

$$q_{mitjà}^2 = -24.77 \cdot 10^{-3}$$

$$r_{màx}^2 = 0.054$$

$$q_{màx}^2 = 0.025$$

Els resultats de la regressió no són substancialment diferents que en model anterior amb tot el conjunt molecular. Tot i que la correlació i la capacitat de predicció han disminuït lleugerament, es continua disposant d'un model acceptable i vàlid per a efectuar les prediccions del conjunt de test. Els resultats de les prediccions es poden veure gràficament a la següent Figura 9:

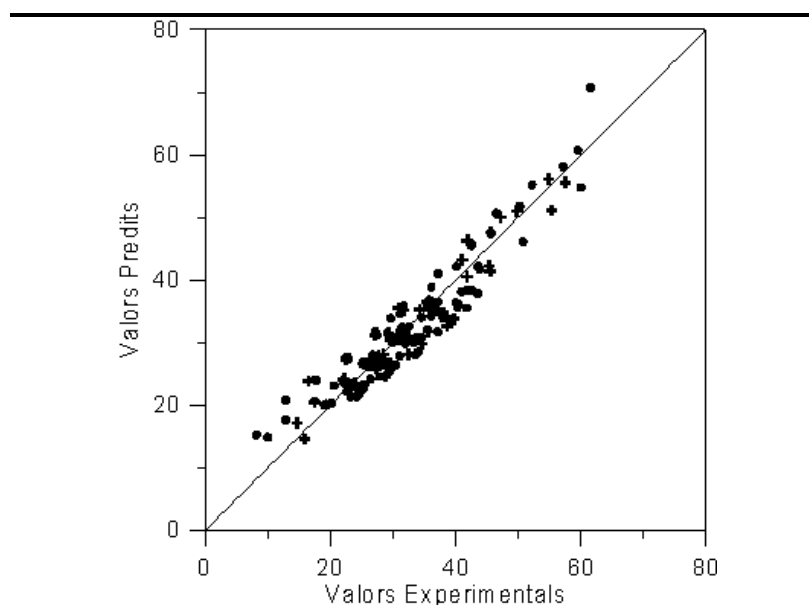


Figura 9.- Representació gràfica dels valors experimentals contra els valors predits per les 122 molècules d'entrenament per validació creuada (●) i les 61 molècules de test veritablement predites (+).

Com es pot veure, els resultats de les prediccions són acurats dins d'un marge d'error petit, ajustant-se força bé als valors experimentals. Al calcular la correlació entre els valors predits i els originals, el resultat és de $r^2=0.91$, confirmant la validesa i estabilitat del model creat emprant $\langle V_{ee} \rangle$ com a descriptor molecular.

6.7.7. Exemple: $\langle V_{ee} \rangle$ com a descriptor de la inhibició de creixement

Aquest exemple introdueix la capacitat de $\langle V_{ee} \rangle$ com a descriptor de activitats biològiques. En aquests cas, l'activitat estudiada és la inhibició del creixement de poblacions de *Tetrahymena Pyriformis* [188], un ciliat sovint emprat per a mesures d'ecotoxicitat.

L'activitat es mesura en termes de $-\log GC_{50}$, o concentració de compost, en escala logarítmica, necessari per a inhibir el creixement de la població al 50%.

El conjunt molecular estudiat, format per 21 alcohols i 9 amines, així com les dades de $\langle V_{ee} \rangle$ i activitat molecular, es mostren a la següent Taula 4:

Taula 4.- Estructures, $\langle V_{ee} \rangle$ i valors de $\log P$ experimentals [188] i predits per a un conjunt d'amides.

Estructures	$\langle V_{ee} \rangle$	$-\log GC_{50}$	Estructures	$\langle V_{ee} \rangle$	$-\log GC_{50}$
Metanol	81.22	-2.77	3-pentanol	350.49	-1.33
Etanol	135.16	-2.41	2-metil-1-butanol	350.05	-1.13
Propanol	196.61	-1.84	3-metil-1-butanol	346.51	-1.13
Butanol	264.22	-1.52	3-metil-2-pentanol	356.16	-1.08
Pentanol	336.77	-1.12	(tert)pentanol	359.82	-1.27
Hexanol	413.71	-0.47	(neo)pentanol	359.80	-0.96
Heptanol	494.15	0.02	Propilamina	188.54	-0.85
Octanol	577.89	0.50	Butilamina	255.95	-0.70
Nonanol	664.56	0.77	Pentilamina	328.48	-0.61
Decanol	753.89	1.10	Hexilamina	405.20	-0.34
Undecanol	845.56	1.87	Heptilamina	485.68	0.10
Dodecanol	939.47	2.07	Octilamina	569.33	0.51
Tridecanol	1035.39	2.28	Nonilamina	656.08	1.59
2-propanol	200.66	-1.99	Decilamina	745.25	1.95
2-pentanol	347.32	-1.25	Unidecilamina	836.97	2.26

El model QSAR obtingut, així com els respectius paràmetres estadístics, es mostren a la següent equació (32):

$$-\log GC_{50} = 1.367 \langle V_{ee} \rangle^{(N)} - 0.258 \quad (32)$$

$$n = 30 \quad r^2 = 0.924 \quad \sigma = 0.415 \quad q^2 = 0.909$$

Probabilitat de correlació aleatòria: $0.363 \cdot 10^{-14}$

Random Test, 100 permutacions

$$r_{mitjà}^2 = 50.45 \cdot 10^{-3} \quad q_{mitjà}^2 = -0.136 \cdot 10^{-3} \quad r_{màx}^2 = 0.307 \quad q_{màx}^2 = 0.209$$

Els resultats es poden veure gràficament a la següent Figura 10, on es representen els valors experimentals contra els predits per validació creuada:

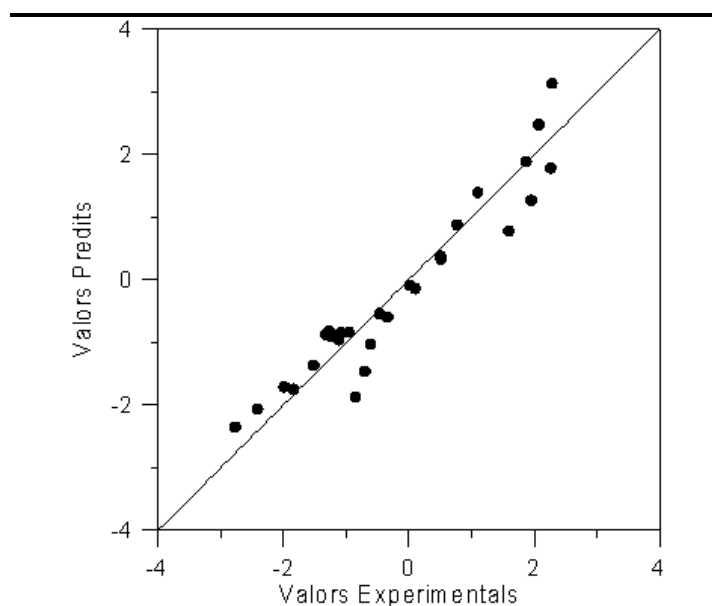


Figura 10.- Representació gràfica dels valors experimentals contra els valors predits per les 30 molècules.

Tal i com es pot veure a la figura anterior, el sistema queda descrit de manera satisfactòria emprant $\langle V_{ee} \rangle$ com a únic descriptor molecular, obtenint-se prediccions força afins a les dades experimentals. Si es jutgen les dades estadístiques derivades del model proposat, $r^2=0.924$ i $q^2=0.909$, la correlació obtinguda relaciona de manera vàlida l'activitat estudiada a $\langle V_{ee} \rangle$.

En aquest exemple, $\langle V_{ee} \rangle$ es comporta principalment com a descriptor dels efectes estèrics, discriminant bàsicament la mida molecular. D'altra banda, la inclusió d'efectes electrònics en $\langle V_{ee} \rangle$ permet una diferenciació entre diferents isòmers estructurals i entre les funcionalitzacions amino i hidroxil.

Tanmateix, l'ús de $\log P$ en aquest sistema milloraria els resultats de correlació, tal i com es mostrava en l'estudi previ [188], assolint-se valors de $r^2=0.951$ (en front a 0.924) i $q^2=0.943$ (en front a 0.909). Aquest estudi fa intuir que, tot i que $\log P$ es comporta com un descriptor de la mida molecular en sèries homogènies, la hidrofobicitat pot tenir una important rellevància en aquest sistema.

6.7.8. Exemple: $\langle V_{ee} \rangle$ com a descriptor de toxicitat

A fi d'il·lustrar el comportament de $\langle V_{ee} \rangle$ com a descriptor de la toxicitat, es presenta com a exemple un conjunt de 36 benzens substituïts, subconjunt de l'exemple anterior de toxicitat presentat en la secció 6.6, sobre els quals s'ha avaluat la seva toxicitat aquàtica [176]. La toxicitat, expressada en termes de $-\log EC_{50}$, correspon a la concentració nanomolar capaç de exterminar al 50% d'una població de *P. Reticulata* (un peix de mida petita) en 96 hores. Les dades dels compostos d'aquesta sèrie, així com el valor de $\langle V_{ee} \rangle$ i toxicitat per a cada compost, es presenten a la següent Taula 5:

Taula 5.- Dades estructurals, de $\langle V_{ee} \rangle$ i de toxicitat relatives a un conjunt de 36 derivats del benzè.

Estructura	$\langle V_{ee} \rangle$	$-\log EC_{50}$	Estructura	$\langle V_{ee} \rangle$	$-\log EC_{50}$
Clorobenzè	563.87	-3.77	2-xylè	443.42	-3.48
1,2-diclorobenzè	890.67	-4.40	3-xylè	439.54	-3.45
1,3-diclorobenzè	874.41	-4.28	4-xylè	438.61	-3.48
1,4-diclorobenzè	871.07	-4.56	Nitrobenzè	557.51	-2.97
1,2,3-triclorobenzè	1243.04	-4.89	2-nitrotoluenè	660.82	-3.59
1,2,4-triclorobenzè	1222.56	-4.83	3-nitrotoluenè	651.68	-3.65
1,3,5-triclorobenzè	1209.43	-4.74	4-nitrotoluenè	649.25	-3.67
1,2,3,4-tetraclorobenzè	1616.76	-5.35	2,3-dimetilnitrobenzè	765.26	-4.39
1,2,3,5-tetraclorobenzè	1599.44	-5.43	3,4-dimetilnitrobenzè	753.58	-4.21
1,2,4,5-tetraclorobenzè	1595.48	-5.85	2-cloronitrobenzè	895.94	-3.72
3-clorotoluenè	653.96	-3.84	3-cloronitrobenzè	876.00	-4.01
4-clorotoluenè	652.23	-4.33	4-cloronitrobenzè	871.45	-4.42
2,4-diclorotoluenè	983.16	-4.54	2,3-dicloronitrobenzè	1255.93	-4.66
2,4,5-triclorotoluenè	1343.82	-5.06	2,4-dicloronitrobenzè	1234.64	-4.46
3,4-diclorotoluenè	991.52	-4.60	2,5-dicloronitrobenzè	1235.61	-4.59
Pentaclorobenzè	2015.06	-6.15	3,5-dicloronitrobenzè	1219.06	-4.58
Benzè	277.72	-3.09	2-cloro-6-nitrotoluenè	999.56	-4.52
Toluenè	355.54	-3.13	4-cloro-2-nitrotoluenè	990.12	-4.44

El model obtingut, amb les dades estadístiques rellevants, es mostra a continuació a l'equació (33):

$$-\log EC_{50} = -0.688 \langle V_{ee} \rangle^{(N)} - 4.309 \quad (33)$$

$$n = 36$$

$$r^2 = 0.889$$

$$\sigma = 0.241$$

$$q^2 = 0.877$$

Probabilitat de correlació aleatòria: $0.814 \cdot 10^{-15}$

Random Test, 100 permutacions

$$r_{mitjà}^2 = 30.15 \cdot 10^{-3}$$

$$q_{mitjà}^2 = -91.56 \cdot 10^{-3}$$

$$r_{màx}^2 = 0.288$$

$$q_{màx}^2 = 0.221$$

A la següent Figura 11 es presenten gràficament els resultats de la validació creuada:

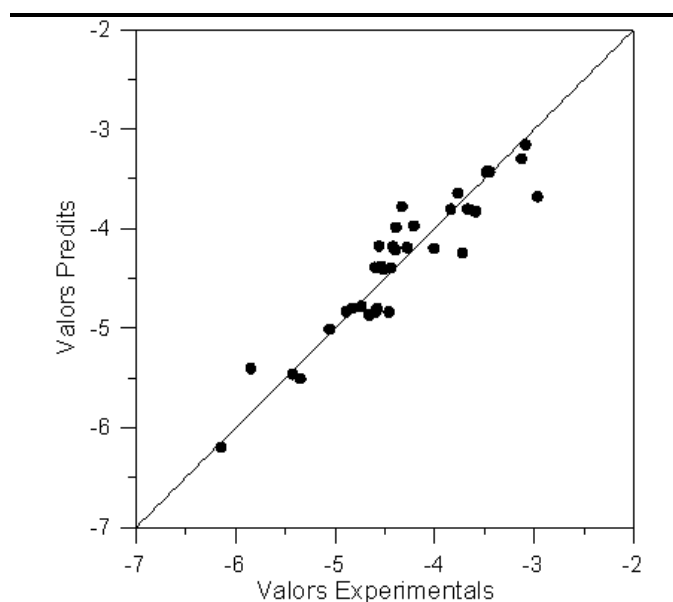


Figura 11.- Representació gràfica dels valors experimentals contra els valors predits pels 36 derivats del benzè.

A partir de l'anàlisi de validació creuada, es confirma que $\langle V_{ee} \rangle$ correlaciona correctament la toxicitat aquàtica d'aquest conjunt, donant un bon índex de correlació i una elevada capacitat de predicció.

Tal i com s'ha comentat en l'exemple de l'apartat anterior, l'estudi previ [176] assolia molt bons resultats $r^2=0.970$ i $q^2=0.930$, emprant 5 descriptors ($\log P$, la càrrega parcial més positiva sobre un àtom d'hidrogen, la càrrega parcial més negativa sobre un àtom no-hidrogen i les energies dels orbitals frontera HOMO i LUMO) sobre un conjunt de 161 molècules. En aquest exemple que s'ha presentat, tan sols s'han pres les molècules que presentaven un anell fenòlic comú.

En el cas de prendre $\log P$ com a únic descriptor d'aquest sub-conjunt, els resultats s'empobreixen considerablement respecte als obtinguts amb $\langle V_{ee} \rangle$, descendant a $r^2=0.694$ i $q^2=0.660$.

A partir dels resultats obtinguts, es pot considerar que els efectes estèrics, però en força mesura els electrònics, prenen part important en la toxicitat aquàtica per aquesta sèrie estudiada. Aquesta afirmació es pot recolzar en els descriptors emprats en el treball original [176], els quals llevat de $\log P$, són tots de natura electrònica.

6.8. Exemple d'aplicació III: Semblances sobre densitats de Forats de Fermi

6.8.1. Funcions Densitat de Forats de Fermi

Els anomenats Forats de Fermi promitjats en un domini (de l'anglès domain-averaged Fermi Holes) foren introduïts [189,190] com a un concepte útil per a la interpretació i visualització de l'estructura molecular. Discutits en profunditat en treballs previs [189-192], en aquesta secció es resumiran les idees bàsiques necessàries per aquests tipus d'estudis.

Aquests Forats de Fermi es defineixen segons la fórmula següent:

$$g_{\theta}(\mathbf{r}_1) = N_{\theta}\rho(\mathbf{r}_1) - 2 \int \rho(\mathbf{r}_1, \mathbf{r}_2) d\mathbf{r}_2, \quad (34)$$

on N_{θ} és el nombre mitjà d'electrons en la regió (o domini) definit per θ , matemàticament expressat com a:

$$N_{\theta} = \int \rho(\mathbf{r}) d\mathbf{r}, \quad (35)$$

i $\rho(\mathbf{r}_1, \mathbf{r}_2)$ és la densitat bielectrònica. I $g_{\theta}(\mathbf{r})$ esdevé una nova funció densitat, la qual es troba majoritàriament definida al voltant del domini θ . Aquests forats estan íntimament lligats a les magnituds $F(\theta, \theta')$, definides com :

$$F(\theta, \theta') = \int g_{\theta}(\mathbf{r}) d\mathbf{r}, \quad (36)$$

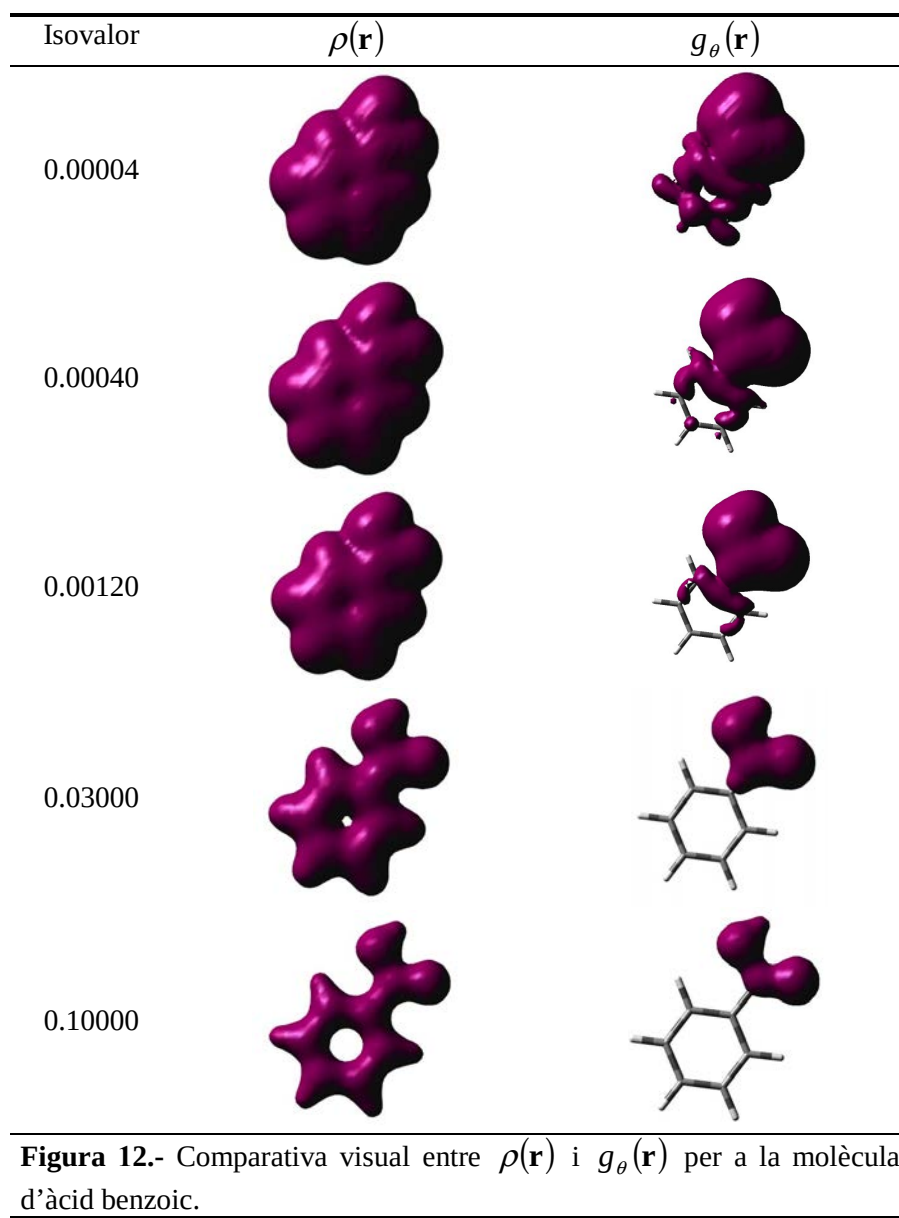
les quals tot i definides en primera instància per Bader [193], la utilitat de les magnituds $g_{\theta}(\mathbf{r}_1)$ per a elucidacions estructurals fou recentment descrita [191,192].

Una característica interessant d'aquestes densitats $g_{\theta}(\mathbf{r})$ és que aconsegueixen la condició universal de normalització:

$$\int g_{\theta}(\mathbf{r}) d\mathbf{r} = N_{\theta}. \quad (37)$$

Tot i que aquesta normalització es manté invariant respecte la mida i la forma de la regió θ , hi ha alguns dominis, la rellevància dels quals en química és de particular importància. Aquests dominis poden identificar-se, per exemple, amb les conques de Bader en la partició virial de la funció densitat electrònica. Aquest procediment es troba detallat en treballs previs [189-192], els quals involucren el pas crucial *isopycnic transformation* [194] del forat (34).

Així, aquesta nova densitat $g_{\theta}(\mathbf{r})$, derivada de $\rho(\mathbf{r})$, es comporta matemàticament igual que la densitat electrònica, llevat que es troba definida en una regió molecular θ . A l'igual que $\rho(\mathbf{r})$, $g_{\theta}(\mathbf{r})$ pot ser representada gràficament i incorporada dins del marc de la semblança quàntica. A títol il·lustratiu, es presenta una comparació visual d'ambdues densitats, calculades a nivell AM1 [171] amb el programa Gaussian [183] sobre geometries optimitzades a nivell AM1 amb l'Ampac [169], per a la molècula d'àcid benzoïc, on en el cas de $g_{\theta}(\mathbf{r})$ s'ha definit la regió $\theta = \text{COOH}$, a diferents isovalors de densitat a la següent figura:



Tal i com es pot veure a l'anterior figura, $\rho(\mathbf{r})$ es comporta de manera regular, disminuint el volum del contorn a mesura que augmenta l'isovalor representat, passant de mostrar-nos característiques merament volumètriques a un patró d'enllaços i posicions atòmiques tal i com s'ha vist extensament en el Capítol 3. D'altra banda, $g_\theta(\mathbf{r})$, definida en la regió del grup carboxílic, ens mostra una forma totalment diferent. La presència d'un electró promitjat en aquesta regió col·lapsa la densitat al voltant d'aquest domini, tal i com es veu isovalors elevats, mentre que per la resta de l'anell benzènic, aquesta densitat es distribueix de manera irregular. Considerant que $g_\theta(\mathbf{r})$ és la distribució en l'espai d'un electró influenciat per la presència promitjada d'un altre electró en el domini θ , aquest electró lliure tendeix majoritàriament a aparellar-se amb l'electró promitjat. A partir d'aquest concepte es pot interpretar la distribució de $g_\theta(\mathbf{r})$ al voltant de l'anell a partir de les formes ressonants de l'àcid benzoïc, tal i com es mostren a la següent figura:

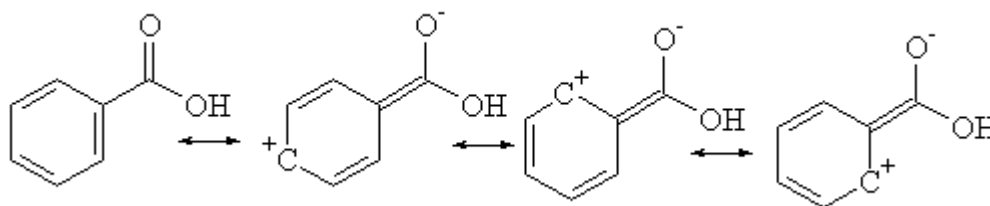
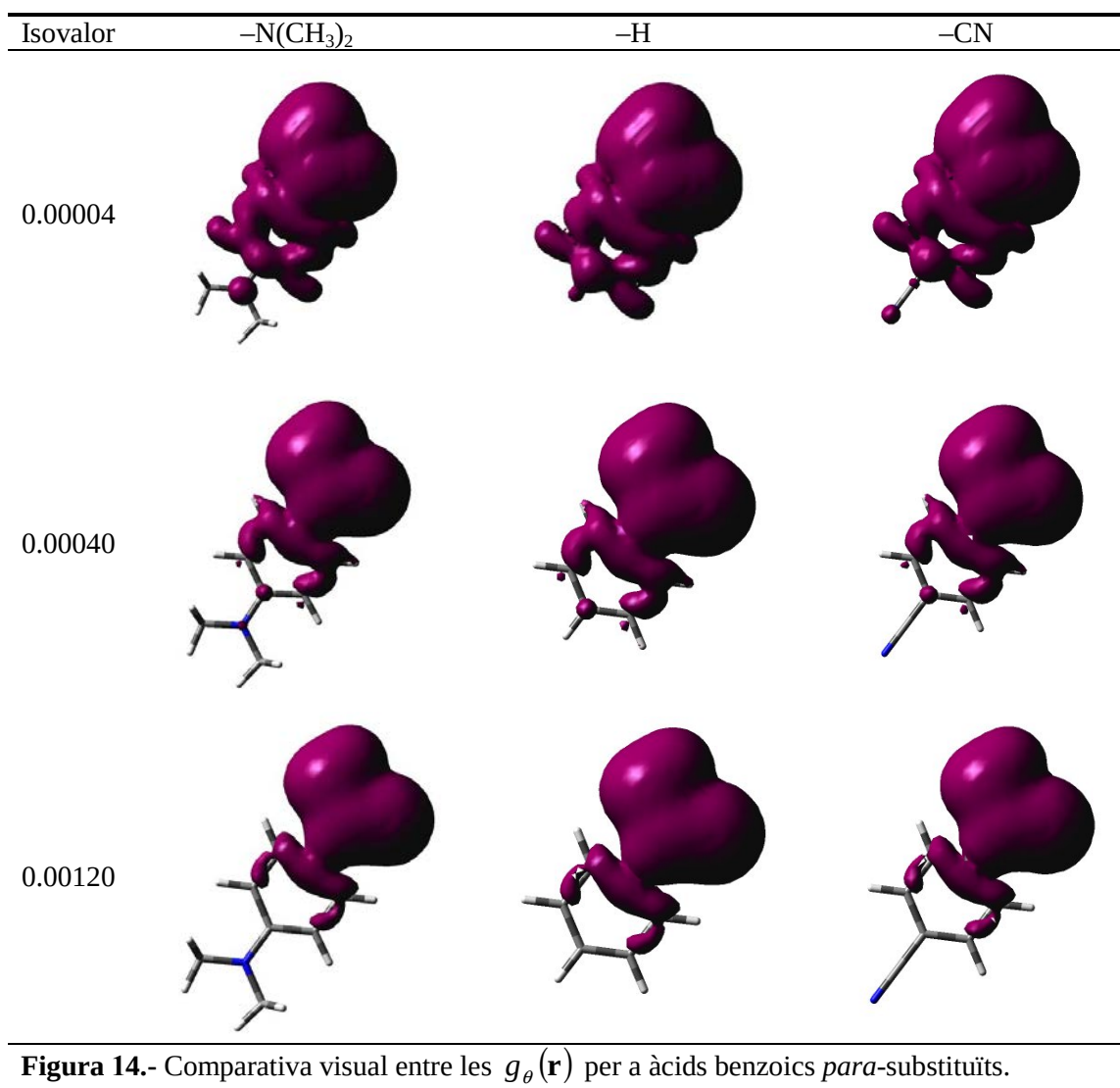


Figura 13.- Formes ressonants de l'àcid benzoïc.

Així es pot observar com $g_\theta(\mathbf{r})$ es distribueix sobre les regions *orto* i *para* de l'anell, sient la *para* aquella més estable per la màxima separació de càrregues. Tal i com es pot veure, tot i que la regió θ està definida fora de l'anell, aquest exerceix una gran influència sobre la regió fixada, i seria d'esperar que substitucions en aquest anell es reflectissin també sobre $g_\theta(\mathbf{r})$, tal i com es pot veure a la següent Figura 8 amb les substitucions $-\text{N}(\text{CH}_3)_2$ i $-\text{CN}$ en posició *para*.

Tal i com s'observa (amb certa dificultat) a la figura 14, la presència d'un grup electró-donador, com és el $-\text{N}(\text{CH}_3)_2$, incrementa la densitat, o concentració de càrrega, en el grup COOH , mentre que un grup electró-atractor, com és el $-\text{CN}$, la retira. Ja que la concentració de càrrega pot quantificar-se amb mesures de semblança, s'obre una possible via per a modelar els efectes electrònics del substituent, tradicionalment modelats amb la sigma de Hammett (σ) [195,196] tal i com s'ha tractat al Capítol 4, emprant semblança.



6.8.2. Mesures de Semblança sobre Densitats de Forats de Fermi

Tal i com s'ha vist en l'apartat anterior, $g_\theta(\mathbf{r})$ es comporta com qualsevulla funció densitat, podent-se incloure dins del marc de la semblança quàntica. Si es suposa la mesura d'autosemblança més senzilla, com a l'expressió (14), suposant l'operador de recobriment $\Omega(\mathbf{r}_1, \mathbf{r}_2) = \delta(\mathbf{r}_1 - \mathbf{r}_2)$, l'expressió (14) es simplifica a:

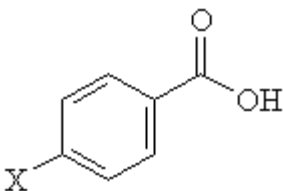
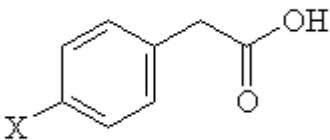
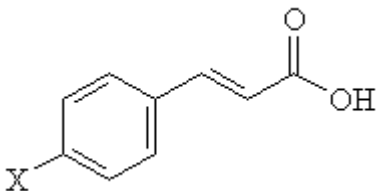
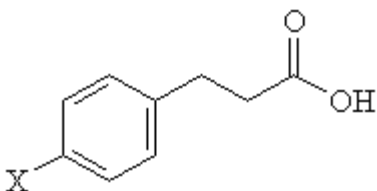
$$Z_{AA}^\theta(\Omega) = \int g_{A;\theta}(\mathbf{r}) g_{A;\theta}(\mathbf{r}) d\mathbf{r}, \quad (38)$$

permetent obtenir diferents mesures d'autosemblança molecular quàntica (MQS-SM) per a la mateixa molècula segons com es defineixi el domini θ .

6.8.3. Exemple d'aplicació: Àcids Aromàtics

Com a aplicació en el camp de QSAR, s'intentarà modelar la sigma de Hammett (σ) [195,196] per a 4 sèries d'àcids aromàtics que es mostren a la següent taula. Les molècules han estat modelades a nivell AM1 [171] amb el programa Gaussian [183] sobre geometries optimitzades a nivell AM1 amb l'Ampac [169], com en l'exemple de representacions gràfiques anteriors. S'ha triat el mètode AM1 per la seva excel·lent relació entre cura i velocitat de càlcul en front als *ab initio*, tanmateix, tot i que s'ha emprat l'Ampac per a realitzar l'optimització de geometria, s'ha dut a terme el càlcul puntual posterior amb el Gaussian, el qual treballa amb funcions gaussianes en lloc de funcions d'Slater, per la facilitat amb que comporta treballar amb gaussianes.

Taula 6.- Estructures i pendents relatius de 4 sèries d'àcids aromàtics.

Tipus	Estructura	ρ
I		1.0 (per definició)
II		0.56
III		0.42
IV		0.21

on ρ és la relació entre el pendent d'una sèrie d'àcids substituïts i el pendent per àcids benzoics substituït (veure Capítol 4). Les substitucions estudiades (X a la Taula 6) són les següents:

Taula 7.- Substituents i sigmes de Hammett per a 12 substituents.

-X	σ
-NO ₂	0.81
-CN	0.71
-CF ₃	0.53
-CCl ₃	0.46
-Br	0.22
-Cl	0.22
-F	0.06
-H	0.00 (per definició)
-CH ₃	-0.14
-CH ₂ CH ₃	-0.13
-OCH ₃	-0.28
-N(CH ₃) ₂	-0.63

Si s'aplica l'equació (38), els valors resultants de la qual es troben a la Taula 8, construint les $g_\theta(\mathbf{r})$ per a la sèrie de 12 àcids benzoics *para*-substituïts, i definint θ en la regió del grup àcid carboxílic, s'obté el següent model de correlació:

Tipus I

$$r^2 = 0.9597$$

$$\rho/\rho_0 = 1 \text{ (per definició)}$$

$$Z_{AA}^\theta = -0.0205 \cdot \sigma + 5.534, \quad (39)$$

el qual, representat gràficament:

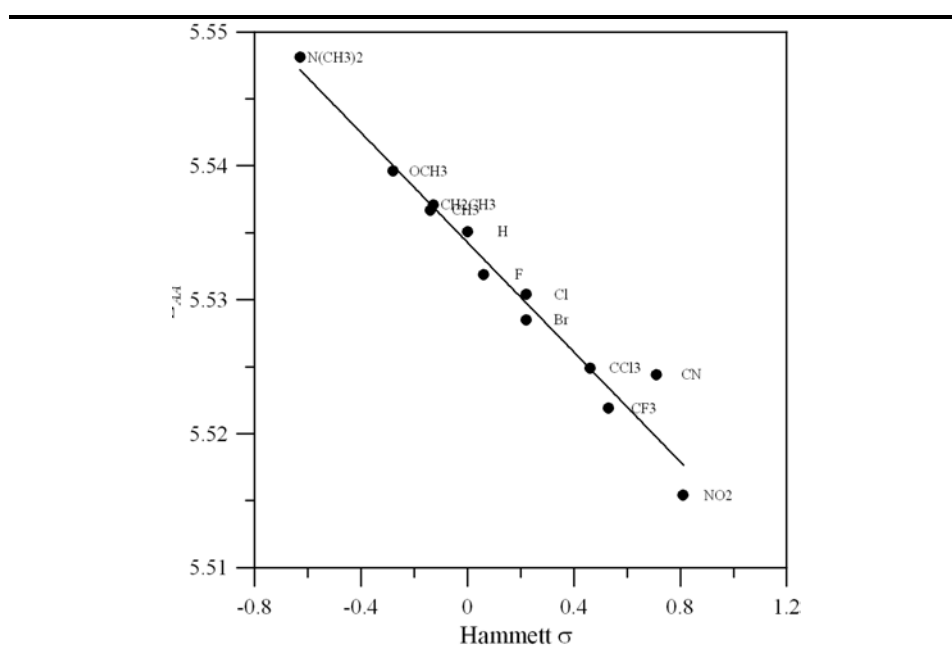


Figura 15.- Autosemblança contra sigma de Hammett per a 12 derivats de l'àcid benzoic.

Bé pel model de correlació com per la representació gràfica anterior, s'observa una bona correlació entre ambdues magnituds. Però més enllà de la simple correlació, és important observar els pendents dels models, i seguint el treball de Hammett [195,196], al aplicar la metodologia a 3 sèries d'àcids més, les dades referents a Z_{AA}^{θ} dels quals es troben a la Taula 8, s'obtenen els següents models:

$$\begin{array}{lll} \text{Tipus II} & r^2 = 0.9520 & \rho/\rho_0 = 0.46 \\ & Z_{AA}^{\theta} = -0.0095 \cdot \sigma + 5.552. & (40) \end{array}$$

$$\begin{array}{lll} \text{Tipus III} & r^2 = 0.9572 & \rho/\rho_0 = 0.63 \\ & Z_{AA}^{\theta} = -0.0129 \cdot \sigma + 5.538. & (41) \end{array}$$

$$\begin{array}{lll} \text{Tipus IV} & r^2 = 0.9251 & \rho/\rho_0 = 0.30 \\ & Z_{AA}^{\theta} = -0.0061 \cdot \sigma + 5.555. & (42) \end{array}$$

Es pot veure com en tots els casos s'obtenen models força acurats, els quals al representar-los gràficament tenen desviacions lleugerament semblants.

Taula 8.- Valors de Z_{AA}^{θ} , $\theta = \text{COOH}$ per a quatre sèries d'àcids aromàtics.

-X	Z_{AA}^{θ} , I	Z_{AA}^{θ} , II	Z_{AA}^{θ} , III	Z_{AA}^{θ} , IV
-NO ₂	5.5154	5.5423	5.5261	5.5485
-CN	5.5244	5.5468	5.5320	5.5516
-CF ₃	5.5219	5.5459	5.5306	5.5507
-CCl ₃	5.5249	5.5475	5.5328	5.5519
-Br	5.5285	5.5490	5.5349	5.5531
-Cl	5.5304	5.5498	5.5361	5.5534
-F	5.5319	5.5503	5.5366	5.5534
-H	5.5351	5.5525	5.5390	5.5555
-CH ₃	5.5367	5.5531	5.5400	5.5559
-CH ₂ CH ₃	5.5371	5.5533	5.5402	5.5559
-OCH ₃	5.5396	5.5541	5.5414	5.5556
-N(CH ₃) ₂	5.5481	5.5572	5.5469	5.5580

Al mateix temps si es comparen gràficament els pendents obtinguts, tal i com es presenten a la següent figura:

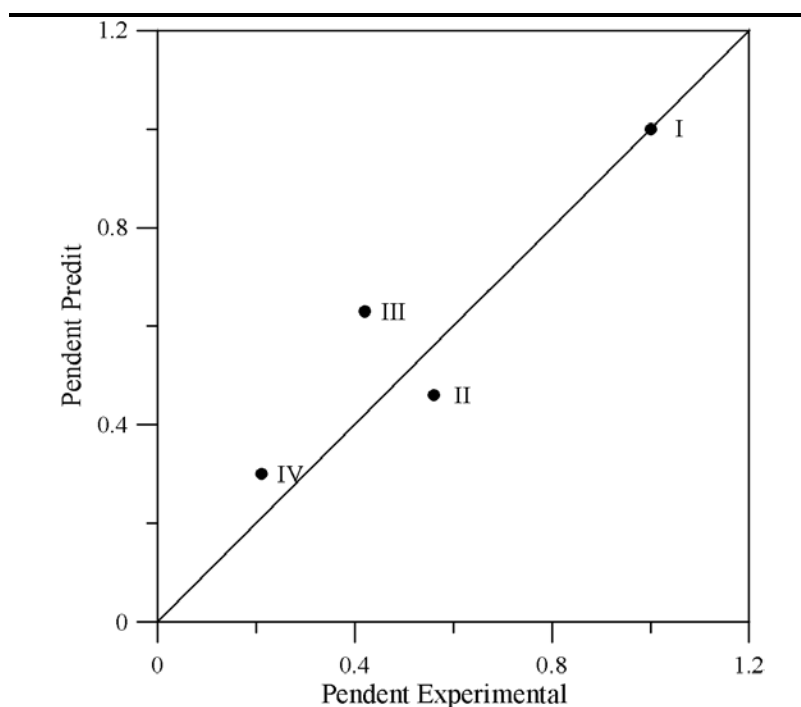


Figura 16.- Relació de pendents de correlació per a 4 sèries d'àcids aromàtics.

es pot veure com els pendents obtinguts a partir de càlculs de semblança s'ajusten força als experimentals, tal i com es d'esperar per l'augment de la distància des l'anell al grup carboxílic, sent l'efecte de la substitució menor.

6.9. Conclusions

- S'ha descrit un protocol genèric, basat en la Semblança Molecular Quàntica, per a la generació de models de predicció QSAR. Aquest protocol avarca el modelatge molecular, la creació dels descriptors i les tècniques estadístiques de correlació i validació.
- S'ha estudiat la possibilitat d'emprar, per assimilació dins del marc teòric de la Semblança Quàntica, magnituds derivades de càlculs mecanoquàntics. En el cas concret de l'energia de repulsió electró-electró, aquesta s'ha emprat com a únic descriptor en models QSAR senzills donant bons resultats.
- Les metodologies descrites anteriorment permeten l'acurada descripció de diferents propietats associades a conjunt moleculars, bé siguin propietats físiques, activitats biològiques o toxicitats moleculars, emprant un nombre relativament de descriptors i amb un cost computacional associat relativament petit.

- La nova mesura de semblança, basada en densitats de Forats de Fermi, ha demostrat ser aplicable per a la determinació dels efectes del substituent en reactivitat química. Aquesta mesura, tot i donar resultats equiparables o lleugerament millors a les autosemblances simples de fragments, cobreix el buit existent entre l'estudi de molècules senceres i l'estudi de fragments aïllats d'aquestes.

6.10. Contribucions i articles adjunts

- Robert, D. Gironés, X., and Carbó-Dorca, R. *Facet diagrams for quantum similarity data*. J. Comput.-Aided Mol. Des., 13, 1999, pp 597-610.
- Gironés, X., Amat, L., and Carbó-Dorca, R. *Descripció de propietats moleculars i activitats biològiques emprant l'energia de repulsió electró-electró*. Scientia Gerundensis, 24, 1999, pp 197-208. Inclós en aquest volum.
- Gironés, X., Amat, L., and Carbó-Dorca, R. *Using molecular quantum similarity measures as descriptors in quantitative structure-toxicity relationships*. SAR & QSAR Envir. Research, 10, 1999, pp 545-556. Inclós en aquest volum.
- Carbó-Dorca, R., Robert, D., Amat, L., Gironés, X., and Besalú, E. *Molecular quantum similarity in QSAR and drug design*. Springer Lecture Notes in Chemistry, Springer, 73, 2000.
- Robert, D., Gironés, X., and Carbó-Dorca, R. *Quantification of the influence of single-point mutations on the haloalkane dehalogenase activity: a molecular quantum similarity study*. J. Chem. Inf. Comput. Sci., 40, 2000, pp 839-846.
- Gironés, X., Amat, L., Robert, D., and Carbó-Dorca, R. *Use of electron-electron repulsion energy as a molecular descriptor in QSAR and QSPR studies*. J. Comput.-Aided. Molec. Des., 14, 2000, pp 477-485. Inclós en aquest volum.
- Carbó-Dorca, R., Amat, L., Besalú, E., Gironés, X., and Robert, D. *Quantum mechanical origin of QSAR: theory and applications*. J. Mol. Struct. (Theochem), 504, 2000, pp 181-228.
- Renners, I., Carbó-Dorca, R., Grauel, A., Ludwig, L.A., Robert, D., and Gironés, X. *Toxicity prediction by using genetically optimized B-spline networks based on molecular quantum similarity*. 2nd symposium on neural computation-NC'2000.
- Gironés, X., Gallegos, A., and Carbó-Dorca, R. *Antimalarial activity of synthetic 1,2,4-trioxanes and cyclic peroxy ketals, a quantum similarity study*. J. Comput.-Aided Mol. Des., in press. Inclós en aquest volum.

- Gironés, X., Gallegos, A., and Carbó-Dorca, R. *Modelling antimalarial activity: application of kinetic energy density quantum similarity measures as descriptors in QSAR*. J. Chem. Inf. Comp. Sci., 40, 2000, pp 1400-1407. Inclós en aquest volum.
- Renners, I., Ludwig, L.A., Grauel, A., Benfenati, E., Palagatti, S., Robert, D., Carbó-Dorca, R., and Gironés, X. *Modeling toxicity with molecular descriptors and similarity measures via B-spline networks*. 8th conference on information management of uncertainty in knowledge-based systems (IPMU'2000).
- Robert, D., Gironés, X., and Carbó-Dorca, R. *Molecular quantum similarity measures as descriptors for quantum QSAR*. Polycycl. Aromat. Comp. (ISPAC17), 19, 2000, pp 51-71.
- Gallegos, A., Robert, D., Gironés, X., and Carbó-Dorca, R. *Structure-toxicity relationships of polycyclic aromatic hydrocarbons using molecular quantum similarity*. J. Comput.-Aided Mol. Des., 15, 2001, 67-80.
- Carbó-Dorca, R., Amat, L., Besalú, E., Gironés, X., and Robert, D. *Quantum molecular similarity: theory and applications to the evaluation of molecular properties, biological activities and toxicity*. In *The fundamentals of molecular similarity*. Kluwer Academic / Plenum Publishers, New York, 2001, pp 197-320.
- Gironés, X., and Carbó-Dorca, R. *Using Molecular Quantum Similarity Measures under Stochastic Transformation to describe physical properties of molecular systems*. J. Chem. Inf. Comput. Sci., in press. Inclós en aquest volum.
- Gironés, X., and Carbó-Dorca, R. *Sobre les relacions lineals d'energia lliure: mesures de semblança molecular quàntica sobre funcions de densitat electrònica de Forats de Fermi*. Scientia Gerundensis, 25, 2001, 5-15. Inclós en aquest volum.
- Gironés, X., and Carbó-Dorca, R. *Molecular Quantum Similarity-based QSARs for Binding Affinities of several Steroid Sets*. J. Chem. Inf. Comput. Sci., en viat. Inclós en aquest volum.
- Besalú, E., Gironés, X., Amat, L., Carbó-Dorca, R. *Molecular Quantum Similarity and the Fundamentals of QSAR*. Acc. Chem. Res., in press.

Nota: Al final del Capítol 7 es llista el total de publicacions generades durant aquest treball. Tot i que algunes d'elles pertanyen a aquest camp de recerca, no s'inclouen en aquest llistat degut al seu estat embrionari.

6.11. Referències

1. Carbó, R., Leyda, L., and Arnau, M. *How similar is a molecule to another? An electron density measure of the similarity between two molecular structures*. Int. J. Quant. Chem., 17, 1980, pp 1185-1189.
2. Carbó, R., and Domingo, L. *LCAO-MO similarity measures and taxonomy*. Int. J. Quant. Chem., 23, 1987, pp 517-545.
3. Carbó, R., and Calabuig, B. *MOLSIMIL-88: molecular similarity calculations using CNDO-like approximation*. Comput. Phys. Commun., 55, 1989, 117-126.
4. Carbó, R., and Calabuig, B. *Molecular quantum similarity measures and n-dimensional representation of quantum objects. I. Theoretical foundations*. Int. J. Quant. Chem., 42, 1992, pp 1681-1693.
5. Carbó, R., and Calabuig, B. *Molecular quantum similarity measures and n-dimensional representation of quantum objects. II. Practical applications*. Int. J. Quant. Chem., 42, 1992, pp 1695-1709.
6. Carbó, R., and Calabuig, B. *Molecular quantum similarity measures and n-dimensional representation of a molecular set: phenyldimethylthiazines*. J. Mol. Struct. (Theochem), 254, 1992, pp 517-531.
7. Carbó, R., Calabuig, B., Besalú, E., and Martínez, A. *Triple density molecular quantum similarity measures: a general connection between theoretical calculations and experimental results*. Mol. Engineering., 2, 1992, pp 43-64.
8. Carbó, R., Calabuig, B., Besalú, E., and Vera, L. *Molecular quantum similarity: theoretical framework, ordering principles and visualization techniques*. Adv. Quantum Chem., 25, 1994, pp 255-313.
9. Besalú, E., Carbó, R., Mestres, J., and Sola, M. *Foundations and recent developments on molecular quantum similarity*. Top. Curr. Chem., 173, 1995, pp 31-62.
10. Carbó-Dorca, R., and Besalú, E. *A general survey of quantum similarity*. J. Mol. Struct. (Theochem), 451, 1998, pp 11-23.
11. Carbó-Dorca, R., Amat, L., Besalú, E., and Lobato, M. *Quantum Similarity*. In *Advances in Molecular Similarity*. Vol 2. Carbó-Dorca, R., and Mezey, P.G. (eds.), JAI Press, Greenwich, 1998, pp 1-42.
12. Carbó-Dorca, R., Amat, L., Besalú, E., Gironés, X., and Robert, D. *Quantum mechanical origin of QSAR: theory and applications*. J. Mol. Struct. (Theochem), 504, 2000, pp 181-228.
13. Cioslowski, J., and Fleishmann, E.D. *Assessing molecular similarity from results of ab initio electronic structure calculations*. J. Am. Chem. Soc., 113, 1991, pp 64-67.
14. Cioslowski, J., and Challacombe, M. *Maximum similarity orbitals for analysis of the electronic excited states*. Int. J. Quant. Chem., 25, 1991, pp 81-93.
15. Ortiz, J.V., Cioslowski, J. *Molecular similarity indices in electron propagator theory*. Chem. Phys. Lett., 185, 1991, pp 270-275.
16. Cioslowski, J. *Differential density matrix overlap: an index for assessment of electron correlation in atoms and molecules*. Theor. Chim. Acta, 81, 1992, pp 319-327.
17. Cooper, D.L., and Allan, N.L. *Bond formation in momentum space*. J. Chem. Soc. Faraday Trans., 83, 1987, pp 449-460.
18. Cooper, D.L., and Allan, N.L. *A novel approach to molecular similarity*. J. Comput.-Aided Mol. Des., 3, 1989, pp 253-259.
19. Cooper, D.L., and Allan, N.L. *Molecular dissimilarity: a momentum-space criterion*. J. Am. Chem. Soc., 114, 1992, pp 4773-4776.
20. Allan, N.L., and Cooper, D.L. *A momentum space approach to molecular similarity*. J. Chem. Inf. Comput. Sci., 32, 1992, pp 587-590.

21. Allan, N.L., and Cooper, D.L. *Momentum-space electron densities and quantum molecular similarity*. Top. Curr. Chem., 173, 1995, pp 85-111.
22. Cooper, D.L., and Allan, N.L. *Molecular similarity and momentum space*. In *Molecular similarity and reactivity: from quantum chemical to phenomenological approaches*. Carbó, R. (ed.), Kluwer, Amsterdam, 1995, pp 31-55.
23. Measures, P.T., Allan, N.L., and Cooper, D.L. *Momentum-space similarity: some recent applications*. In *Advances in molecular similarity*. Vol 1. Carbó-Dorca, R., and Mezey, P.G. (eds.), JAI Press, Greenwich, 1996, pp 61-87.
24. Allan, N.L., and Cooper, D.L. *Quantum molecular similarity via momentum-space indices*. J. Math. Chem., 23, 1998, pp 51-60.
25. Bowen-Jenkins, P.E., Cooper, D.L., and Richards, W.G. *Ab initio computation of molecular similarity*. J. Phys. Chem., 89, 1985, pp 2195-2197.
26. Bowen-Jenkins, P.E., and Richards, W.G. *Molecular similarity in terms of valence electron density*. J. Chem. Soc. Chem. Comm., 133, 1986, pp 133-135.
27. Bowen-Jenkins, P.E., and Richards, W.G. *Quantitative measures of similarity between pharmacological active compounds*. Int. J. Quantum Chem., 30, 1986, pp 763-768.
28. Hodgkin, E.E., and Richards, W.G. *Molecular similarity based on electrostatic potential and electric field*. Int. J. Quant. Chem., 14, 1987, pp 105-110.
29. Richards, W.G., and Hodgkin, E.E. *Molecular similarity*. Chem. in Britain, 24, 1988, pp 1141-1143.
30. Meyer, A.M., Richards, W.G. *Similarity molecular shape*. J. Comput.-Aided Mol. Des., 5, 1991, pp 426-439.
31. Good, A.C. *The calculation of molecular similarity: alternative formulas, data manipulation and graphical display*. J. Mol. Graph., 10, 1992, pp 144-151.
32. Good, A.C., Hodgkin, E.E., and Richards, W.G. *Similarity screening of molecular data sets*. J. Comput.-Aided Mol. Des., 6, 1992, pp 513-520.
33. Seri-Levy, A., West, S., Richards, W.G. *Molecular similarity, quantitative quirkality and QSAR for quiral drugs*. J. Med. Chem., 37, 1994, pp 1727-1732.
34. Seri-Levy, A., West, S., Richards, W.G. *Shape similarity as a single independent variable in QSAR*. Eur. J. Med. Chem., 29, 1994, pp 687-694.
35. Richards, W.G. *Molecular similarity and dissimilarity*. In *Modelling of biomolecular structures and mechanisms*. Pullman, A. (ed.), Kluwer, The Netherlands, 1995.
36. Richards, W.G. *The dominant role of shape similarity and dissimilarity in QSAR*. In *QSAR and molecular modelling: concepts, computational tools and biological applications*. Sanz, F., and Manaut, F. (eds.), Prous Science, Barcelona, 1995, pp 364-373.
37. Fratev, F., Polansky, O.E., Mehlhorn, A., and Monev, V. *Application of distance and similarity measures. The comparison of molecular electronic structures in arbitrary electronic states*. J. Mol. Struct., 56, 1979, pp 245-253.
38. Fratev, F., Monev, V., Mehlhorn, A., and Polansky, O.E. *Application of similarity measures. An estimation of the degree of fragmentation of a molecule in ground and excited states*. J. Mol. Struct., 56, 1979, pp 255-266.
39. Amovilli, C., and McWeeny, R. *Shape and similarity: two aspects of molecular recognition*. J. Mol. Struct. (Theochem), 227, 1991, pp 1-9.
40. Pepperrell, C.A., and Willett, P. *Techniques for the calculation of three-dimensional structural similarity using interatomic distances*. J. Comput.-Aided Mol. Des., 5, 1991, pp 455-474.
41. Petke, J.D. *Cumulative and discrete similarity analysis of electrostatic potentials and fields*. J. Comput. Chem. 14, 1993, pp 928-933.
42. Lee, C., and Smithline, S. *An approach to molecular similarity using density functional theory*. J. Phys. Chem., 98, 1994, pp 1135-1138.

43. Turner, D.B., Willett, P., Ferguson, A.M., Heritage, T.W. *Similarity searching in files of three-dimensional structures: evaluation of similarity coefficients and standarization methods for field-based similarity searching*. SAR QSAR Environ. Res., 3, 1995, pp 101-130.
44. Burgess, E.M. Ruell, J.A., Zalkow, L.H., Haugwitz, R.D. *Molecular similarity from atomic electrostatic multipole comparisons. Application to anti-HIV drugs*. J. Med. Chem., 38, 1995, pp 1635-1640.
45. Sheridan, R.P., Miller, M.D. Underwood, D.J. Kearsley, S.K. *Chemical similarity using geometric atom pair descriptions*. J. Chem. Inf. Comput. Sci., 36, 1996, pp 128-136.
46. Boon, G. De Proft, F., Langenäker, W., and Geerlings, P. *The use of density functional theory-based reactivity descriptors in molecular similarity calculations*. Chem. Phys. Lett., 295, 1998, pp 122-128.
47. Herndon, W.C., and Leonard, J.E. Inorg. Chem., 22, 1983, pp 554.
48. Herndon, W.C., and Bertz, S.H. *Linear notations and molecular graph similarity*. J. Comput. Chem., 8, 1987, pp 367-374.
49. Herndon, W.C. *Graph codes and a definition of structural similarity*. Comput. Math. Applic., 15, 1988, pp 303-309.
50. Herndon, W.C., and Bruce, A.J. J. Math. Chem., 2, 1988, pp 155.
51. Mezey, P.G. *Shape in chemistry: an introduction to molecular shape and topology*. VCH, New York, 1993.
52. Mezey, P.G. *Shape group studies of molecular similarity: shape groups and shape graphs of molecular contour surfaces*. J. Math. Chem., 2, 1988, pp 299-323.
53. Arteca, E.A., Jammal, V.B., and Mezey, P.G. *Shape group studies of molecular similarity and regioselectivity in chemical reactions*. J. Comput. Chem., 9, 1988, pp 608-619.
54. Arteca, G.A., and Mezey, P.G. *Molecular similarity and molecular shape changes along reactions paths: a topological analysis and consequences on the Hammond postulate*. J. Phys. Chem., 93, 1989, pp 4746-4751.
55. Arteca, G.A., and Mezey, P.G. *A quantitative approach to structural similarity from molecular topology of reaction paths*. Int. J. Quant. Chem. Symp., 24, 1990, pp 1-13.
56. Harary, F., and Mezey, P.G. *Similarity and complexity of the shapes of square-cell configurations*. Theor. Chim. Acta, 79, 1991, pp 379-387.
57. Mezey, P.G. *The degree of similarity of three-dimensional bodies: applications to molecular shapes*. In *Mathematical modeling in chemistry*. Mezey, P.G. (ed.), VCH, New York, 1991, pp 39-49.
58. Luo, X., and Mezey, P.G. *A global characterization and similarity analysis of two-dimensional potential energy surfaces*. Int. J. Quant. Chem., 41, 1992, pp 557-579.
59. Mezey, P.G. *Shape-similarity measures of molecular bodies: a 3D topological approach to QShAR*. J. Chem. Inf. Comput. Sci., 32, 1992, pp 650-656.
60. Mezey, P.G. *Similarity analysis in two and three dimensions using lattice animals and polycubes*. J. Math. Chem., 11, 1992, pp 27-45.
61. Arteca, G.A., and Mezey, P.G. *Similarity between the effects of configurational changes and applied electric fields on the shape of electron densities*. J. Mol. Struct. (Teechem), 256, 1991, pp 125-134.
62. Ponec, R. *Topological aspects of chemical reactivity. On the similarity of molecular structures*. Collect. Czech. Chem. Commun., 52, 1987, pp 555-561.
63. Ponec, R. *Similarity measures, the last motion principle and selection rules in chemical reactivity*. Z. Phys. Chemie, 268, 1987, pp 1180-1188.
64. Ponec, R., and Strnad, M. *Similarity approach to chemical reactivity. Specificity of multibond reactions*. Collect. Czech. Chem. Commun., 55, 1990, pp 2583-2589.
65. Ponec, R., and Strnad, M. *Topological aspects of chemical reactivity. Evans/Dewar principle in terms of molecular similarity approach*. J. Phys. Org. Chem., 4, 1991, pp 701-705.

66. Ponec, R., and Strnad, M. *Electron correlation in pericyclic reactivity: a similarity approach*. Int. J. Quant. Chem., 42, 1992, pp 501-508.
67. Ponec, R., and Strnad, M. *Similarity ideas in the theory of pericyclic reactivity*. J. Chem. Inf. Comput. Sci., 32, 1992, pp 693-699.
68. Ponec, R., and Strnad, M. *Position invariant index for assesment of molecular similarity*. Croat. Chem. Acta, 66, 1993, pp 123-127.
69. Tsai, C.C., Johnson, M.A., Nicholson, V., and Naim, M. *A topological approach to molecular similarity and its application*. In *Graph theory and topology in chemistry*. King, R.B., and Rouvray, D.H. (eds.), Elsevier, Amsterdam, 1987, pp 231-236.
70. Takahashi, Y., Sukekawa, M., and Sasaki, S.-I. *Automatic identification of molecular similarity using reduced-graph representation of chemical structure*. J. Chem. Inf. Comput. Sci., 32, 1992, pp 639-643.
71. Uraguchi, R., Sato, Y., Nakayama, A., Sukekawa, M., Iwataki, I., Boeger, P., Wakabayashi, K. *Molecular shape similarity of cyclic imides and protoporphyrinogen-IX*. J. Pesticide Sci., 22, 1997, pp 314-320.
72. Durand, P.J., Pasari, R., Baker, J.W., and Tsai, C.-C. *An efficient algorithm for similarity analysis of molecules*. <http://baker.mcs.kent.edu/paper/>
73. Carbó-Dorca, R., and Mezey, P.G. (eds.) *Advances in Molecular Similarity*. Vol. 1. JAI Press, Greenwich, 1996.
74. Carbó-Dorca, R., and Mezey, P.G. (eds.) *Advances in Molecular Similarity*. Vol. 2. JAI Press, Greenwich, 1998.
75. Carbó-Dorca, R. Gironés, X., and Mezey, P.G. (eds.) *The fundamentals of molecular similarity*. Math. Comp. Chem., Kluwer Academic Publishers, New York, 2001.
76. Carbó, R. (ed.) *Molecular similarity and reactivity: from quantum chemical to phenomenological approaches*. Kluwer, Amsterdam, 1995.
77. Johnson, A.M., and Maggiora, G.M. (eds.) *Concepts and applications of molecular similarity*. Wiley, New York, 1990.
78. Dean, P.M. (ed.) *Molecular similarity in drug design*. Blackie academic, London, 1995.
79. Kubinyi, H., Martin, Y.C., and Folkers, G. (eds.) *3D QSAR in drug design: ligand-protein interactions and molecular similarity (quantitative structure-activity relationships)*. Vol. 2. Kluwer, Dordrecht, 1998.
80. Willett, P. *Similarity and clustering in chemical information systems*. Wiley, New York, 1987.
81. Pàgina web del V Girona Seminar on molecular similarity: <http://iqc.udg.es/gsms2001>.
82. Von Neumann, J. *Mathematical foundations of quantum mechanics*. Princeton university press, Princeton, 1955.
83. Messiah, A. *Mécanique quantique*. Dunod, Paris, 1959.
84. Bohm, D. *Quantum theory*. Dover publications, New York, 1989.
85. Frisch, M.J., Trucks, G. W., Schlegel, H.B., Scuseria, G.E., Robb, M.A., Cheeseman, J.R., Zakrzewski, V.G., Montgomery, Jr., J.A., Stratmann, R.E., Burant, J.C., Dapprich, S., Millam, J.M., Daniels, D., Kudin, K.N., Strain, M.C., Farkas, O., Tomasi, J., Barone, V., Cossi, M., Cammi, R., Mennucci, B., Pomelli, C., Adamo, C., Clifford, S., Ochterski, J., Petersson, G.A., Ayala, P.Y., Cui, Q., Morokuma, K., Malick, D.K., Rabuck, A.D., Raghavachari, K., Foresman, J.B., Cioslowski, J., Ortiz, J.V., Stefanov, B.B., Liu, G., Liashenko, A., Piskorz, P., Komaromi, I., Gomperts, R., Martin, R.L., Fox, D.J., Keith, M. T., Al-Laham, A., Peng, C.Y., Nanayakkara, A., Gonzalez, C., Challacombe, M., Gill, P.M.W., Johnson, B., Chen, W., Wong, M.W., Andres, J.L., Gonzalez, C., Head-Gordon, M., Replogle, E.S., and Pople, J.A. *Gaussian 98, Revision A.6*, Gaussian, Inc., Pittsburgh PA, 1998.
86. Schmidt, M.W., Baldridge, K.K., Boatz, J.A., Elbert, S.T., Gordon, M.S. Jensen, J.H., Koseki, S., Matsunaga, N., Nguyen, K.A., Su, S., Windus, T.L., Dupuis, M., and Montgomery, J.A. *General Atomic and Molecular Electronic Structure System*. J. Comput. Chem., 14, 1993, pp 1347-1363.

87. Constants, P., and Carbó-Dorca, R. *Atomic shell approximation: electron density fitting algorithm restricting coefficients to positive values*. J. Chem. Inf. Comput. Sci., 35, 1995, pp 1046-1053.
88. Constants, P., Amat, L., Fradera, X., and Carbó-Dorca, R. *Quantum molecular similarity measures (QMSM) and atomic shell approximation (ASA)*. In *Advances in Molecular Similarity. Vol. 1*. Carbó-Dorca, R., and Mezey, P.G. (eds.), JAI Press, Greenwich, 1996, pp 187-211.
89. Amat, L., and Carbó-Dorca, R. *Quantum similarity measures under atomic shell approximation: first order density fitting using elementary jacobi rotations*. J. Comput. Chem., 18, 1997, pp 2023-2039.
90. Carbó-Dorca, R. *Tagged sets, convex sets and QS measures*. J. Math. Chem., 23, 1998, pp 353-364.
91. Carbó-Dorca, R. *On the statistical interpretation of density functions: ASA, convex sets, discrete quantum chemical molecular representations, diagonal vector spaces and related problems*. J. Math. Chem., 23, 1998, pp 365-375.
92. Carbó-Dorca, R., Besalú, E., and Gironés, X. *Extended density functions*. Adv Quantum Chem, 38, 2000, pp 3-63.
93. Gironés, X., and Carbó-Dorca, R. *Kinetic energy density function: theoretical foundation and practical QSAR application*. Pòster presentat a *Electron Densities and Electron Distributions Workshop*, celebrat a Donosti, 27/28 Abril 1999.
94. Gironés, X., Gallegos, A., and Carbó-Dorca, R. *Modeling Antimalarial Activity: Application of Kinetic Energy Density Quantum Similarity Measures as Descriptors in QSAR*. J. Chem. Inf. Comput. Sci., 40, 2000, pp 1400-1407.
95. Fradera, X., Duran, M., and Mestres, J. *Second-order quantum similarity measures from intracule and extracule densities*. Theor. Chem. Acc., 99, 1998, pp 44-52.
96. Hyman, A.S., Yaniger, S.I., and Liebman, J.F. *Interrelationship among X-ray scattering , electron densities and ionization potentials*. Int. J. Quantum Chem., 14, 1978, pp 757-766.
97. Hyman, A.S., Yaniger, S.I., Kramer, I., Bartolotti, L.J., and Liebman, J.F. *Interrelationship among X-ray scattering , electron densities and ionization potentials, II: upper bound results*. J. Chem. Phys., 81, 1984, pp 575-576.
98. Ponec, R., Amat, L., and Carbó-Dorca, R. *Molecular basis of quantitative structure-property relationships (QSPR): a quantum similarity approach*. J. Comput.-Aided Mol. Des., 13, 1999, pp 259-270.
99. Ponec, R., Amat, L., and Carbó-Dorca, R. *Quantum similarity approach to LFER: substituent and solvent effects on the acidities of carboxylic acids*. J. Phys. Org. Chem., 12, 1999, 447-454.
100. Amat, L., Besalú, E., Carbó-Dorca, R., and Ponec, R. *Identification of Molecular Sites using Quantum-Self-Similarity Measures*. J. Chem. Inf. Comput Sci., 41, 2001, pp 978-991.
101. Carbó, R., Besalú, E., Amat, L., and Fradera, X. *On quantum molecular similarity measures (QMSM) and indices (QMSI)*. J. Math. Chem., 19, 1996, pp 47-56.
102. Robert, D., and Carbó-Dorca, R. *A formal comparison between molecular quantum similarity measures and indices*. J. Chem. Inf. Comput. Sci., 38, 1998, pp 469-475.
103. Tou, J.T., and González, R.C. *Pattern recognition principles*. Addison-Wesley, Reading, 1974.
104. Amat, L., Fradera, X., and Carbó, R. *Sobre els mapes de semblança quàntica molecular*. Scientia Gerundensis, 22, 1996, pp 97-107.
105. Constants, P., Amat, L., and Carbó-Dorca, R. *Toward a global maximization of the molecular similarity function: superposition of two molecules*. J. Comput. Chem., 18, 1997, pp 826-846.
106. Gironés, X., Robert, R., and Carbó-Dorca, R. *TGSA: A Molecular Superposition Program Based on Topo-Geometrical Considerations*. J. Comput. Chem. 22, 2001, pp 255-263.

107. Atai, A., Tomioka, N., Yamada, M., Inoue, A., and Kato, Y. *Molecular superposition for rational drug design*. In *3D QSAR in drug design*. Kubunyi, H. (ed.), ESCOM, Leiden, 1993, pp 200-225.
108. Nyburg, S.C. *Some uses of a best molecular fit routine*. Acta Cryst., B30, 1974, pp 251-253.
109. Martin, Y.C. *3D database searching in drug design*. J. Med. Chem., 35, 1992, pp 2145-2154.
110. Gavuzzo, E., Pagliuca, S., Pavel, V., and Quagliata, C. *Generation and best fitting of molecular models*. Acta Cryst., B28, 1972, pp 1968-1968.
111. McLachan, A.D. *A mathematical procedure for superimposing atomic coordinates of proteins*. Acta Cryst., A28, 1972, pp 656-657.
112. Gerber, P.R., and Müller, K. *Superimposing several sets of atomic coordinates*. Acta Cryst., A41, 1987, pp 426-428.
113. Redington, P.K. *Molfit: a computer program for molecular superposition*. Comput. Chem., 16, 1992, pp 217-222.
114. Kearsly, S.K., and Smith, G.M. *An alternative method for alignment of molecular structures: maximizing electrostatic and steric overlap*. Tetrahedron Comput. Method., 3, 1990, pp 615-633.
115. Manaut, M., Sanz, F., Jose, J., and Milesi, M. *Automatic search for maximum similarity between molecular electrostatic potential distributions*. J. Comput.-Aided Mol. Des., 5, 1991, pp 371-380.
116. Clark, M., Cramer III, R.D., Jones, D.M., Patterson, D.E., and Simeroth, P.E. *Comparative molecular field analysis (CoMFA). 2. Towards its use with 3D-structural databases*. Tetrahedron Comput. Method., 3, 1990, 47-59.
117. Good, A.C., Hodgkin, E.E., and Richards, W.G. *Utilization of gaussian functions for the rapid evaluation of molecular similarity*. J. Chem. Inf. Comput. Sci., 32, 1992, pp 188-191.
118. Mestres, J., Rohrer, D.C., and Maggiora, G.M. *MIMIC: a molecular-field matching program. Exploiting applicability of molecular similarity approaches*. J. Comput. Chem., 18, 1997, pp 934-954.
119. Parretti, M.F., Kroemer, R.T., Rothman, J.H., and Richards, W.G. *Alignment of molecules by the Monte Carlo optimization of molecular indices*. J. Comput. Chem., 18, 1997, pp 1344-1353.
120. McMahon, A.J., and King, P.M. *Optimization of Carbó molecular similarity index using gradient methods*. J. Comput. Chem., 18, 1997, pp 151-158.
121. Lemmen, C., Lengauer, T., and Klebe, G. *A method for fast flexible ligand superposition*. J. Med. Chem., 23, 1998, pp 4502-4520.
122. Dean, P.M. Callow, P., and Chau, P.L. *Molecular recognition: blind searching for regions of strong structural match on the surfaces of two dissimilar molecules*. J. Mol. Graphics, 6, 1988, pp 28-34.
123. Torrent, M., Duran, M., and Solà, M. *How Similar are HF, MP2, and DFT Charge Distributions in the Cr(CO)₆ Complex?* In *Advances in Molecular Similarity*. Vol. 1. Carbó-Dorca, R., and Mezey, P.G. (eds.), JAI Press, Greenwich, 1996, pp 167-186.
124. Poater, J., Duran, M., and Solà, M. *Parametrization of the Becke3-LYP Hybrid Functional for a Series of Small Molecules Using Quantum Molecular Similarity Techniques*. J. Comput. Chem., 22, 2001, pp 1666-1678.
125. Mestres, J., Solà, M., Carbó, R., Luque, F.J., and Orozco, M. *Effect of Solvation on the Charge Distribution of a Series of Anionic, Neutral, and Cationic Species. A Quantum Molecular Similarity Study*. J. Phys. Chem., 100, 1996, pp 606-610.
126. Cioslowsky, J., Stefanov, B.B., and Constants, P. *Efficient algorithm for quantitative assessment of similarities among atoms in molecules*. J. Comput. Chem., 17, 1996, pp 1352-1358.

127. Solà, M., Mestres, J., Carbó, R., and Duran, M. *A comparative analysis by means of quantum molecular similarity measures of density distributions derived from conventional ab initio and density functionals methods*. J. Chem. Phys., 104, 1996, pp 636-647.
128. Solà, M., Mestres, J., Oliva, J.M., Duran, M., and Carbó, R. *The use of ab initio quantum molecular self-similarity measures to analyze electronic charge distributions*. Int. J. Quantum Chem., 58, 1996, pp 361-372.
129. Bach, A. Xarxes de funcions de base distribuïdes aplicades a càlculs SCF i CI de sistemes bieletrònics. Treball de recerca. Institut de Química Computacional, Girona, 1999.
130. Bach, A., and Carbó-Dorca, R. *Aplicació de la Semblança Molecular Quàntica en la reducció de l'espai configuracional per a l'estat fonamental i primers excitats de l'àtom d'heli*. Scientia Gerundensis, 24, 1999, pp 183-196.
131. Solà, M., Mestres, J., Carbó, R., and Duran, M. *Use of ab Initio Quantum Molecular Similarities as an Interpretative Tool for the Study of Chemical Reactions*. J. Am. Chem. Soc., 116, 1994, pp 5909-5915.
132. Solà, M., and Toro-Labbé, A. *The Hammond Postulate and the Principle of Maximum Hardness in Some Intramolecular Rearrangement Reactions*. J. Phys. Chem. A, 103, 1999, pp 8847-8852.
133. Hammond, G.S. *A Correlation of Reaction Rates. 'Hammond Postulate*. J. Am. Chem. Soc., 77, 1955, pp 334-338.
134. Robert, D., and Carbó-Dorca, R. *Structure-property relationships in nuclei. Prediction of the binding energy per nucleon using a quantum similarity approach*. Il Nuovo Cimento, A111, 1998, pp 1311-1321.
135. Robert, D., and Carbó-Dorca, R. *On the extension of QS to atomic nuclei: Nuclear QS*. J. Math. Chem., 23, 1998, pp 327-351.
136. Robert, D., and Carbó-Dorca, R. *General trends in atomic and nuclear quantum similarity measures*. Int. J. Quantum Chem., 77, 2000, pp 685-692.
137. Vautherin, D., and Brink, D.M. *Hartree-Fock calculations with Skyrme's interactions. I. Spherical nuclei*. Phys. Rev., C5, 1972, pp 626-647.
138. Vautherin, D. *Hartree-Fock calculations with Skyrme's interactions. II. Axially deformed nuclei*. Phys. Rev., C7, 1973, pp 296-316.
139. Martín, M., Sanz, F., Campillo, M., Pardo, L., Pérez, J., Turmo, J., and Aulló, J.M. *Quantum chemical structure-activity relationships on b-carbolines as natural monoamine oxidase inhibitors*. Int. J. Quantum Chem., 23, 1983, pp 1643-1652.
140. Sanz, F., Martín, M., Lapeña, F., and Manaut, F. *Quantitative structure-activity relationships on MAO substrates by means of quantum chemical properties*. Quant. Struct.-Act. Relat., 5, 1986, pp 54-57.
141. Sanz, F., Manaut, F., José, J., Segura, J., Carbó, M., and De la Torre, R. *Automatic determination of MEP patterns of molecules and its application to cafein metabolism inhibitors*. J. Mol. Struct. (Theochem), 170, 1988, pp 171-180.
142. Luque, F.J., Sanz, F., Illas, F., Pouplana, R., and Smeyers, Y.G. *Relationships between the activity of some H2-receptors agonists of histamine and their ab initio molecular electrostatic potencial (MEP) and electron density comparison coeficients*. Eur. J. Med. Chem., 23, 1988, pp 7-10.
143. Sanz, F., Manaut, F., Dot, T., and López de Briñas, E. *Complete or partial comparison of molecular electrostatic potential distributions? Some tests with 5-HT ligands*. J. Mol. Struct. (Theochem), 256, 1992, pp 287-193.
144. Sanz, F., Manaut, F., Rodríguez, J. Lozoya, E., and López de Briñas, E. *MEPSIM: a computational package for analysis and comparison of molecular electrostatic potentials*. J. Comput.-Aided Mol. Des., 7, 1993, pp 337-347.
145. Burt, C., Huxley, P., and Richards, W.G. *The appliction of molecular similarity calculations*. J. Comput. Chem., 11, 1990, pp 1139-1146.

146. Richard, A.M. *Quantitative comparison of molecular electrostatic potentials for structure activity studies*. J. Comput. Chem., 12, 1991, pp 959-969.
147. Rum, G., and Herndon, W.C. *Molecular similarity concepts. 5. Analysis of steroid-protein binding constants*. J. Am. Chem. Soc., 113, 1991, pp 9055-9060.
148. Herndon, W.C., and Rum, G. *Three-dimensional topological descriptors and similarity of molecular structures*. In *QSAR and molecular modeling: concepts, computational tools and biological applications: proceedings of the 10th european symposium on structure-activity relationships, QSAR and molecular modeling*. Sanz, F., Giraldo, J., and Manaut, F. (eds.), Prous Science, Madrid, 1995, pp 380-384.
149. Good, A.C., So, S.-S., and Richards, W.G. *Structure-activity relationships from molecular similarity matrices*. J. Med. Chem., 36, 1993, pp 433-438.
150. Good, A.C., Peterson, S.J., and Richards, W.G. *QSAR's from similarity matrices. Technique validation and application in the comparison of different similarity evaluation methods*. J. Med. Chem., 36, 1993, pp 2929-2937.
151. Good, A.C., and Richards, W.G. *The extension and application of molecular similarity to drug design*. Drug Information Journal, 30, 1996, pp 371-388.
152. Good, A.C., and Richards, W.G. *Rapid evaluation of shape similarity using gaussian functions*. J. Chem. Inf. Comput. Sci., 33, 1993, pp 112-116.
153. Carbó, R., and Calabuig, B. *Quantum similarity measures, molecular cloud description and structure-properties relationships*. J. Chem. Inf. Comput. Sci., 32, 1992, pp 600-606.
154. Besalú, E., Amat, L., Fradera, X., and Carbó, R. *An application of the molecular quantum similarity: ordering of some properties of the hexanes*. In *QSAR and molecular modelling: concepts, computational tools and biological applications*. Sanz, F., and Manaut, M. (eds.), Prous Science, Barcelona, 1995, pp 396-399.
155. Carbó, R., Besalú, E., Amat, L. and Fradera, X. *Quantum molecular similarity measures (QMSM) as a natural way leading towards theoretical foundation of quantitative structure-properties relationship*. J. Math. Chem., 18, 1995, pp 237-246.
156. Fradera, X., Amat, L., Besalú, E., and Carbó-Dorca, R. *Application of molecular quantum similarity to QSAR*. Quant. Struct.-Act. Relat., 16, 1997, pp 25-32.
157. Amat, L., Carbó-Dorca, R., and Poncet, R. *Molecular quantum similarity measures as an alternative to log P values in QSAR studies*. J. Comput. Chem., 19, 1998, pp 1575-1583.
158. Gironés, X., Amat, L., and Carbó-Dorca, R. *Using molecular quantum similarity measures as descriptors in quantitative structure-toxicity relationships*. SAR & QSAR Envir. Research, 10, 1999, pp 545-556.
159. Gironés, X., Amat, L., and Carbó-Dorca, R. *Use of electron-electron repulsion energy as a molecular descriptor in QSAR and QSPR studies*. J. Comput.-Aided Mol. Des., 14, 2000, pp 477-485.
160. Gironés, X., Amat, L., and Carbó-Dorca, R. *Descripció de propietats moleculars i activitats biològiques emprant l'energia de repulsió electró-electró*. Scientia Gerundensis, 24, 1999, pp 545-556.
161. Lobato, M., Amat, L., Besalú, E., and Carbó-Dorca, R. *Structure-activity relationships of a steroid family using quantum similarity measures and topological quantum similarity indices*. Quant. Struct.-Act. Relat., 16, 1997, pp 465-472.
162. Lobato, M., Amat, L., Besalú, E., and Carbó-Dorca, R. *Estudi QSAR d'una família de quinolones utilitzant índexs de semblança i índexs topològics de semblança*. Scientia Gerundensis, 23, 1998, pp 17-27.
163. Amat, L., Robert, D., Besalú, E., and Carbó -Dorca, R. *Molecular quantum similarity measures tuned 3D QSAR: an antitumoral family validation study*. J. Chem. Inf. Comput. Sci., 38, 1998, pp 624-631.
164. Robert, D., Amat, L., and Carbó -Dorca, R. *Three-dimensional quantitative structure-activity relationships from tuned molecular quantum similarity measures. Prediction of the*

- corticosteroid-binding globulin binding affinity for a steroid family*. J. Chem. Inf. Comput. Sci., 29, 1999, pp 333- 344.
165. Robert, D.; Gironés, X., and Carbó-Dorca, R. *Facet diagrams for quantum similarity data*. J. Comput.-Aided Mol. Des., 13, 1999, pp 597-610.
166. Gironés, X., Gallegos, A., and Carbó-Dorca, R. *Antimalarial Activity of Synthetic 1,2,4-Trioxanes and Cyclic Peroxy Ketals, a Quantum Similarity Study*. J. Comput.-Aided Mol. Des., en viat.
167. Robert, D., and Carbó-Dorca, R. *Aromatic compounds aquatic toxicity QSAR using quantum similarity measures*. SAR & QSAR Environ. Res., 10, 1999, pp 401-422.
168. Hyperchem 6.01, Hypercube Inc., Gainesville, FL, 1999.
169. Ampac 6.01, Semichem, 7128 Summit, Shawnee, KS 66216DA, 1994.
170. WebLab ViewerPro version 3.2, Molecular Simulations Inc., 1998. Versions de demostració actualitzades poden ser descarregades de <http://www.msi.com>.
171. Dewar, M. J. S., Zoebisch, E. G., Healy, E. F., and Stewart, J. J. P. *AM1: A New General Purpose Quantum Chemical Molecular Model*. J. Am. Chem. Soc., 107, 1985, pp 3902-3909.
172. PC Spartan Pro, version 1.4, Wavefunction Inc, Irvine, 1997.
173. Rücker, G., and Rücker, C. *On Topological Indices, Boiling Points, and Cycloalkanes*. J. Chem. Inf. Comput. Sci., 39, 1999, pp 788-802.
174. Posner, G.H., O'Dowd, H., Ploypradith, P., Cumming J.N., Xie, S., and Shapiro, T.A. *Antimalarial cyclic peroxy ketals*. J. Med. Chem., 41, 1998, pp 2164-2167.
175. The Malaria Foundation International website, at <http://www.malaria.org>.
176. Urrestarazu E., Vaes W.H.J., Verhaar H.J.M., and Hermens J.L.M. *Quantitative Structure-Activity of the Aquatic Toxicity of Polar and nonpolar Aquatic Pollutants*. J. Chem. Inf. Comput. Sci., 38, 1998, pp 845-852.
177. Amat L., Ponc R., and Carbó-Dorca R. *Molecular Quantum Similarity Measures as an Alternative to log P Values in QSAR studies*. J. Comput. Chem., 19, 1998, pp 1575-1583.
178. Fujita T., Iwasa J., and Hansch C. *A New Substituent Constant, π , Derived from Partition Coefficient*. J. Am. Chem. Soc., 86, 1964, pp 5175-5180.
179. Exner O. *Correlation Analysis of Chemical Data*. SNTL Publishers of Thecnical Literature, Prague, 1988.
180. Hansch C., Kim D., Leo A. J., Novellino E., Silipo C. Vittoria A. *Toward a quantitative comparative toxicology of organic compounds*. CRC Crit. Rev. Toxicol., 19, 1989, pp 185-226.
181. Hansch C., and Leo A. *Exploring QSAR. Fundamentals and Applications in Chemistry and Biology*. ACS Professional Reference Book, Washington, DC, 1995.
182. Parr R.G. *The Quantum Theory of Molecular Electronic Structure*. W.A. Benjamin Inc. New York, 1964.
183. Frisch, M.J., Trucks, G.W., Schlegel, H.B., Gill, P.M.W., Johnson, B.G., Robb, M.A., Cheeseman, J.R., Keith, T., Petersson, G.A., Montgomery, J.A., Raghavachari, K., Al-Laham, M.A., Zakrzewski, V.G., Ortiz, J.V., Foresman, J.B., Cioslowski, J., Stefanov, B.B., Nanayakkara, A., Challacombe, M., Peng, C.Y., Ayala, P.Y., Chen, W., Wong, M.W., Andres, J.L., Replogle, E.S., Gomperts, R., Martin, R.L., Fox, D.J., Binkley, J.S., Defrees, D.J., Baker, J., Stewart, J.P., Head-Gordon, M., Gonzalez C., and Pople, J.A. *Gaussian-94,(Revision E.2)*, Gaussian, Inc., Pittsburgh PA, 1995.
184. Anil Saxena, K. *Physicochemical Significance of Topological Parameters: Molecular Connectivity Index and Information Content: Part 2. Correlation Studies with Molar Refractivity and Lipophylicity*. Quant. Struct-Act. Relat., 14, 1995, pp 142-150.
185. Kier, L.B., and Hall, L.H. *Molecular connectivity in chemistry and drug research*. Academic, New York, 1976.
186. Kier, L.B., and Hall, L.H. *Molecular connectivity in structure-activity analysis*. Research Studies Press, Letchworth, 1986.

187. Randić, M. *On the characterization of molecular branching*. J. Am. Chem. Soc., 97, 1975, pp 6609-6615.
188. Schultz, T.W., Lin, D.T., Wilke, T.S., and Arnald, L.M. *Quantitative structure-activity relationships for the Tetrahymena Pyriformis population growth endpoint: a mechanism of action*. In *Practical applications of quantitative structure-activity relationships (QSAR) in Environmental chemistry and toxicology*. Karcher, W., and Devillers, J. (eds.), Kluwer Academic Publishers, Dordrecht, 1990, pp 241-262.
189. Ponec, R. J. Math. Chem. 21, 1997, pp 323.
190. Ponec, R. *Electron Pairing and Chemical Bonds. Molecular Structure from the Analysis of Pair Densities and Related Quantities*. J. Math. Chem., 23, 1998, pp 85.
191. Ponec, R., and Roithová, J. *Domain averaged Fermi-holes - a new means of visualization of chemical bonds. Bonding in hypervalent molecules*. Theor. Chem. Acc., 105, 2001, pp 383-392.
192. Ponec, R., and Duben A. *Electron pairing and chemical bonds. Bonding in hypervalent molecules from the analysis of the Fermi holes*. J. Comp. Chem., 8, 1999, pp 760-771.
193. Bader R.F.W., and Stephens M.E. J. Am. Chem. Soc., 98, 1975, pp 7391.
194. Cioslowski J. *Isopycnic Orbital Transformations and Localization of Natural Orbitals*. Int. J. Quant. Chem., S24, 1990, pp 15-28.
195. Hammett, L.P. *The effect of structures upon the reactions of organic compounds. Benzene derivatives*. J. Am. Chem. Soc., 59, 1937, pp 96.
196. Hammett, L.P. *Physical organic chemistry*. McGraw-Hill, New York, 1940.

6.12. Taules adjuntes

1.- Estructures i punts d'ebullició dels 529 alcans.

Compound	Boiling Point	Compound	Boiling Point	Compound	Boiling Point
CCC	-32.8	CCCCCCCCC	99.2	CCCC(C)(C)CCCC	172.5
C1CC1	-42.1	CC(C)CCCCC	118.2	CCC(CC)C(CC)CC	166
CCCC	8	CCC(C)CCCC	109.8	CC(C)(C)CC(C)CCC	164.5
CC(C)C	0.7	CCCC(C)CCCC	114.8	CC(C)C(C)C(C)CCC	179.5
C1(C)CC1	12.6	CC(C)CCCC(C)C	106.5	CC(C)CCC(C)(C)CC	171.5
C1CCCC1	-0.5	CCCCC(CC)CC	140.5	CC(C)C(CCC)C(C)C	175.8
C12CC1C2	-11.7	CCCC(CC)CCC	168.3	CCC(C)C(C)C(C)CC	160.5
CCCCC	36	CC(C)CCC(C)CC	164.6	CC(C)(C)CC(CC)CC	165
CC(C)CC	46	CC(C)C(C)CCCC	162.5	CCC(C)(CC)CCCC	172.5
CC(C)(C)C	39	CCC(C)C(C)CCC	166	CC(C)C(C)C(CC)CC	186.2
C1(CC1)CC	33.5	CCC(C)CC(C)CC	161.4	CC(C)C(CC)C(C)CC	174.5
C1(C)CC1C	35.9	CCC(CC)CC(C)C	142.5	CC(C)CC(C)(C)CCC	164
C1(C)(C)CC1	32.6	CC(C)(C)CCCCC	147.8	CC(C)(C)C(C)CCCC	160.5
C1(C)CCC1	20.6	CCC(C)C(C)CCC	140.7	CCC(C)(C)CC(C)CC	159.5
C1CCCC1	36.3	CCCC(CC)C(C)C	161	CCCC(C)(CC)CCC	162
C1C2CC1C2	49.3	CC(C)CC(C)CCC	146	CC(C)(C)CC(C)C(C)C	172.5
C1CC2CC12	36	CCC(C)CC(C)CC	137.5	CC(C)(C)CCC(C)(C)C	165.5
C1CC12CC2	27.8	CC(C)(C)CCC(C)C	137.5	CC(C)C(C)C(C)C(C)C	159.5
C12(C)CC1C2	9.5	CCC(C)(C)CCCC	131	CC(C)C(C)(C)CCCC	163
CCCCCC	71	CCCC(C)(C)CCC	128	CC(C)CC(C)(CC)CC	150
CC(C)CCC	76	CC(C)C(C)C(C)CC	152.5	CC(C)(C)C(CC)CCC	168
CCC(C)CC	83	CC(C)(C)CC(C)CC	169	CC(C)(C)C(C)CC(C)C	153
CC(C)C(C)C	81	CC(C)C(CC)C(C)C	161.8	CCC(CC)(CC)CCC	168.5
CC(C)(C)CC	69.5	CCC(C)(CC)CCC	163	CC(C)C(C(C)C)C(C)C	160
C1(CC1)CCC	60.5	CC(C)CC(C)(C)CC	163	CCC(C)(C)C(C)CCC	152
C1(CC1)C(C)C	55	CC(C)(C)C(C)CCC	164	CCC(C)C(C)(C)CCC	157.5
C1(CC)CC1C	69	CCC(CC)(CC)CC	157.8	CC(C)C(C)(C)CC(C)C	158.5
C1(C)(CC1)CC	58.3	CC(C)C(C)(C)CCC	168	CCC(C)(C)C(CC)CC	159.5
C1(C)C(C)C1C	63	CC(C)(C)C(CC)CC	155	CC(C)C(C)(CC)CCC	152
C1(C)(C)CC1C	57	CCC(C)(C)C(C)C	154	CC(C)(C)CC(C)(C)CC	155.1
C1(CC)CCC1	63	CC(C)(C)C(C)(CC)CC	152.5	CC(C)(C)C(C)C(C)CC	148.3
C1(C)CC(C1)C	52.6	CC(C)(C)CC(C)(C)C	157	CCC(C)(CC)C(C)CC	153.5
C1(C)CCC1C	70.7	CC(C)(C)C(C)C(C)C	159	CC(C)(C)C(CC)C(C)C	146.7
C1(C)(C)CCC1	59	CC(C)C(C)(C)C(C)C	149	CC(C)C(C)C(C)(C)CC	146
C1CCC(C1)C	62	CC(C)(C)C(C)(C)CC	153	CC(C)C(CC)(CC)CC	158
C1CCCCC1	53.6	C1(CC1)CCCCC	145	CC(C)C(C)(C)C(C)CC	155.5
C12CCC(C1)C2	71.8	C1(CC1)C(C)CCCC	152	CC(C)C(C)(CC)C(C)C	154
C1CC1C2CC2	80.7	C1(CC1C)CCCCC	148.2	CC(C)(C)C(C)(C)CC	145.5
C1CC2C1CC2	68.7	C1(CC)CC1CCC	151	CC(C)(C)C(C)C(C)(C)C	180
C12CCCC1C2	60.3	C1CC1(CC)CCCC	146.7	CCC(C)(C)C(C)(C)CC	172
C1CCC12CC2	63.3	CC(C)(C)CC1(C)CCC	151.5	CC(C)(C)C(C)(CC)CC	170.7
C12(C)CCC1C2	58	CC(C)CC1CC1(C)C	149	CC(C)(C)C(C)(C)C(C)C	176.5
C12(C)CC1(C2)C	49.7	C1(C)(C)CC1C(C)(C)C	151	C1(C)(CC1)CCCC(C)C	172
CCCCCCC	108.5	C1(C)(CC)CC1(C)CC	143	CC(C)CC1CC1C(C)C	153
CC(C)CCCC	104	C1(C)(C)C(C)(C)C1(C)C	138.8	C(C)CCCC1CC1(C)C	175.7
CCC(C)CCC	105	C1CCC1C(CC)CC	146.5	CCCC1CC1(C)CCC	173.9
CCC(CC)CC	107	C1CCCC1CCCC	142	CC1(C)CC1C(C)(C)CC	173.5
CC(C)CC(C)C	110	C1CCCC1CC(C)C	148.2	C1(C)C(C1)C(C)C(C)(C)C	160
CC(C)C(C)CC	107.5	C1C(C)CCC1CCC	150	C1(CCC1C(C)C)C(C)C	172
CC(C)(C)CCC	105.5	C1CCCC1C(C)CC	150.5	C1(CC)C(C)C(C)C1CC	173.9
CCC(C)(C)CC	110	C1C(CC)CCC1CC	132.2	C1(C)(C)C(CC)CC1CC	160.5
CC(C)(C)C(C)C	110.5	C1(C)CCCC1CCC	154.5	C1(C)C(CC1(C)C)C(C)C	148.5
C1CC1CCCC	116	C1(CC)CCCC1CC	141	C1CCC(C1)CCCCC	172
C1CC1C(C)CC	96.5	C1C(C)CCC1C(C)C	131	C1CCC(C1)CCC(C)C	171

C1(C)CC1CCC	98.5	C1(C)CCCC1C(C)C	135	C1(C)CCC(C1)CCCC	163
C1(CC)CC1CC	100	C1CCCC1(C)CCC	127.2	C1CCC(C1)C(C)CCC	164
C1(C)(CC1)CCC	103	C1C(C)C(C)CC1CC	149	C1CCC(C1C)CCCC	141.5
C1(CC1C)C(C)C	81.5	C1C(C)CC(C)C1CC	143	C1CC(CC1C)CC(C)C	180.9
C1(CC1)C(C)(C)C	92	C1CC(C)C(C)C1CC	153	C1CCCC1C(CC)CC	171.3
C1(CC1)(CC)CC	71.5	C1CC(C)C(C1C)CC	140.2	C1CCCC1CC(C)CC	173.4
C1(CC)C(C)C1C	74	C1(CCCCC1)C(C)(C)C	125	C1CCCC1C(C)C(C)C	169
C1(C)(CC1)C(C)C	78	C1(CC)(CC)CCCC1	125.5	CC(C)C1CCCC1CC	179.3
C1(C)(C)CC1CC	84	C1(CC)CCC(C1)(C)C	121	C1(CC)CCC(C1C)CC	175.5
C1(C)(CC)CC1C	98	C1(C)CCC(C1)(C)CC	130.8	CC(C)(CC)C1CCCC1	172
C1(C)(C)C(C)C1C	90.3	C1CCC(C1)(C)C(C)C	124.5	CC(C)C1CC(C)C(C)C1	170
C1(C)(C)CC1C(C)C	93	C1CC(C)C(C)C1(C)C	148.7	CC(C)C1CCC(C)(C)C1	176
C1CCC1CCC	90	C1C(C)C(C)C(C)C1C	156.6	C1(C)C(C)CC(C1C)CC	167
C1CCC1C(C)C	84.9	C1CCC(C)C1(C)CC	148	C1(CC)C(C)CC(C)C1C	168.5
C1C(C)CC1CC	81.1	C1C(C)C(C)CC1(C)C	148.3	C1(CCCC1(C)C)C(C)C	171
C1CC(C)C1CC	80.5	C1(C)CC(C)CC1(C)C	154.3	C1(C)C(C)C(C)C(C)C1C	174.3
C1CCCC1CC	88.6	C1(C)C(C)CCC1(C)C	148.2	C1C(C)(C)CC(C)C1(C)C	175
C1C(C)CCC1C	91	C1C(C)(C)CCC1(C)C	149.5	C1CCCC(C1)CCCC	171.5
C1(C)CCCC1C	81.5	C1(C)CCC(C)C1(C)C	150.5	C1CCCC(C1)CC(C)C	179.5
C1CCCC1(C)C	79.1	C1(C)(C)CCCC1(C)C	141	C1CC(CCC)CCC1C	168
C1CCCC(C1)C	85.2	C1CCCCC1CCC	145	C1(C)CCCC(C1)CCC	166.6
C1CCCCC1	78	C1CCCCC1C(C)C	145	C1CCCC(C1)C(C)CC	166.5
C1CC1CCC2CC2	76	C1CC(C)CCC1CC	142.5	C1CC(CC)CCC1CC	177.5
C1CC2CCC1C2	100.7	C1C(C)CCCC1CC	138	C1CC(CC)CC(C1)CC	172.5
C1CC2CC(C1)C2	92.7	C1(C)CCCCC1CC	147	C1(CC)CCCCC1C	153
C12CCCC1CC2	89.5	C1(C)CC(C)CC(C1)C	151	C1CC(CCC(C)C)C(C)C	160.3
C12CCCCC1C2	94	C1(C)CCC(C)C(C1)C	145	C1CCCC(CC)C1CC	158
C1CCC12CCC2	103.5	C1CCC(C)C(C)C1C	151	C1C(C)CCCC1C(C)C	153
C1CCCC12CC2	91.3	C1CCC(C)(CC1)CC	133	C1C(C)CC(C)CC1CC	155
C12CCC(C)C1C2	95.6	C1(C)CCC(C)(C)CC1	135.5	C1(C)CCCCC1C(C)C	167
C12CCCC1C2C	97.9	C1CCC(C)(C)CC1C	143	C1CCCCC1(C)CCC	161.5
C12CCC(C1)(C2)C	101	C1CCC(C)(C)C(C1)C	138	C1(C)C(C)CCCC1CC	182.8
C12CCCC1(C2)C	118.4	C1CCCCC(C1)CC	135.9	C1(CCCCC1)C(C)(C)C	164.2
C1C2(C)CC1(C)C2	98.5	C1CC(C)CCC(C1)C	142.5	C1(CC)(CC)CCCCC1	173
C12(C)CCC1(C2)C	90	C1CCC(C)CC(C1)C	127	C1CC(C)CCC1(C)CC	170.7
C1CC12CC2(C)C	92	C1CCCC(C)C(C1)C	129.5	C1C(C)C(C)CC(C)C1(C)C	177.5
C12(C)CC1C2(C)C	93.5	C1CCCCC(C1)(C)C	134	C1C(C)CCCC1(C)CC	161
C12CC2CC3C1C3	80.5	C1CCCC(C)CCC1	118.2	C1(C)C(C)CC(C)CC1C	174
C12CC1C3CC2C3	89.8	C1CCCCCCCC1	138	C1CCCCC1(C)C(C)C	191.4
C12CCC3C(C1)C3C2	79.2	C1CCC2CC1CCC2	135	C1(C)C(C)CCC(C)C1C	183.5
C12CCCC3C1C23	86.1	C1CCC1CC2CC2	156.7	C1(C)(C)CC(C)CC(C1)C	181.5
C123CC1CCC2C3	80.9	C1CCCC2CC1CC2	154.8	C1(C)(C)CCC(C)C(C1)C	185.5
C12CC3C4C1C2C34	137.5	C1CCCC2CCCC12	150.8	C1(C)CCC(C)CC1(C)C	170.5
C12CC1C3C4C2C34	153	C1CCCC2CCC2C1	150	C1(C)(C)CCC(C)(C)CC1	193.6
CCCCCCCC	106	C1(CCCCC1)C2CC2	154.3	C1(C)(C)CCCC(C1)(C)C	142
CC(C)CCCC	142	C1CCCCC2CC2C1	139.5	C1(C)(C)CCCC(C)C1C	174.1
CCC(C)CCCC	149	C1C(C)C2CCC1C2	144.8	C1(C)(C)CCCCC1(C)C	166.8
CCCC(C)CCC	136	C1CC2CCC(C1C)C2	149.4	C1CCCCC(C1)CCC	167.8
C(C)(C)CCC(C)C	125	C1CC2CC(C1)C(C)C2	152	C1C(C)CCC(C)CC1C	165.7
C(C)CC(CC)CC	120.5	C1C(C)CC2C1CCC2	136	C1CC(C)C(C)CCC1C	160
C(C)(C)CC(C)CC	103	C123CCCC1.C2CCC3	136.6	C1CC(C)CC(C)C(C1)C	165.1
C(C)(C)C(C)CCC	111	C123CCC1.C2CCCC3	145.1	C1CCC(C)C(C)C(C1)C	168
C(C)C(C)C(C)CC	115	C1CCC2CCC(C)C12	163.7	C1CCC(C)(C)CC(C1)C	158.5
CC(C)(C)CCCC	102	C123CC1.C2CCCCC3	153.8	C1CCC(C)C(C)(C)CC1	164
CCC(CC)C(C)C	129	C1CC2(C)CCC1CC2	151	C1CCCCC(C1)CC	157
CC(C)C(C)C(C)C	137	C1CC2CC1CC2CC	157	C1CC(C)CCCCC1C	161.8
CCC(C)(C)CCC	136	C12CCCC(C2)(C)CC1	150	C1(C)CCCCC(C1)C	153
CC(C)(C)CC(C)C	133	C1CC2CC1C(C)C2C	168.2	C1CCCCC(C)C1C	159.5
CCC(C)(CC)CC	141	C1CC2(CC)CC1CC2	175	C1CCCCC(C1)(C)C	164
CC(C)(C)C(C)CC	125	C1CC2CC(C2)C1(C)C	157.8	C1CCCCC(C)CCC1	159.7
CC(C)C(C)(C)CC	128	C1CCCC(C)C12CC2	174.5	C1CCCCCCCC1	159
C(C)(C)(C)C(C)(C)C	128	C1CCCC2(C)CC2C1	150.8	C12CCCCC2CCCC1	155

C(CCCC)C1CC1	130.5	C1C2CCC(C2)C1(C)C	142.8	C1CCC(C1)C2CCCC2	160
CC(CCC)C1CC1	125	C1C(C)C2(C)CC1CC2	144	C1CCCCC2CCCC12	145
C1(CC1C)CCCC	117	C1CCCCC12CC2C	142.4	C12CCCC(C2)CC(C1)C	166
C1CC1CC(C)(C)C	138	C1CC2(C)CCC1C2C	134	C12CCCC(C2)C(C1)C	160
C1CC1C(C)C(C)C	125	C1CC2CCC1C2(C)C	143	C12CCCC(C1)CCC2C	155.7
C1(CCC)CC1CC	115	C1(C)(C)C2CCCC1C2	142.1	C1C(C)CCC2C1CCC2	166
CC1CC1CC(C)C	91	C1CCCC2CC12CC	136	C1CCCC2C1CC(C)C2	162.4
C1(C)(C)CC1CCC	126.1	C1(C)CCC12CCC2C	133.5	C1CCC2CCCC1C2C	159.5
C1(C)(CC)CC1CC	104	C1CCCC2C1C2(C)C	140.5	C1CCCC2(C1)CCCC2	148.2
C1(C)(C)CC1C(C)C	105	C1(C)CCC2C1C2(C)C	136	C1(C)CCCC2C1CCC2	167
C1(C)(C)CC1(C)CC	159	C12(C)CCC(C2)C1(C)C	133.8	C1CCCCC2(C1)CCC2	159.7
C1(C)(C)C(C)C1(C)C	128	C1CC2(C)CC2(C1)CC	132.7	C12CCC(C2)CC1CC	167
C1(CCC1)CC(C)C	117.7	C1(C)CC2(C)CC2(C1)C	140.6	C1CCCC2CCC(C)C12	161.2
C1(CCC)CC(C1)C	124	C12CC3CC(C1)C(C2)C3	138	C12(C)CCCC(C1)CC2	157
C1(CCC1)C(C)CC	106	C1C2CC3CCC1C3C2	131.3	C1CCC2CCC(C)C12	147
C1(CC)CCC1CC	115.5	C12CCC(C2)CC3C1C3	140.4	C12CCC(C2)CC(C)C1C	157
C1(C)C(C)C(C)C1C	108	C1CC2C3C1CC2CC3	124	C12CCC(C2C)CC1C	160
C1(C)(C)CC(C1)(C)C	110	C1CC2(CC2)CC13CC3	137.3	CC1CC2CC(C)CC2C1	162
C1CCCC1CCC	105.9	C1(CC1)(C2CC2)C3CC3	135.2	C12CC(C)C(C2)CCC1C	147.7
C1CCCC1C(C)C	108.9	C1CC2CCC1C23CC3	139	C12(CC)CCC(C1)CC2	163
C1C(C)CCC1CC	94.4	C12C(C)C2C3CC1CC3	126.5	C12CCC(C2)C(C)C1C	152.8
C1(C)CCCC1CC	104.5	C1CCC2(CC2)C13CC3	136.7	C1CCCC2CCCC12C	163
C1(C)CC(C)CC1C	100.5	C1C2CC3C1C3(CC)C2	140.6	C1CC(C)C2CCC(C)C12	164
C1CCCC1(C)CC	120.1	C1C2C3CC1C(C)(C)C23	130.7	C1CCC2C1C(C)CC2C	147
C1(C)C(C)CCC1C	117.4	C1C2CC3C1C3(C)C2C	133.6	C12CCC(C2)CCC1(C)C	163.8
C1C(C)CCC1(C)C	123	C1C2CC3C1(C)C3(C)C2	145	C123CCCC1.C3CCC2C	164
C1(C)CCCC1(C)C	119	C12CC(C3C1C3)C4C2C4	137.7	C12CCC(C2)CC1(C)C	164
C1CCCC(C1)CC	114.5	C12CCC3C2C4C1C34	133.8	C12(C)CCC(C)C(C2)CC1	152
C1CC(C)CCC1C	86	CCCCCCCCCCC	122.3	C1CCC2CC(C)CC12C	158
C1CC(C)CC(C1)C	131	CC(C)CCCCCCC	133	C12(C)CCCC(C2)(C)CC1	155.7
C1CCC(C)C(C1)C	126.4	CCC(C)CCCCCC	141.5	C12CC(C)C(C2)CC1(C)C	166
C1CCCC(C1)(C)C	121	CCCC(C)CCCCC	140.2	C1CCCC2CC12CCC	147.9
C1CCCC(C)CC1	124.7	CC(C)CCCC(C)C	160	C12CCC(C2)C(C)C1(C)C	137
C1CCCCCCC1	115	CCCC(C)CCCC	174	C12(C)CCC(C2)C(C)C1C	158
C1(CC1)CCC2CC2	121.5	CCC(CC)CCCCC	155	C12CCC(C1C)C2(C)C	160
C12CCCC1CCC2	117	CC(C)CCCC(C)CC	175.5	C12(C)CCC(C2)C(C1)(C)C	158
C1(CCC1)C2CCC2	104.9	CCCC(CC)CCCC	219	C12CC(CCC1C)C2(C)C	159
C1CCC2CCC2C1	114	CC(C)CCC(C)CCC	189.5	C12(C)CCC(C)(CC1)C2C	148.7
C1CCCCC2C1C2	131.8	CCCC(CCC)CCC	158.3	C1C(C)CCC2C1C2(C)C	166.3
C1CC2CC1C(C)C2	121.8	CC(C)CC(C)CCCC	192.7	C12(C)CCC(C1)C2(C)C	157
C1CCCC12CCC2	122.3	CCC(C)CCC(C)CC	160	C1CCCCC12CC2(C)C	164
C12CCC(C1)C2C	126.6	CC(C)C(C)CCCCC	151	C1CC(C)C2CC12C(C)C	164
C1(C)CCC2C1CC2	119.5	CCC(CC)CCC(C)C	143.5	C1CCC2(C)CC12CCC	153
C123CC1.C2CCCC3	134	CCCC(C)CC(C)CC	153	C1C(CC)CC2(C)CC12C	165
C12(C)CCC(C2)CC1	149	CCCCCCCC(C)(C)C	191.5	C1CCC2(C)CC12C(C)C	169
C1CCCC2C1C2C	142	CC(C)CC(CC)CCC	190	C12CC1CCC(C2)C3CC3	153
C1CCCC2CC12C	202	CC(C)CC(C)CC(C)C	193	C1CC2C1C3CCC2CC3	155
C12CC1CC(C2)(C)C	171	CCC(C)C(C)CCCC	182	C12CCC(CC2)C3C1CCC3	170
C1CC2(C)CC1(C)C2	125.7	CCC(C)CC(C)CC	190	C1(C)CC2CCC1C3CC23	155.3
C12CCCC1C2(C)C	117.6	CC(C)CCC(C)C(C)C	187	C1C(C1(C)C2CC2)C3CC3	162
C12C(C2(C)C)C1(C)C	118.9	CC(C)C(CC)CCCC	178	C1CCCC12C3CCCC23	174
C12(C)CC1(C)C2(C)C	117.7	CCCC(C)C(C)CCC	173.5	C1CCCC2(CC2)C13CC3	164
C12CC1CC3CC3C2	109.1	CCCC(CCC)C(C)C	189.9	C12CCC(C2C)C3C1C3	169.4
C12CC1CCC3C2C3	118.5	CC(C)(C)CCCC(C)C	185	C1(C)(C)C2C3CC1(C)CC23	159.3
C12CCC(C2)C3C1C3	109.4	CCC(CC)C(C)CCC	182	C1C2CC3C1(C)C3C2(C)C	170.5
C1CC2C3CCC2C13	115.6	CC(C)C(C)CC(C)CC	183	C12(C)C3C2CC(C3)C1(C)C	168
C1(C)C2CC3C1C3C2	117.7	CCC(C)C(CC)CCC	183	C12CC2C3CC1C4(CC4)C3	166
C123CC1.C2C4(C3)CC4	106.8	CCC(C)(C)CCCCC	182	C1CC2C3CC4C(C3)C24C1	188
C1C2CC3C1(C)C3C2	115.6	CC(C)CC(C)C(C)CC	178	C12CC1C3C4C2C5C3C45	167
C1CC2(CC2)C13CC3	113.5	CC(C)(C)CCC(C)CC	174		
C12CC3C4CC1C2C34	112	CC(C)CC(CC)C(C)C	174.5		

2.- Estructures i activitat antimalarial dels 20 antimalarials

#	Estructura	log IC ₅₀	#	Estructura	log IC ₅₀
1		3.041	11		1.929
2		2.279	12		1.892
3		2.447	13		1.491
4		2.342	14		2.255
5		2.204	15		2.204
6		2.255	16		1.748
7		2.322	17		1.663
8		2.079	18		2.000
9		1.785	19		2.301
10		1.763	20		2.146

3- Estructures i toxicitats aquàtiques dels 92 derivats del benzè.

Compound	-log EC ₅₀	Compound	-log EC ₅₀	Compound	-log EC ₅₀
c1ccccc1Cl	-3.77	c1(Cl)cc(Cl)ccc1n(o)o	-4.46	c1(N)ccccc1C	-3.12
c1cccc(Cl)c1Cl	-4.4	c1(Cl)ccc(Cl)cc1n(o)o	-4.59	c1(N)cccc(c1)C	-3.47
c1ccc(Cl)cc1Cl	-4.28	c1c(Cl)cc(Cl)cc1n(o)o	-4.58	c1(N)ccc(C)cc1	-3.72
c1cc(Cl)ccc1Cl	-4.56	c1(C)c(Cl)cccc1n(o)o	-4.52	c1(ccccc1)N(C)C	-3.33
c1ccc(Cl)c(Cl)c1Cl	-4.89	c1(C)ccc(Cl)cc1n(o)o	-4.44	c1(N)cccc1CC	-3.21
c1cc(Cl)cc(Cl)c1Cl	-4.83	c1ccccc1O	-3.45	c1(N)cccc(c1)CC	-3.65
c1c(Cl)cc(Cl)cc1Cl	-4.74	c1cccc(C)c1O	-3.77	c1(N)ccc(cc1)CC	-3.52
c1cc(Cl)c(Cl)c(Cl)c1Cl	-5.35	c1ccc(C)cc1O	-3.48	c1(N)ccc(cc1)CCCC	-4.16
c1c(Cl)cc(Cl)c(Cl)c1Cl	-5.43	c1cc(C)ccc1O	-3.74	c1(N)c(cccc1C(C)C)C(C)C	-4.06
c1c(Cl)c(Cl)cc(Cl)c1Cl	-5.85	c1cc(C)cc(C)c1O	-3.86	c1(N)ccccc1Cl	-4.31
c1ccc(Cl)cc1C	-3.84	c1(C)cccc(C)c1O	-3.75	c1(N)cccc(c1)Cl	-3.98
c1cc(Cl)ccc1C	-4.33	c1cc(C)c(C)cc1O	-3.92	c1(N)ccc(Cl)cc1	-3.67
c1cc(Cl)cc(Cl)c1C	-4.54	c1(C)ccc(C)c(C)c1O	-4.21	c1(N)ccc(Cl)cc1Cl	-4.41
c1c(C)c(Cl)cc(Cl)c1Cl	-5.06	c1cc(CC)ccc1O	-4.07	c1(N)c(Cl)cccc1Cl	-4.99
c1cc(Cl)c(Cl)cc1C	-4.60	c1cc(CCC)ccc1O	-4.09	c1(N)ccc(Cl)c(c1)Cl	-4.39
c1(Cl)c(Cl)c(Cl)c(Cl)c(Cl)c1C	-6.15	c1cc(ccc1O)CCCC	-4.47	c1(N)cc(Cl)ccc(c1)Cl	-4.62
c1ccccc1	-3.09	c1cc(ccc1O)C(C)(C)C	-4.46	c1(N)ccc(Cl)c(Cl)c1Cl	-5.15
c1ccccc1C	-3.13	c1(cc(C)ccc1O)C(C)(C)C	-4.90	c1(N)c(Cl)ccc(Cl)c1Cl	-4.73
c1cccc(C)c1C	-3.48	c1cc(ccc1O)CCCC	-5.12	c1(N)cc(Cl)c(Cl)cc1Cl	-4.92
c1ccc(C)cc1C	-3.45	c1cc(ccc1O)CC(C)(C)C	-4.81	c1(N)ccc(Br)cc1	-3.56
c1cc(C)ccc1C	-3.48	c1(CC=C)ccccc1O	-3.96	c1(N)ccc(F)c(c1)C(F)(F)F	-3.77
c1ccccc1n(o)o	-2.97	c1(ccccc1O)c2ccccc2	-4.76	c1(N)ccc(F)cc1C(F)(F)F	-3.78
c1cccc(C)c1n(o)o	-3.59	c1cccc2cccc(O)c12	-4.50	c1(N)c(F)c(F)c(F)c(F)c1F	-3.69
c1ccc(C)cc1n(o)o	-3.65	c1cc(Cl)ccc1O	-4.18	c1(N)cccc(c1)OCc2ccccc2	-4.34
c1cc(C)ccc1n(o)o	-3.67	c1cc(Cl)c(C)cc1O	-4.33	c1(N)ccc(cc1)OCCCCC	-4.78
c1ccc(C)c(C)c1n(o)o	-4.39	c1c(O)cc(C)c(Cl)c1C	-4.66	c1(N)ccccc1n(o)o	-4.15
c1cc(C)c(C)cc1n(o)o	-4.21	c1ccc(OC)cc1O	-3.22	c1(N)cccc(c1)n(o)o	-3.24
c1(Cl)ccccc1n(o)o	-3.72	c1cc(OC)ccc1O	-3.05	c1(N)ccc(cc1)n(o)o	-3.23
c1c(Cl)cccc1n(o)o	-4.01	c1cc(ccc1O)Oc2ccccc2	-4.58	c1(N)c(Cl)cccc1n(o)o	-3.93
c1cc(Cl)ccc1n(o)o	-4.42	c1cccc2c1nccc2	-3.63	c1(N)c(OC)cccc1n(o)o	-3.85
c1(Cl)c(Cl)cccc1n(o)o	-4.66	c1(N)ccccc1	-2.91		

4- Estructures i refractivitat molar (MR) per a 183 compostos.

Estructura	MR	Estructura	MR	Estructura	MR
CO	8.20	C(=O)OCC	17.79	C=O	15.94
C(O)C	12.85	C(=O)OCCCC	22.49	C(=O)(C)CC	20.59
C(O)CC	17.51	CC(=O)OC	17.50	C(=O)(C)CCC	25.25
C(O)CCC	22.16	CC(=O)OCC	22.41	C(C)C(=O)CC	25.25
C(O)CCCC	26.81	CC(=O)OCCCC	27.11	CC(=O)C(C)C	25.25
C(O)CCCCC	31.43	CC(=O)OC(C)C	27.11	CC(=O)CCCC	29.90
C(O)CCCCC	36.08	CC(=O)OCCCC	31.46	CCC(=O)CCC	29.61
C(O)CCCCC	40.94	CC(=O)OCC(C)C	30.93	CC(=O)C(C)CC	29.87
C(O)CCCCC	45.39	C(C)C(=O)OC	22.77	CC(=O)CC(C)C	29.60
CC(C)O	17.47	C(CC)C(=O)OC	27.42	CCC(=O)C(C)C	29.60
C(O)C(C)C	22.13	C(CCCC)C(=O)OCC	41.02	CC(=O)CCCCC	29.60
CC(C)(O)C	22.17	C(CCCCC)C(=O)OCC	45.64	CCC(=O)CCCC	34.55
CC(C)CCO	28.34	C(CCCCCC)C(=O)OCC	50.26	CC(C)C(=O)C(C)C	34.26
CCC(C)CO	26.78	C(CCCCCC)C(=O)OCC	54.88	CCCCC(=O)CCCC	43.56
CCCC(O)C	26.78	C(CCCCCC)C(=O)OCC	59.50	CC(=O)CCCCCCC	43.82
CCC(O)CC	26.74	C(C)C(=O)OCC	27.06	c1cccc1C(C)=O	35.61
CC(C)C(C)O	26.71	C(CC)C(=O)OCC	31.72	CCCCC	25.29
CCC(C)(O)C	27.74	C(C)(C)C(=O)OCC	31.72	CCC(C)C	24.99
CC(C)(C)CO	26.79	CC(=O)OCCCCC	36.11	C(C)CC(C)C	29.65
CCCCC(C)O	31.43	C(CCC)C(=O)OCCCC	45.64	C(C)C(C)CC	28.84
CCCC(O)CC	31.44	CC(=O)OCc1cccc1	42.48	C(CC)CCC	29.92
CCC(O)(C)CC	31.39	c1(C)cccc(c1)OC(=O)CC	46.48	C(CCC)CCC	34.57
CCCC(C)(C)O	31.13	c1cccc(c1C)OC(C)=O	41.86	C(C)(C)CC(C)C	34.54
CCC(O)C(C)C	31.40	c1cc(ccc1C)OC(=O)CCC	52.17	CCCCCCCC	39.22
CCC(C)C(C)O	31.40	OC(C)=O	12.88	C1CCCC1	23.23
CC(C)CC(C)O	31.40	OC(=O)CC	16.52	C1CCCCC1	27.86
CC(C)C(C)(C)O	31.44	OC(=O)CCC	22.80	C1CCCC1C	27.85
CC(C)(C)CCO	31.44	OC(=O)OCCCCC	31.17	C1CCCCC1	32.51
CC(C)(C)C(C)O	31.44	OC(=O)OCCCCCCCC	49.78	C1CCCCC1C	32.48
CC(C)(O)CCCC	36.09	OC(=O)Cc1cccc1	37.21	C1CCCCCCC1	37.16
CCCC(C)(O)CC	36.05	OC(=O)CCc1cccc1	42.48	C1CCCC(C)C1C	37.10
CCC(O)(CC)CC	36.05	OC(=O)CCCc1cccc1	47.13	CCCC#C	23.21
CC(O)(C)C(C)CC	35.75	OC(=O)C(c1cccc1)c2cccc2	61.54	C(C)CCC#C	27.86
CC(C)C(C)(O)CC	35.75	CN	10.02	C(CC)CCC#C	32.51
CC(O)(C)CC(C)C	35.75	C(N)C	14.67	C(CCC)CCC#C	37.14
CC(C)C(O)C(C)C	35.52	C(N)CC	19.03	C(CCC#C)CCCC	41.79
CC(C)(C)C(O)CC	35.52	C(N)CCC	23.98	C(CCC#C)CCC#C	39.73
CC(C)(C)C(C)(O)CC	40.14	C(N)CCCC	28.63	C#CCCCC#C	30.43
C1(O)CCCCC1	29.24	C(N)CCCCC	33.25	CCCC=C	24.75
C=CCCCO	26.00	C(N)CCCCC	38.87	CCC=CC	24.71
CC=CC(C)O	26.23	C(N)(C)CC	23.95	C(C)CCC=C	29.40
CCC(O)C=C	26.00	C(N)C(C)C	23.63	C(CC)CC=CC	34.02
C(C)CC(O)C=C	30.92	CCCCCCCC(C)N	42.56	C(CCC)CCC=C	38.67
C(C)C(O)C=CC	30.89	C1CCCC(C1)N	31.19	CC(C)CC=C	29.37
C=CC(O)C(C)C	30.93	CC(C)N	19.33	C=CCCC=C	33.51
c1cccc1CO	32.55	CCNC	19.41	C=CCCC=C	28.85
c1cccc1CCO	37.18	C(C)CNCCC	33.37	C=CCC=C	24.20
c1cccc1CCCCO	41.84	CCN(CC)CC	29.09	C1CC=CC1	22.69
c1cccc1C(O)c2cccc2	57.19	CCCCNCCCC	43.57	C1CCC=CC1	27.34
COCC	22.23	CCNCC	24.06	C1CCCC=CC1	31.96
COCCCC	26.77	CCCCNCCC	38.02	c1cccc1C	31.01
COCC(C)C	26.91	CCCCNC	28.71	c1cccc1CC	35.66
COC(C)CC	27.03	C1CCNCC1	26.57	c1cccc1C(C)C	40.32
COC(C)(C)C	28.45	CCNC(C)C	28.72	c1cccc1CCC	40.32
CCOCCC	26.82	C(C)CNCC(C)C	38.01	c1cccc1Cc2cccc2	55.34
CCOC(C)C	26.82	C(C)CNC(C)CC	37.99	c1cccc1CCc2cccc2	59.99
C(C)COCCC	31.48	N(C)(C)C	20.17	c1cccc1c2cccc2	50.72

C(C)COC(C)C	31.48	N(C)(C)CCCC	33.75	c1cc(C)ccc1C	35.63
C(C)COC	22.17	N(C)(C)Cc1cccc1	45.53	c1cccc1	26.39
CC(C)OC	22.17	NCc1cccc1	35.45	c1ccc2cccc2c1	41.92
C1CC1OC	25.46	NCCc1cccc1	40.01	c1ccc2c(c1)ccc3cccc23	57.46

6.13. Llistat d'Articles Adjunts

- Gironés, X., Amat, L., and Carbó-Dorca, R. *Descripció de propietats moleculars i activitats biològiques emprant l'energia de repulsió electró-electró*. Scientia Gerundensis, 24, 1999, pp 197-208.
- Gironés, X., Amat, L., and Carbó-Dorca, R. *Using molecular quantum similarity measures as descriptors in quantitative structure-toxicity relationships*. SAR & QSAR Envir. Research, 10, 1999, pp 545-556.
- Gironés, X., Amat, L., Robert, D., and Carbó-Dorca, R. *Use of electron-electron repulsion energy as a molecular descriptor in QSAR and QSPR studies*. J. Comput.-Aided. Molec. Des., 14, 2000, pp 477-485.
- Gironés, X., Gallegos, A., and Carbó-Dorca, R. *Antimalarial activity of synthetic 1,2,4-trioxanes and cyclic peroxy ketals, a quantum similarity study*. J. Comput.-Aided Mol. Des., acceptat.
- Gironés, X., Gallegos, A., and Carbó-Dorca, R. *Modelling antimalarial activity: application of kinetic energy density quantum similarity measures as descriptors in QSAR*. J. Chem. Inf. Comp. Sci., 40, 2000, pp 1400-1407.
- Gironés, X., and Carbó-Dorca, R. *Using Molecular Quantum Similarity Measures under Stochastic Transformation to describe physical properties of molecular systems*. J. Chem. Inf. Comput. Sci., in press.
- Gironés, X., and Carbó-Dorca, R. *Sobre les relacions lineals d'energia lliure: mesures de semblança molecular quàntica sobre funcions de densitat electrònica de Forats de Fermi*. Scientia Gerundensis, 25, 2001, pp 5-15.
- Gironés, X., and Carbó-Dorca, R. *Molecular Quantum Similarity-based QSARs for Binding Affinities of several Steroid Sets*. J. Chem. Inf. Comput. Sci., enviat.

DESCRIPCIÓ DE PROPIETATS MOLECULARS I ACTIVITATS BIOLÒGIQUES EMPRANT L'ENERGIA DE REPULSIÓ ELECTRÓ-ELECTRÓ

Xavier Gironés, Lluís Amat i Ramon Carbó-Dorca

Institut de Química Computacional, Universitat de Girona, Catalunya, 17071 Espanya
E-mail: director@iqc.udg.es . Fax: 34 (9)72 418356

RESUM

En aquest treball es descriu l'ús de les mesures de semblança molecular quàntica (MSMQ) per a caracteritzar propietats i activitats biològiques moleculars, i definir descriptors emprables per a construir models QSAR i QSPR. L'estudi que es presenta consisteix en la continuació d'un treball recent, on es descrivien relacions entre el paràmetre log P i MSMQ, donant així una alternativa a aquest paràmetre hidrofòbic empíric. L'actual contribució presenta una nova mesura, capaç d'estendre l'ús de les MSMQ, que consisteix en l'energia de repulsió electró-electró ($\langle V_{ee} \rangle$). Aquest valor, disponible normalment a partir de programari de química quàntica, considera la molècula com una sola entitat, i no cal recórrer a l'ús de contribucions de fragments. La metodologia s'ha aplicat a cinc tipus diferents de compostos on diferents propietats moleculars i activitats biològiques s'han correlacionat amb $\langle V_{ee} \rangle$ com a únic descriptor molecular. En tots els casos estudiats, s'han obtingut correlacions satisfactòries.

RESUMEN

En este trabajo se describe el uso de las medidas de semejanza molecular cuántica (MSMQ) para caracterizar propiedades i actividades biológicas moleculares, y definir descriptores aptos para la construcción de modelos QSAR y QSPR. El estudio que se presenta consiste en la continuación de un trabajo reciente, donde se describían relaciones entre el parámetro log P y MSMQ, proporcionando así una alternativa a este parámetro hidrofóbico empírico. La actual contribución presenta una nueva medida, capaz de extender el uso de las MSMQ, consistente en la energía de repulsión electrón-electrón ($\langle V_{ee} \rangle$). Este valor, disponible normalmente a partir de software de química cuántica, considera la molécula como un solo ente, y no se requiere el uso de contribuciones de fragmentos. La metodología se ha aplicado a cinco tipos diferentes de compuestos donde propiedades y actividades biológicas se han correlacionado con $\langle V_{ee} \rangle$ como único descriptor molecular. En todos los casos estudiados, se han obtenido correlaciones satisfactorias.

ABSTRACT

In this paper it is reported the use of molecular quantum similarity measures (MQSM) to describe molecular activities and to define descriptors used when constructing QSAR and QSPR models. The present study consists of a continuation of a recent work where relationships between the log P parameter and MQSM were described, so an alternative to this empirical hydrophobic parameter was given. The present work presents a new measure, able to universalise the use of MQSM, consisting of the expectation value of the electron-electron repulsion energy ($\langle V_{ee} \rangle$). This value, commonly available from any current quantum chemical software package, considers the molecule as a whole, and there is no need to employ isolated fragments contributions. This methodology has been tested over three different sets of compounds where different biological activities have been correlated using $\langle V_{ee} \rangle$ as a parameter. In all studied cases, excellent linear relationships have been obtained.

Keywords: Electron-electron repulsion energy, log P, QSAR, QSPR, molecular quantum similarity measures.

INTRODUCCIÓ

L'entorn teòric com el desenvolupament matemàtic de les mesures de semblança molecular quàntica (MSMQ), ha estat àmpliament estudiat durant els últims 25 anys (1-12), i les seves implementacions computacionals són emprades per a solucionar problemes químics de forma general. Un exemple d'aquestes aplicacions pràctiques, actualment estudiades al nostre laboratori, relaciona les MSMQ a relacions quantitatives estructura-propietat (QSPR) (13-17). A més a més, fou demostrat que la base teòrica de les QSPR pot ésser derivada a partir del formalisme de les MSMQ (18). Aquest treball dóna la possibilitat de caracteritzar propietats i activitats biològiques moleculars mitjançant l'energia de repulsió electró-electró ($\langle V_{ee} \rangle$), considerada com un nou tipus de MSMQ. Aquesta metodologia pretén introduir el possible ús de les MSMQ a qui no té programari específic per a computar altres tipus de MSMQ.

Les MSMQ estan fonamentades sobre bases mecanoquàntiques; així, les molècules són considerades com un ens únic, i el problema d'emprar descriptors construïts a partir de contribucions atòmiques i de fragments, com és el cas del log P (19-25), és evitat, i no es requereix cap mena d'informació prèvia de tipus empíric o experimental.

La immensa quantitat de nous compostos que se sintetitzen cada any, dels quals la major part de les propietats queden desconegudes, requereix mètodes per a poder predir les seves activitats biològiques tan acuradament com sigui possible. A causa d'aquest fet, s'han realitzat molts estudis de càlculs de propietats moleculars i activitats biològiques sobre grans quantitats de conjunts moleculars. Entre els descriptors usats, el log P ha estat particularment emprat en un nombre considerable d'estudis, per exemple (26). En aquest sentit, proporcionar alternatives a aquests tipus de descriptors pot resultar de gran utilitat, i és aquí on rau el propòsit d'aquest estudi.

MATERIALS I MÈTODES

MQSM

La formulació general de les MSMQ (1) a partir del seu fonament teòric pot ésser escrita com a:

$$Z_{AB}(\Omega) = \iint \rho_A(\mathbf{r}_1) \Omega(\mathbf{r}_1, \mathbf{r}_2) \rho_B(\mathbf{r}_2) d\mathbf{r}_1 d\mathbf{r}_2 \quad (1)$$

on $\rho_A(\mathbf{r}_1)$ i $\rho_B(\mathbf{r}_2)$ són les respectives funcions densitat de les molècules A i B, $\Omega(\mathbf{r}_1, \mathbf{r}_2)$ és un operador definit positiu i $Z_{AB}(\Omega)$ és la resultant MSMQ. Quan les dues molècules comparades resulten ser la mateixa, s'obté una formulació més simplificada de l'equació (1):

$$Z_{AA}(\Omega) = \int \Omega(\mathbf{r}) \rho_A^2(\mathbf{r}) d\mathbf{r} \quad (2)$$

anomenada mesura d'autosemblança molecular quàntica (MAMQ), que és vàlida per a qualsevol operador no diferencial. Així, l'equació (2) pot ésser reformulada considerant el cas més simple com a:

$$Z_{AA}^{(1)}(\omega) = \int \omega(\mathbf{R}) \rho_A(\mathbf{R}) d\mathbf{R} \quad (3)$$

on $\mathbf{R}$ representa el conjunt de les coordenades de les partícules de qualsevol sistema microscòpic. L'equació (3) representa una formulació alternativa al valor esperat de qualsevol operador no diferencial, el qual pot ésser considerat com a un cas particular de MAMQ llavors:

$$\langle \omega(\mathbf{R}) \rangle = Z_{AA}^{(1)}(\omega) \quad (4)$$

Valors esperats dels operadors de Coulomb i Bescanvi: energia de repulsió electró-electró.

Els valors esperats dels operadors de Coulomb i Bescanvi per a un sistema de n electrons pot ser escrit, al nivell Hartree-Fock monoconfiguracional per a l'estat fonamental (27), com a:

$$\langle V_{ee} \rangle = \sum_i \sum_j [2 \cdot (ii|jj) - (ij|ij)] \quad (5)$$

on $(ii|jj)$ i $(ij|ij)$ són les respectives integrals de Coulomb i Bescanvi sobre la base dels orbitals moleculars (OM). D'aquesta manera, $\langle V_{ee} \rangle$ pot ser considerat com una suma de mesures d'autosemblança sobre els OM ocupats (2), els quals involucren l'operador de Coulomb, definit com:

$$C(\mathbf{R}) = \sum_{i>j} |\mathbf{r}_i - \mathbf{r}_j|^{-1} \wedge \mathbf{R} = (\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_n) \quad (6)$$

emprant la funció d'ona del sistema, formulada habitualment:

$$\rho_A(\mathbf{R}) = |\Psi_A(\mathbf{R})|^2 \quad (7)$$

on $\Psi_A(\mathbf{R})$ és la funció d'ona aproximada. D'aquesta manera, una expressió equivalent a l'equació (5) es pot assimilar a:

$$\langle V_{ee} \rangle = \int C(\mathbf{R}) \rho_A(\mathbf{R}) d\mathbf{R} \quad (8)$$

L'equació (8) adopta una forma anàloga a l'equació (3), la qual fou deduïda d'un entorn de MAMQ. Amb aquest simple tractament matemàtic, es pot afirmar

que l'avaluació dels operadors de Coulomb i Bescanvi, computats de manera metòdica en els programaris de química quàntica, constitueix un tipus de MAMQ i obre un nou camp per a l'aplicació d'aquesta mesura com a descriptor molecular.

Models QSAR lineals simples emprant $\langle V_{ee} \rangle$

Donat un conjunt de molècules, qualsevol propietat o activitat molecular pot ser associada al valor esperat d'un operador desconegut. Quan es consideren sèries homòlogues, es pot emprar un vector construït a partir de MAMQ com a descriptor molecular (17), en lloc de la matriu simètrica $n \times n$ derivada de MSMQ entre parells de molècules.

La intenció d'aquest treball és utilitzar $\langle V_{ee} \rangle$ com una alternativa basada en MAMQ a descriptors clàssics relacionats amb els efectes estèrics i electrònics, i trobar relacions lineals entre aquest paràmetre i propietats moleculars, emprant la formulació:

$$\pi_A = a \cdot \langle V_{ee} \rangle_A + b \quad (9)$$

Es considera que $\langle V_{ee} \rangle$ està relacionat amb els efectes estèrics, ja que la seva magnitud és creixent amb el nombre d'àtoms presents en la molècula. Així mateix, el valor de $\langle V_{ee} \rangle$ també es veu modificat per la geometria molecular, reflectint així els efectes electrònics.

Conjunts moleculars i propietats i activitats

Per a provar la utilitat d'aquesta metodologia proposada, es presenten cinc sèries de compostos dividits en dues categories: propietats moleculars i activitat biològica.

Els exemples de propietats moleculars consisteixen en els punts de fusió i ebullició d'un conjunt de 19 alcans (28) i en l'índex de refractivitat molar d'un conjunt de 183 molècules orgàniques de petita grandària (29).

Pel que fa a les activitats biològiques, el primer conjunt de molècules es refereix a l'afinitat d'unió d'un conjunt de 6 n-alquil carbamats al citocrom P-450 de cèl·lules hepàtiques de rata (30). El següent exemple consisteix en un conjunt de 13 alcohols on es mesura la concentració capaç d'inhibir el 50 % del creixement de la *Tetrahymena pyriformis*, un ciliat emprat freqüentment per a quantificar l'ecotoxicitat, avaluada per $-\log GC_{50}$ (31). L'últim exemple consisteix en la toxicitat aquàtica, mesurada en termes de $-\log GC_{50}$, d'un conjunt de derivats del fenol (32).

Modelatge molecular

Les geometries de tots els compostos d'aquest estudi s'han considerat en fase gas. L'optimització de geometria s'ha dut a terme amb l'aproximació AM1 del progra-

ma Ampac 6.0 (33), i el $\langle V_{ee} \rangle$ s'ha calculat amb el Gaussian 94 (34) a nivell HF/3-21g sobre la geometria molecular optimitzada.

Construcció i validació dels models

Les regressions lineals basades en l'equació (9) s'han dut a terme amb programari propi (35). Per a contrastar la validesa dels models, es calcula el coeficient de regressió (r^2), la desviació estàndard dels valors predits respecte als experimentals (s) i el coeficient de predicció (q^2) provenint d'una validació creuada (36).

RESULTATS

Punts de fusió i ebullició dels alcans

El primer exemple d'aplicació de $\langle V_{ee} \rangle$ com a descriptor de propietats moleculars consisteix en un conjunt de 19 alcans lineals on es troba una correlació amb els punts de fusió i ebullició. Les dades sobre els compostos, $\langle V_{ee} \rangle$ i valors de les temperatures de fusió i ebullició (28) es troben a la taula 1. Les equacions de les relacions lineals obtingudes, així com els paràmetres estadístics rellevants, es mostren a continuació, pel punt de fusió (T_f):

$$T_f^2 = 64,5 \cdot \langle V_{ee} \rangle + 10,1 \cdot 10^3$$

$$n = 19$$

$$r^2 = 0,955$$

$$s = 7,19 \cdot 10^3$$

$$q^2 = 0,940$$

i pel punt d'ebullició (T_e):

$$T_e^2 = 230 \cdot \langle V_{ee} \rangle + 37,1 \cdot 10^3$$

$$n = 19$$

$$r^2 = 0,983$$

$$s = 16,4 \cdot 10^3$$

$$q^2 = 0,977$$

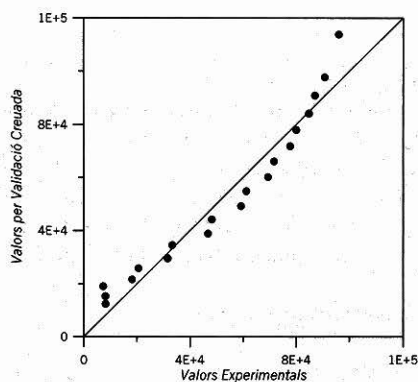


Figura 1. Representació gràfica dels punts de fusió d'un conjunt de 19 alcans calculats per validació creuada contra els valors experimentals.

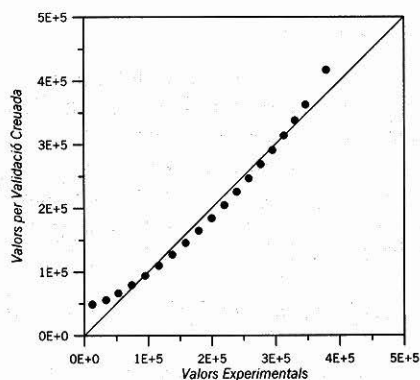


Figura 2. Representació gràfica dels punts d'ebullició d'un conjunt de 19 alcans calculats per validació creuada contra els valors experimentals.

<i>Compost</i>	$\langle V_{ee} \rangle$	T_f	T_e
Metà	26,06	90,76	111,66
Età	67,36	90,36	184,56
Propà	120,21	85,56	231,06
Butà	180,77	134,96	272,66
Pentà	247,51	143,46	309,16
Hexà	319,19	177,86	341,86
Heptà	395,07	182,56	371,66
Octà	474,58	216,36	398,76
Nonà	557,28	219,66	423,96
Decà	642,81	243,46	447,26
Undecà	730,94	247,56	469,06
Dodecà	821,38	263,56	489,46
Tridecà	913,99	267,86	508,56
Tetradecà	1008,6	278,96	526,66
Pentadecà	1105,04	283,06	543,76
Hexadecà	1203,23	291,26	559,96
Heptadecà	1303,06	295,16	575,16
Octadecà	1404,38	301,36	589,46
Icosà	1611,33	309,96	616,16

Taula 1. Dades de $\langle V_{ee} \rangle$, temperatures de fusió i ebullició (en K) referides a un conjunt de 19 alcans.

Tal i com es pot veure a partir dels resultats, $\langle V_{ee} \rangle$ dona una correlació correcta amb una notable capacitat de predicció per ambdues propietats, que depenen majoritàriament del pes molecular, i $\langle V_{ee} \rangle$ és capaç de relacionar-lo satisfactòriament gràcies al fet que inclou els efectes estèrics. D'altra banda, cal dir que hi ha efectes d'interaccions intermoleculars que també afecten les propietats, i no estan inclosos dins del descriptor proposat. Tanmateix, i tal i com es pot veure gràficament a les figures 1 i 2, l'ús d'un sol descriptor proporciona una bona descripció dels sistemes.

Índex de refractivitat molar d'un conjunt de molècules orgàniques

El segon exemple d'aplicació de $\langle V_{ee} \rangle$ com a descriptor de propietats moleculars consisteix en un conjunt de 183 molècules orgàniques senzilles (29), on la propietat avaluada és l'índex de refractivitat molar (*RM*). *RM* està relacionat amb la capacitat de les distribucions electròniques moleculars de desviar feixos de llum en interaccionar. Les dades de *RM*, així com el conjunt de molècules, han estat extretes de (29).

La correlació obtinguda en aquest cas, així com els paràmetres estadístics rellevants, es mostren a continuació:

$$RM = 46,7 \cdot 10^{-3} \cdot \langle V_{ee} \rangle + 11,3$$

$$n = 183$$

$$r^2 = 0,900$$

$$s = 2,88$$

$$q^2 = 0,897$$

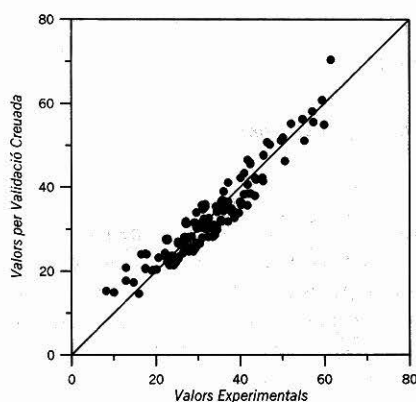


Figura 3. Representació gràfica dels índex de refractivitat molar d'un conjunt de 183 molècules orgàniques calculats per validació creuada contra els valors experimentals.

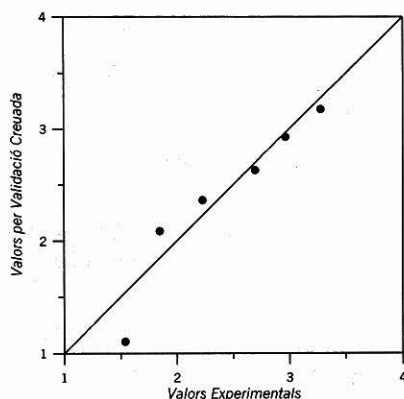


Figura 4. Representació gràfica dels log K_s d'un conjunt de 6 carbamats calculats per validació creuada contra els valors experimentals.

En aquest cas, $\langle V_{ee} \rangle$ proporciona una satisfactòria correlació de RM . Aquest exemple posa en evidència la inclusió dels efectes electrònics en $\langle V_{ee} \rangle$, ja que és capaç de discriminar entre molècules de la mateixa fórmula empírica però distinta connectivitat. Tal i com es pot observar a la figura 3, els valors predits per validació creuada s'ajusten correctament als valors experimentals, fet que no podria ésser possible si només els efectes estèrics estiguessin continguts en $\langle V_{ee} \rangle$.

Afinitat d'unió d'un conjunt de 6 n-alquil carbamats

Aquest primer exemple de tractament d'activitats biològiques involucra un conjunt de 6 n-alquil carbamats, on la propietat estudiada és l'afinitat d'unió al citocrom P-450 de cèl·lules hepàtiques de rata. Aquesta afinitat és avaluada per constant d'unió K_s (30). Les dades dels compostos i l'activitat biològica en termes de $\log K_s$ es mostren a la taula 2. La relació lineal calculada en aquest cas, així com els paràmetres estadístics, es presenten a continuació:

$$\log K_s = -3,40 \cdot 10^{-3} \cdot \langle V_{ee} \rangle + 5,01$$

$$n = 6$$

$$r^2 = 0,967$$

$$s = 0,210$$

$$q^2 = 0,875$$

Substituent	$\langle V_{ee} \rangle$	$\log K_s$
$(\text{CH}_2)_3\text{CH}_3$	524,93	3,28
$(\text{CH}_2)_4\text{CH}_3$	609,19	2,97
$(\text{CH}_2)_5\text{CH}_3$	696,07	2,7
$(\text{CH}_2)_6\text{CH}_3$	785,02	2,23
$(\text{CH}_2)_7\text{CH}_3$	876,18	1,85
$(\text{CH}_2)_9\text{CH}_3$	1064,22	1,54

Taula 2: Dades de $\langle V_{ee} \rangle$ i $\log K_s$ per a un conjunt de 6 carbamats.

Tal i com es pot veure en aquest exemple, els efectes estèrics tenen un paper important i $\langle V_{ee} \rangle$ actua principalment com un descriptor de dimensió molecular. Si s'ha de jutjar pels valors de r^2 i q^2 , el model obtingut per aquest conjunt de molècules és satisfactori. La figura 4 mostra una gràfica dels valors predits per validació creuada en comparació amb els valors experimentals, i es pot veure que els punts s'ajusten força bé, reforçant la consistència del model obtingut.

Inhibició de creixement per 13 alcohols

Aquest exemple està relacionat amb la toxicitat d'un conjunt de 13 alcohols. La toxicitat és un cas puntual d'activitat biològica de gran interès entre la comunitat científica, amb vista al disseny de nous productes. En aquest cas, es tracta la concentració capaç d'inhibir el 50 % del creixement de la *Tetrahymena pyriformis*, avaluada per $-\log GC_{50}$, la qual es correlaciona amb $\langle V_{ee} \rangle$. Les dades referides a compostos, $\langle V_{ee} \rangle$ i $-\log GC_{50}$ es mostren a la taula 3. El model obtingut emprant $\langle V_{ee} \rangle$, així com els paràmetres estadístics, es mostren a continuació:

$$-\log GC_{50} = 5.21 \cdot 10^{-3} \cdot \langle V_{ee} \rangle - 2.84$$

$$n = 13$$

$$r^2 = 0,968$$

$$s = 0,404$$

$$q^2 = 0,939$$

Els resultats per aquest sistema proporcionen una excel·lent descripció de la toxicitat emprant $\langle V_{ee} \rangle$. Tal i com es pot veure a la figura 5, els valors predits s'apropen molt als valors experimentals. També en aquest cas, $\langle V_{ee} \rangle$ és un clar descriptor dels efectes estèrics, ja que la toxicitat varia de forma evident d'acord amb la grandària molecular.

Toxicitat aquàtica deguda a derivats del fenol

L'últim exemple que es presenta també està relacionat amb la toxicitat. En aquest cas s'estudia la toxicitat aquàtica, avaluada per $-\log EC_{50}$ relacionada a un conjunt de 21 derivats del fenol. Les dades referides als compostos, $\langle V_{ee} \rangle$ i $-\log EC_{50}$

Alcohol	$\langle V_{ee} \rangle$	$-\log GC_{50}$
Metanol	81,07	-2,77
Etanol	135,21	-2,41
Propanol	196,33	-1,84
Butanol	263,38	-1,52
Pentanol	335,28	-1,12
Hexanol	411,32	-0,47
Heptanol	490,92	0,02
Octanol	573,71	0,50
Nonanol	659,3	0,77
Decanol	747,48	1,10
Unidecanol	837,97	1,87
Dodecanol	930,62	2,07
Tetradecanol	1121,73	2,28

Taula 3. Dades de $\langle V_{ee} \rangle$ i $-\log GC_{50}$ per a un conjunt de 13 alcohols.

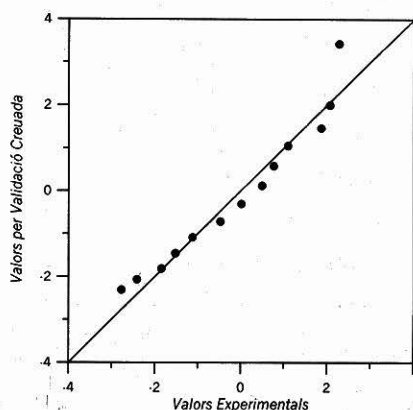


Figura 5. Representació gràfica dels $-\log GC_{50}$ d'un conjunt de 13 alcohols calculats per validació creuada contra els valors experimentals.

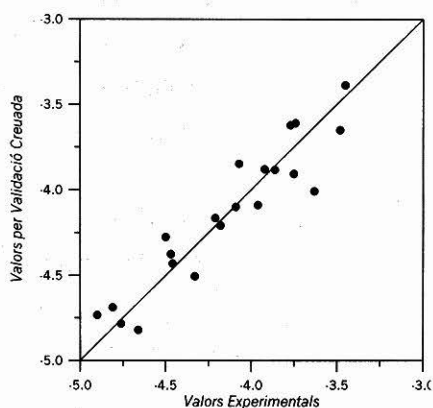


Figura 6. Representació gràfica dels $-\log EC_{50}$ d'un conjunt de 21 derivats del fenol calculats per validació creuada contra els valors experimentals.

es mostren a la taula 4. El model obtingut, així com els paràmetres estadístics, es presenten a continuació:

$$-\log EC_{50} = -2.712 \cdot 10^{-3} \cdot \langle V_{ee} \rangle - 2.389$$

$$n = 21$$

$$r^2 = 0,897$$

$$s = 0,130$$

$$q^2 = 0,877$$

Nom	$\langle V_{ee} \rangle$	$-\log EC_{50}$
Fenol	372,9	-3,45
2-metilfenol	461,22	-3,77
3-metilfenol	457,24	-3,48
4-metilfenol	456,29	-3,74
2,4-dimetilfenol	550,81	-3,86
2,6-dimetilfenol	556,09	-3,75
3,4-dimetilfenol	550,74	-3,92
2,3,6-trimetilfenol	655,91	-4,21
4-etilfenol	543,76	-4,07
4-propilfenol	630,77	-4,09
4-n-butilfenol	734,99	-4,47
4-tert-butilfenol	753,84	-4,46
2-tert-butil-4-metilfenol	874,57	-4,9
4-tert-pentilfenol	854,39	-4,81
2-allilfenol	624,8	-3,96
2-phenilfenol	882,1	-4,76
1-naftol	700,45	-4,5
4-clorofenol	670,2	-4,18
4-cloro-3-metilfenol	775,06	-4,33
4-cloro-3,5-dimetilfenol	886,62	-4,66
Quinolina	590	-3,63

Taula 4. Dades de $\langle V_{ee} \rangle$ i $-\log EC_{50}$ per a un conjunt de 21 fenols.

Es pot veure també com en aquest cas, la relació obtinguda, emprant $\langle V_{ee} \rangle$ com a descriptor, és satisfactòria, assolint una correcta correlació i una notable capacitat

de predicció. El gràfic presentat a la figura 6 verifica el model i presenta uns valors predits força propers als experimentals. En aquest exemple, al igual que l'anterior de 183 molècules on es relaciona el RM , $\langle V_{ee} \rangle$ descriu els efectes estèrics i els electrònics, ja que també és capaç de discriminar entre compostos on l'única diferència estructural és la connectivitat.

DISCUSSIÓ

Un grup de conjunts moleculars compostos per alcans, molècules orgàniques petites, carbamats, alcohols i derivats del fenol han estat estudiats per correlacionar les seves propietats moleculars o activitats biològiques amb la repulsió coulombiana, $\langle V_{ee} \rangle$, considerada com una mesura d'autosemblança. En tots els sistemes estudiats s'han obtingut bones relacions lineals entre els paràmetres analitzats i $\langle V_{ee} \rangle$, donant també una notable capacitat de predicció. Aquests resultats indiquen que les MSMQ, un mètode basat en densitats electròniques en lloc de contribucions atòmiques o de fragments, constitueix una metodologia apropiada per quantificar o complementar models QSPR i QSAR en general.

AGRAÏMENTS

Aquesta recerca ha estat parcialment subvencionada per una bonificació del CICYT (SAF 96-0158), la *Fundació Maria Francisca de Roviralta* i un contracte de la Unió Europea: #ENV4-CT97-0508. Un de nosaltres (X. Gironés) vol agrair la Universitat de Girona per la concessió d'una beca predoctoral.

Bibliografia

1. CARBÓ R., LEYDA L. i ARNAU M., *Int. J. Quantum Chem.*, 1980, 17, 1185.
2. CARBÓ R. i DOMINGO LL., *Int. J. Quantum Chem.*, 1987, 23, 517.
3. BESALÚ E., CARBÓ R., MESTRES J. i SOLÀ M., *Topics Curr. Chem.*, 1995, 173, 31.
4. CARBÓ R., BESALÚ E., CALABUIG B. i VERA L., *Adv. in Quantum Chem.*, 1994, 25, 253.
5. CARBÓ R. (ed.), *Molecular Similarity and Reactivity: From Quantum Chemical to Phenomenological Approaches*, Kluwer, Amsterdam, 1995.
6. CARBÓ-DORCA R. i MEZEY P. G. (eds.), *Advances in Molecular Similarity*, vol. 1, JAI Press, Greenwich, CT, 1996.
7. CIOŚLOWSKY J. i FLEISCHMANN E. D., *J. Am. Chem. Soc.*, 1991, 113, 64.
8. BURT C., RICHARDS W. G. i HUXLEY P., *J. Comput. Chem.*, 1990, 10, 1139.
9. MEZEY P. G., *Topics Curr. Chem.*, 1995, 173, 63.
10. ALLAN N. L. i COOPER D. L., *Topics Curr. Chem.*, 1995, 173, 85.
11. PONEC R., *Topics Curr. Chem.*, 1995, 174, 1.
12. CARBÓ-DORCA R. i BESALÚ E., *J. Mol. Struct.*, 1998, 451, 11.
13. FRADERA X., AMAT L., BESALÚ E. i CARBÓ-DORCA R., *Quant. Struct.-Act. Relat.*, 1997, 16, 25.

14. LOBATO M., AMAT L., BESALÚ E. i CARBÓ-DORCA R., *Quant. Struct.-Act. Relat.*, 1997, 16, 465.
15. AMAT L., ROBERT D., BESALÚ E. i CARBÓ-DORCA R., *J. Chem. Inf. Comput. Sci.*, 1998, 38, 624.
16. PONEC R., AMAT L., i CARBÓ-DORCA R., *J. Comp.-Aided Molec. Design*, 1999, 13, 259.
17. AMAT L., PONEC R. i CARBÓ-DORCA R., *J. Comput. Chem.*, 1998, 19, 1575.
18. CARBÓ R., BESALÚ E., AMAT L. i FRADERA X., *J. Math. Chem.*, 1995, 18, 237.
19. FUJITA T., IWASA J. i HANSCH C., *J. Am. Chem. Soc.*, 1964, 86, 5175.
20. REKKER R. F. (ed.), *The hydrophobic Fragmental Constants. Its Derivation and Applications. A Means of Characterizing Membrane Systems*, Elsevier, New York, 1977.
21. REKKER R. F. i KORT H. M., *Eur. J. Med. Chem.*, 1979, 14, 479.
22. HANSCH C. i LEO A. (ed.), *Substituent Constants for Correlation Analysis in Chemistry and Biology*, Wiley, New York, 1979.
23. KLOPMAN G. i IROFF L., *J. Comput. Chem.*, 1981, 2, 157.
24. KLOPMAN G., NAMBOODIRI K. i SCHOCHET M., *J. Comput. Chem.*, 1985, 1, 28.
25. GHOSE A. K. i CRIPPEN G. M., *J. Comput. Chem.*, 1986, 7, 565.
26. HANSCH C., KIM D., LEO A. J., NOVELLINO E., SILIPO C. i VITTORIA A., *CRC Crit. Rev. Toxicol.*, 1989, 19, 185.
27. ROTHMAN C. C. J., *Revs. Mod. Phys.*, 1951, 23, 69.
28. LIDE D. R. (ed.), *Handbook of Chemistry and Physics 76th Edition*, CRC Press, Boca Raton, 1995.
29. ANIL SAXENA K., *Quant. Struct.-Act. Relat.*, 1995, 14, 142.
30. SARGENT N. S. E., UPSHALL D. G. i BRIDGES J. W., *Bio.Pharm.*, 1982, 31, 1309.
31. KARCHER W. i DEVILLERS J. (eds.), *Practical Applications of Quantitative Structure-Activity Relationships (QSAR) in Environmental Chemistry and Toxicology*, Kluwer Academic Publishers, Dordrecht, 1990.
32. URRESTARAZU E., VAES W. H. J., VERHAAR H. J. M. i HERMENS J. L. M., *J. Chem. Inf. Comput. Sci.*, 1998, 38, 845.
33. Ampac 6.0, 1994 Semichem, 7128 Summit, Shawnee, KS 66216DA.
34. FRISCH M. J., TRUCKS G. W., SCHLEGEL H. B., GILL P. M. W., JOHNSON B. G., ROBB M. A., CHEESEMAN J. R., KEITH T., PETERSSON G. A., MONTGOMERY J. A., RAGHAVACHARI K., AL-LAHAM M. A., ZAKRZEWSKI V. G., ORTIZ J. V., FORESMAN J. B., CIOŚLOWSKI J., STEFANOV B. B., NANAYAKKARA A., CHALLACOMBE M., PENG C. Y., AYALA P. Y., CHEN W., WONG M. W., ANDRES J. L., REPLOGLE E. S., GOMPERTS R., MARTIN R. L., FOX D. J., BINKLEY J. S., DEFREES D. J., BAKER J., STEWART J. P., HEAD-GORDON M., GONZÁLEZ C. i POPLE J. A., *Gaussian-94, (Revision E.2)*, Gaussian, Inc., Pittsburgh PA, 1995.
35. GIRONÉS X. i CARBÓ-DORCA R. *Programa NDIM*, Institut de Química Computacional, Girona, Catalunya, Espanya, 1999.
36. ALLEN D. M. *Technometrics*, 1974, 16, 125.

USING MOLECULAR QUANTUM SIMILARITY MEASURES AS DESCRIPTORS IN QUANTITATIVE STRUCTURE-TOXICITY RELATIONSHIPS

X. GIRONÉS, L. AMAT and R. CARBÓ-DORCA*

*Institute of Computational Chemistry, University of Girona,
Catalonia, 17071 Spain*

(Received 16 April 1999; In final form 26 June 1999)

In this paper molecular quantum similarity measures (MQSM) are used to describe molecular toxicity and to construct Quantitative Structure-Toxicity Relationships (QSTR) models. This study continues the recently described relationships between MQSM and $\log P$ values, which permits to use the theoretical MQSM as an alternative to the empirical hydrophobic parameter in QSPR studies. In addition a new type of MQSM is presented in this work: it is based on the expectation value of electron–electron repulsion energy. The molecular properties studied here, as application examples are aquatic toxicity, toxicology on Bacteria and inhibition of a macro-molecule employing four different molecular sets.

Keywords: Molecular quantum similarity measures; molecular toxicology; toxicity; electron–electron repulsion energy; QSAR; QSTR; $\log P$

INTRODUCTION

Correlations and predictions of biological, or toxicological activity, have become an extensively studied field due to the large quantity of new compounds synthesized nowadays. So, *a priori* knowledge of any molecular activity plays an important role in current chemistry. Large collections of papers on the QSAR subject have been published up to date and, as a result, huge amounts of relationships between the hydrophobicity character, normally represented by the octanol-water partition coefficient ($\log P$), and

*Corresponding author.

diverse biological properties of many series of compounds have been obtained, see for example references [1–4].

The present study can be partly considered as a continuation of previous work [5, 6], where the relationship between molecular quantum similarity measures (MQSM) and $\log P$ was demonstrated. Here is analyzed a new possible way to characterize molecular toxicity using MQSM based on the quantum expectation value of electron–electron repulsion energy calculations ($\langle V_{ee} \rangle$). This work pretends to describe a complementary quantum mechanical point of view for classical QSAR studies based on empirical parameters. Our goal consists to prove that $\langle V_{ee} \rangle$ can be considered a new type of MQSM and that a simple relationship exists between relevant toxicological properties and $\langle V_{ee} \rangle$ in some molecular families.

In this work, MQSM methodology is applied over three series of biological compounds. The first studied group consists in benzene derivatives, in which the aquatic toxicity is studied [7]. Next, alcohols both with straight and branched chain are considered, and the QSAR example presented consists in the inhibition of *Bacillus subtilis* spore germination [8]. The last analyzed molecular group relates its activity to inhibition of Human Sickel Hemoglobin Gelation [9], and is formed by ureas and amides. In all cases a good linear QSTR using MQSM is found.

THEORETICAL BACKGROUND

The theoretical framework of MQSM, its mathematical development and computational implementations, have been largely developed and studied in the past 20 years [10–21]. MQSM provide a well-defined mathematical tool, useful to be applied in varied practical chemical problems. One of these applications, which is currently studied in our laboratory, corresponds to the use of MQSM in the determination of quantitative-structure property relationships (QSPR) [5, 6, 22–26]. This methodology is based on the quantification of the similarity between two molecules using first order density functions of both studied systems. Once density functions have been calculated, MQSM between two molecules *A* and *B*, can be computed by direct quantitative comparison, generally defined by the integral:

$$Z_{AB}(\Omega) = \iint \rho_A(\mathbf{r}_1) \Omega(\mathbf{r}_1, \mathbf{r}_2) \rho_B(\mathbf{r}_2) d\mathbf{r}_1 d\mathbf{r}_2, \quad (1)$$

where $\{\rho_A(\mathbf{r}_1), \rho_B(\mathbf{r}_2)\}$ are the respective molecular first order density functions and $\Omega(\mathbf{r}_1, \mathbf{r}_2)$ is a positive definite operator used as a weight.

In general, MQSM depend on the relative position of the involved molecules in space. As a consequence, an optimization process is needed to obtain the optimal superposition of both structures, which corresponds to the maximum value of integral (1) [27]. When both molecules coincide, then the corresponding MQSM: Z_{AA} is called a molecular quantum self-similarity measure (MQS-SM). In this case no integral optimization has to be performed. Here, only MQS-SM will be considered, and consequently the cumbersome part required for the optimization of molecular superposition is skipped. In this context then, *ab initio* MQSM can be used, and thus no approximate density functions are required as in usual MQSM procedures [22–25]. The practical implementation of MQSM in the present context consists to use first-order LCAO density functions, corresponding to the expression

$$\rho_A(\mathbf{r}) = \sum_{\mu, \nu \in A} P_{\mu\nu} \chi_{\mu}^*(\mathbf{r}) \chi_{\nu}(\mathbf{r}), \quad (2)$$

where $P_{\mu\nu}$ is the charge-bond order matrix, and $\{\chi_{\mu}\}$ are the AO basis set functions.

The most useful form of the two-electron operator Ω used in Eq. (1) corresponds to a Dirac delta function: $\Omega(\mathbf{r}_1, \mathbf{r}_2) = \delta(\mathbf{r}_1 - \mathbf{r}_2)$, which gives the so-called overlap-like MQSM. Other possibilities are available for MQSM practical implementation [12, 13]. A particular case may be illustrated by the expectation value of the electron–electron repulsion potential operator for a p -particle system, which may be written as:

$$\Omega = V_{ee} = \sum_{i=1}^{p-1} \sum_{j=i+1}^p (|r_i - r_j|)^{-1}. \quad (3)$$

It is well known that use of the Coulomb operator (3) produces an expectation value, in the closed shell mono-configurational case, where Coulomb $\{\langle ii|jj \rangle\}$ and exchange $\{\langle ij|ij \rangle\}$ two-electron integrals play a leading role [28]:

$$\langle V_{ee} \rangle = \sum_j \sum_j (2\langle ii|jj \rangle - \langle ij|ij \rangle) > 0. \quad (4)$$

Although both parts of the expression (4) can be used by themselves as MQS-SM measures [11], the positive definite nature of the molecular quantum electron repulsion energy $\langle V_{ee} \rangle$, as a whole, can be considered such a MQSM too. This opens the way to the potential use of $\langle V_{ee} \rangle$ as a molecular descriptor. The advantage of using $\langle V_{ee} \rangle$ consists of the fact that it is a customarily computed parameter in the available quantum chemical programs.

SIMPLE LINEAR QSPR MODEL USING MQS-SM

A symmetric ($n \times n$) similarity matrix associated to a molecular set may be defined as: $\mathbf{Z} = \{Z_{IJ}\}$. The $\mathbf{Z}$ elements are computed using MQSM between pairs of molecules of the set and the diagonal is made of self-similarity measures. Each similarity matrix column can be considered as a source of descriptors, like a vector containing some molecular information which can be used to construct QSAR models, $\mathbf{Z} = \{\mathbf{z}_I\}$ [22–26]. For any given molecule in a studied set, a molecular property can be associated to the expectation value of some unknown operator [29]:

$$\pi_I \approx \mathbf{a}^T \mathbf{Z}_I \quad (5)$$

where $\mathbf{a}$ is a n -dimensional vector, representing the unknown operator associated to the molecular property in the n -dimensional discrete space of MQSM vectors $\mathbf{z}_I$. The property-MQSM pairs $\{\pi_I, \mathbf{z}_I\}$ are well defined, and the elements of $\mathbf{a}$ can be obtained by least-squares techniques. In this way Eq. (5) becomes the theoretical basis of QSAR, QSPR or QSTR. Equation (5) can also be written as:

$$\pi_I \approx a_I Z_{II} + \sum_{J \neq I} a_J Z_{JI}, \quad (6)$$

where the MQS-SM have been isolated from the rest of the similarity matrix. The terms: $a = a_I$, $b = \sum_{J \neq I} a_J Z_{JI}$ can be considered roughly constant for a given homogeneous molecular series [5, 6]. Then, a simple linear relationship may be present between some molecular properties and MQS-SM:

$$\pi_I \approx a Z_{II} + b. \quad (7)$$

Owing to the MQS-SM nature of $\langle V_{ee} \rangle_I$ for a given I -th molecule, as discussed before, Eq. (7) could be also written as:

$$\pi_I \approx a \langle V_{ee} \rangle_I + b. \quad (8)$$

RESULTS AND DISCUSSION

As mentioned before, among the literature one can found a considerable amount of QSAR between biological activities and empirical parameters, such as hydrophobicity, steric and electronic effects over a great number of

compounds [1–4]. In this work, three series of biologically active compounds have been considered. The first one accounts for aquatic toxicity related to the guppy *Poecilia reticulata*, which inversely correlate to $\log P$ and their ability to form hydrogen bonds [7]. The second one is related to Inhibition of *Bacillus subtilis* spore germination, which has been demonstrated to be produced by the interaction between a hydrophobic region on the spore and the hydrophobic chain of alcohols [8]. The last series is based on inhibition of Human Sickel Hemoglobin Gelation by ureas and amides, which is related to two different hydrophobic effects: the electrostatic interaction of the aliphatic chain with the polymer, and the modulation of the functional group in the expression of such hydrophobic character [9]; in this last case, ureas and amides have been studied separately.

All obtained QSAR models are submitted to a cross-validation analysis with the leave-one-out method, which consists of reconstructing the model after leave an element out for later predicting its activity. This process is carried out for all molecules. This validation method allows the computation of the cross-validation coefficient for prediction q^2 [30] parameter, which informs about the predictive capacity of the model.

All studied molecules have been considered in completely optimized geometries. MO calculations have been carried out at the HF/3-21G* computational level, using the Gaussian 94 program [31]. $\{\langle V_{ee} \rangle_I\}$ have been extracted from the output files of this program. All $\langle V_{ee} \rangle_I$ values have been standardized in the statistical sense and then, symbolized by $\langle V_{ee} \rangle_I^{(N)}$. That is, a new variable is defined: $\langle V_{ee} \rangle^{(N)} = (\langle V_{ee} \rangle - \overline{\langle V_{ee} \rangle}) \cdot s(\langle V_{ee} \rangle)^{-1}$, where $\overline{\langle V_{ee} \rangle}$ is the arithmetic mean and s is the standard deviation of the $\langle V_{ee} \rangle$ values associated to the molecular set.

Aquatic Toxicity to *Poecilia Reticulata*

The first example presented here consists of a set of 36 benzene derivatives, where the aquatic toxicity related to *Poecilia Reticulata* [7], evaluated by $\log EC_{50}$, has been correlated by $\langle V_{ee} \rangle$. All data concerning molecules, $\langle V_{ee} \rangle$ and biological activity are listed in Table I. The QSTR, as well as all relevant statistical parameters, obtained are:

$$\log EC_{50} = 0.687 \cdot \langle V_{ee} \rangle^{(N)} + 4.309 \quad (9)$$

$$n = 36 \quad r^2 = 0.889 \quad s = 0.241 \quad q^2 = 0.877$$

TABLE I $\langle V_{ee} \rangle$ and $\log EC_{50}$ values for 36 benzene derivatives

Compound	$\langle V_{ee} \rangle$	$\log EC_{50}^a$
Chlorobenzene	563.87	3.77
1,2-Dichlorobenzene	890.67	4.40
1,3-Dichlorobenzene	874.41	4.28
1,4-Dichlorobenzene	871.07	4.56
1,2,3-Trichlorobenzene	1243.04	4.89
1,2,4-Trichlorobenzene	1222.56	4.83
1,3,5-Trichlorobenzene	1209.43	4.74
1,2,3,4-Tetrachlorobenzene	1616.76	5.35
1,2,3,5-Tetrachlorobenzene	1599.44	5.43
1,2,4,5-Tetrachlorobenzene	1595.48	5.85
3-Chlorotoluene	653.96	3.84
4-Chlorotoluene	652.23	4.33
2,4-Dichlorotoluene	983.16	4.54
2,4,5-Trichlorotoluene	1343.82	5.06
3,4-Dichlorotoluene	991.52	4.60
Pentachlorobenzene	2015.06	6.15
Benzene	277.72	3.09
Toluene	355.54	3.13
2-Xylene	443.42	3.48
3-Xylene	439.54	3.45
4-Xylene	438.61	3.48
Nitrobenzene	557.51	2.97
2-Nitrobenzene	660.82	3.59
3-Nitrobenzene	651.68	3.65
4-Nitrobenzene	649.25	3.67
2,3-Dimethylnitrobenzene	765.26	4.39
3,4-Dimethylnitrobenzene	753.58	4.21
2-Chloronitrobenzene	895.94	3.72
3-Chloronitrobenzene	876.00	4.01
4-Chloronitrobenzene	871.45	4.42
2,3-Dichloronitrobenzene	1255.93	4.66
2,4-Dichloronitrobenzene	1234.64	4.46
2,5-Dichloronitrobenzene	1235.61	4.59
3,5-Dichloronitrobenzene	1219.06	4.58
2-Chloro-6-nitrotoluene	999.56	4.52
4-Chloro-2-nitrotoluene	990.12	4.44

^a from reference [7], but with signs changed.

As it can be observed from the above results, $\langle V_{ee} \rangle$ is able to correlate the aquatic toxicity of the benzene derivatives yielding a $r^2 = 0.889$. In addition, the predictive capacity of the model is also high, $q^2 = 0.877$, obtaining values fairly close to experimental ones when a cross-validation is performed. Figure 1 shows a plot of cross-validated values *versus* experimental ones.

The results arising from the cross-validation analysis confirm the linear relationship between $\langle V_{ee} \rangle$ and the aquatic toxicity of the homogeneous subset. Most of the estimated toxicities are close to the experimental values with errors ranging below 10%. A previous study [7] involved a set of 161

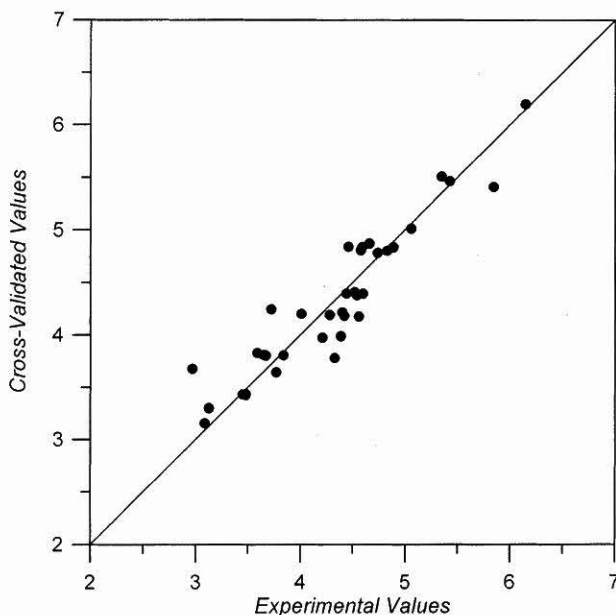


FIGURE 1 Representation of Experimental vs predicted values of the aquatic toxicity ($\log EC_{50}$) for a set of 36 benzene derivatives.

molecules using 5 descriptors ($\log P$, largest positive partial charge on a hydrogen atom, largest negative partial charge on any non-hydrogen atom, HOMO and LUMO energy) yielding to a good correlation, $r^2=0.970$ and $q^2=0.930$. In this work, only the molecules that present a common benzene ring have been tested. If only $\log P$ is used as molecular descriptor, the results are: $r^2=0.694$ and $q^2=0.659$. $\langle V_{ee} \rangle$ contains information about steric and electronic effects, whereas $\log P$ may be related to hydrophobicity and steric effects. As it can be seen in this case, description of electronic effects, as brought out in the previous study by HOMO/LUMO energies, plays an important role. However, it should also be said, that description of the entire molecular set requires the use of more than one descriptor, as performed in the previous study. A similar pattern is found when the same family was studied in our laboratory using a full MQSM technique [32]. A higher molecular subset involving 92 molecules was studied giving $r^2=0.740$ and $q^2=0.716$ using 3 PC. The molecular subset as in the present paper example was also analyzed in the same computational context using 5 PC and yielding to $r^2=0.892$ and $q^2=0.840$. As shown in the mentioned previous studies, description of the entire molecular set, or refinement of the present subset, requires the use of more than one descriptor.

Inhibition of *Bacillus Subtilis* Spore Germination

The next example presented, in order to demonstrate the applicability of $\langle V_{ee} \rangle$ as a MQSM descriptor of toxicity, consists in a set of 19 alcohols, from which 10 of them present straight chain and 9 have a branched chain. The values of toxicity related to inhibition of *Bacillus subtilis* spore germination ($-\log I_{50}$) are described in Table II, and have been obtained from [8]. The results of the linear relationship between calculated $\langle V_{ee} \rangle$ and $-\log I_{50}$ are summarized in Eq. (10). As it can be seen, a fairly good linear QSTR has been obtained using $\langle V_{ee} \rangle$ as a molecular descriptor:

$$-\log I_{50} = 0.882 \cdot \langle V_{ee} \rangle^{(N)} + 2.191 \quad (10)$$

$$n = 19 \quad r^2 = 0.863 \quad s = 0.363 \quad q^2 = 0.828$$

As it can be seen, a sound relationship between molecular toxicity and $\langle V_{ee} \rangle$ can be considered present, and analyzing the value of q^2 , the model in Eq. (10) presents a noticeable predictive capacity. A cross-validation analysis has been performed, as in the previous section, whose results are plotted in Figure 2.

The cross-validation analysis reflects a good linear tendency between experimental and predicted values; even if the residual error is higher than

TABLE II $\langle V_{ee} \rangle$ and $-\log I_{50}$ values for 19 alcohols

Alcohol	$\langle V_{ee} \rangle$	$-\log I_{50}^a$
Methyl	81.07	0.10
Ethyl	134.84	0.59
<i>n</i> -Propyl	196.33	1.03
<i>n</i> -Butyl	263.38	1.77
<i>n</i> -Pentyl	335.28	2.19
<i>n</i> -Hexyl	411.32	2.47
<i>n</i> -Heptyl	490.92	3.10
<i>n</i> -Octyl	573.70	3.40
<i>n</i> -Nonyl	659.30	3.46
<i>n</i> -Decyl	747.47	3.52
2-Butyl	272.50	1.66
Isobutyl	271.06	1.80
tert-Butyl	277.88	1.19
Isopentyl	348.98	2.31
Cyclohexyl	433.79	2.31
2-Octyl	589.36	2.89
3-Octyl	595.93	2.68
2-Ethyl-1-Hexyl	611.28	3.05
Phenethyl	550.50	2.11

^a from reference [8].

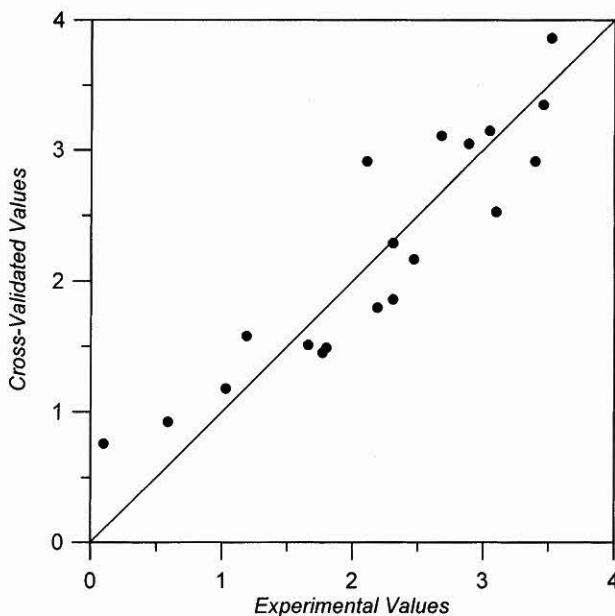


FIGURE 2 Representation of Experimental vs predicted values of the inhibition of spore germination ($-\log I_{50}$) for a set of 19 alcohols.

the previous example. It should be pointed out that when the same molecular set was studied by Hachisuka *et al.* [8] a correlation yielding $r^2 = 0.920$ was obtained using $\log P$ as a single molecular descriptor, however cross-validation was missing in this reference.

Inhibition of Human Sickle Hemoglobin Gelation

The last example presented here consists in two sets of 6 ureas and 6 amides. Biological activity data related to inhibition of Human Sickle Hemoglobin Gelation are obtained from [9] and resumed in Table III along the values of $\langle V_{ee} \rangle$. The studied molecular toxicity, K_S , or Setshenow constant, is a quantitative measure of molar effectiveness. The results of the two linear QSTR between calculated $\langle V_{ee} \rangle$ and K_S for both sets are summarized for ureas in Eq. (11) and for amides in the following Eq. (12).

$$-K_S = 0.293 \cdot \langle V_{ee} \rangle^{(N)} + 0.628 \quad (11)$$

$$n = 6 \quad r^2 = 0.928 \quad s = 0.116 \quad q^2 = 0.831$$

TABLE III $\langle V_{ee} \rangle$ and $-K_S$ values for 6 ureas and 6 amides

Compound	$\langle V_{ee} \rangle$	$-K_S^a$
Ureas		
Urea	203.66	0.25
Methylurea	275.08	0.45
Ethylurea	353.17	0.54
Propylurea	431.83	0.89
Butylurea	513.86	1.16
1,3-Dimethylurea	352.91	0.48
Amides		
Formamide	131.34	0.07
Acetamide	194.51	0.14
Propionamide	264.51	0.23
Butyramide	337.13	0.45
N-methyl-formamide	194.13	0.13
N-ethyl-formamide	264.87	0.26

^a from reference [9].

$$-K_S = 0.119 \cdot \langle V_{ee} \rangle^{(N)} + 0.213 \quad (12)$$

$$n = 6 \quad r^2 = 0.936 \quad s = 0.056 \quad q^2 = 0.748$$

In both sets of molecules, a fairly good correlation is found, as it can be seen from the high values of r^2 , and a considerable predictive capacity, as it is pointed out by q^2 . In addition, these systems were studied by Hansch *et al.* [3] using $\log P$ as molecular descriptor, yielding to $r^2 = 0.951$ for the ureas and $r^2 = 0.983$ for the amides, however in this previous study only 4 molecules in each set were used.

CONCLUSIONS

A series of benzene derivatives, alcohols, amides and ureas have been analyzed for correlation search between experimental toxicity data and Coulomb repulsion expectation values $\langle V_{ee} \rangle$ considered as a MQSM. Very satisfactory correlations have been obtained for all studied series. The results indicate that MQSM, a method to obtain molecular descriptors based on electron densities, constitutes a suitable way for obtaining QSTR and QSPR in general.

Acknowledgments

This research has been partially supported by a CICYT grant: SAF 96-0158, the *Fundació Maria Francisca de Roviralta* and an EU Project: #ENV4-

CT97-0508. One of us, X. Gironés acknowledges University of Girona for a pre-doctoral fellowship.

References

- [1] Fujita, T., Iwasa, J. and Hansch, C. (1964). A new substituent constant, π , derived from partition coefficient. *J. Am. Chem. Soc.*, **86**, 5175–5180.
- [2] Exner, O. (1988). *Correlation Analysis of Chemical Data*, SNTL Publishers of Technical Literature, Prague, p. 275.
- [3] Hansch, C., Kim, D., Leo, A. J., Novellino, E., Silipo, C. and Vittoria, A. (1989). Toward a quantitative comparative toxicology of organic compounds. *CRC Crit. Rev. Toxicol.*, **19**, 185–226.
- [4] Hansch, C. and Leo, A. (1995). *Exploring QSAR. Fundamentals and Applications in Chemistry and Biology*; ACS Professional Reference Book: Washington, DC, p. 557.
- [5] Amat, L., Ponec, R. and Carbó-Dorca, R. (1998). Molecular quantum similarity measures as an alternative to log P values in QSAR studies. *J. Comput. Chem.*, **19**, 1575–1583.
- [6] Ponec, R., Amat, L. and Carbó-Dorca, R. (1999). Molecular basis of quantitative structure-properties relationship (QSPR): A quantum similarity approach. *J. Comp. Aided Molec.*, **13**, 259–270.
- [7] Urrestarazu, E., Vaes, W. H. J., Verhaar, H. J. M. and Hermens, J. L. M. (1998). Quantitative structure-activity of the aquatic toxicity of polar and nonpolar aquatic pollutants. *J. Chem. Inf. Comput. Sci.*, **38**, 845–852.
- [8] Yasuda-Yasaki, Y., Namiki-Kanie, S. and Hachisuka, Y. (1978). Inhibition of *Bacillus subtilis* spore germination by various hydrophobic compounds: Demonstration of hydrophobic character of the *L*-alanine receptor site. *J. Bacteriol.*, **136**, 484–490.
- [9] Poillon, W. N. (1980). Noncovalent inhibitors of sickle hemoglobin gelation: Effects of aliphatic alcohols, amides and ureas. *Biochemistry*, **19**, 3194–3199.
- [10] Carbó, R., Leyda, L. and Arnau, M. (1980). How similar in a molecule to another? An electron density measure of similarity between two molecular structures. *Int. J. Quantum Chem.*, **17**, 1185–1189.
- [11] Carbó, R. and Domingo, L. (1987). LCAO-MO similarity measures and taxonomy. *Int. J. Quant. Chem.*, **23**, 517–545.
- [12] Besalú, E., Carbó, R., Mestres, J. and Solà, M. (1995). Foundations and recent developments on molecular quantum similarity. *Topics Curr. Chem.*, **173**, 31–62.
- [13] Carbó, R., Calabuig, B., Vera, L. and Besalú, E. (1994). Molecular quantum similarity: Theoretical framework, ordering principles and visualization techniques. *Adv. in Quantum Chem.*, **25**, 253–313.
- [14] Carbó, R. (1995). *Molecular Similarity and Reactivity: From Quantum Chemical to Phenomenological Approaches*, Kluwer, Amsterdam, p. 321.
- [15] Carbó-Dorca, R. and Mezey, P. G. (1996). *Advances in Molecular Similarity*, Vol. 1, JAI Press, Greenwich, CT, p. 287.
- [16] Carbó-Dorca, R. and Mezey, P. G. (1998). *Advances in Molecular Similarity*, Vol. 2, JAI Press, Greenwich, CT, p. 297.
- [17] Cioslowski, J. and Fleischmann, E. D. (1991). Assessing molecular similarity from results of *ab initio* electronic structure calculations. *J. Am. Chem. Soc.*, **113**, 64–67.
- [18] Good, A. C., So, S.-S. and Richards, W. G. (1993). Structure-Activity relationships from molecular similarity matrices. *J. Med. Chem.*, **36**, 433–438.
- [19] Mezey, P. G. (1995). Density domain bonding and molecular similarity measures. *Topics Curr. Chem.*, **173**, 63–83.
- [20] Allan, N. L. and Cooper, D. L. (1995). Momentum-space electron density and quantum similarity measures. *Topics Curr. Chem.*, **173**, 85–111.
- [21] Ponec, R. (1995). Similarity models in the theory of pericyclic reactions. *Topics Curr. Chem.*, **174**, 1–26.
- [22] Fradera, X., Amat, L., Besalú, E. and Carbó-Dorca, R. (1997). Application of molecular quantum similarity to QSAR. *Quant. Struct.-Act. Relat.*, **16**, 25–32.

- [23] Lobato, M., Amat, L., Besalú, E. and Carbó-Dorca, R. (1997). Structure-activity relationships of a steroid family using quantum similarity measures and topological quantum similarity indices. *Quant. Struct.-Act. Relat.*, **16**, 465–472.
- [24] Amat, L., Robert, D., Besalú, E. and Carbó-Dorca, R. (1998). Molecular quantum similarity measures. Tuned 3D QSAR: An antitumoral family validation study. *J. Chem. Inf. Comput. Sci.*, **38**, 624–631.
- [25] Robert, D., Amat, L. and Carbó-Dorca, R. (1999). Three-dimensional quantitative structure-activity relationships from tuned molecular quantum similarity measures. Prediction of the corticosteroid-binding globulin binding affinity for a steroid family. *J. Chem. Inf. Comput. Sci.*, **39**, 333–344.
- [26] Carbó-Dorca, R. and Besalú, E. (1998). A general survey of molecular quantum similarity. *Teochem.*, **451**, 11–23.
- [27] Constans, P., Amat, L. and Carbó-Dorca, R. (1997). Towards a global maximization of the molecular similarity function: Superposition of two molecules. *J. Comput. Chem.*, **18**, 826–846.
- [28] See for example: Parr, R. G. (1963). *The Quantum Theory of Molecular Electronic Structure*, W.A. Benjamin, Inc. New York, p. 510.
- [29] Carbó, R., Besalú, E., Amat, L. and Fradera, X. (1995). Quantum molecular similarity measures (QMSM) as a natural way to leading towards a theoretical foundation of quantitative structure-properties relationships (QSPR). *J. Math. Chem.*, **18**, 237–246.
- [30] Allen, D. M. (1974). A relationship between variable selection and data augmentation and a method for prediction. *Technometrics*, **16**, 125–127.
- [31] Frisch, M. J., Trucks, G. W., Schlegel, H. B., Gill, P. M. W., Johnson, B. G., Robb, M. A., Cheeseman, J. R., Keith, T., Petersson, G. A., Montgomery, J. A., Raghavachari, K., Al-Laham, M. A., Zakrzewski, V. G., Ortiz, J. V., Foresman, J. B., Cioslowski, J., Stefanov, B. B., Nanayakkara, A., Challacombe, M., Peng, C. Y., Ayala, P. Y., Chen, W., Wong, M. W., Andres, J. L., Replogle, E. S., Gomperts, R., Martin, R. L., Fox, D. J., Binkley, J. S., Defrees, D. J., Baker, J., Stewart, J. P., Head-Gordon, M., Gonzalez, C. and Pople, J. A. (1995). *Gaussian-94*, (Revision E.2), Gaussian, Inc., Pittsburgh PA.
- [32] Robert, D. and Carbó-Dorca, R. (1999). Aromatic compounds aquatic toxicity. QSAR using molecular quantum similarity measures. *SAR QSAR Environ. Res.*, **10**.



Use of electron-electron repulsion energy as a molecular descriptor in QSAR and QSPR studies

Xavier Gironés, Lluís Amat, David Robert & Ramon Carbó-Dorca

Institute of Computational Chemistry, University of Girona, Campus Montilivi, E-17071 Girona, Spain

Received 29 July 1999; Accepted 7 January 2000

Key words: electron-electron repulsion energy, molecular descriptor, molecular quantum similarity measures, QSAR, QSPR

Summary

Electron-electron repulsion energy ($\langle V_{ee} \rangle$) is presented as a new molecular descriptor to be employed in QSAR and QSPR studies. Here it is shown that this electronic energy parameter is connected to molecular quantum similarity measures (MQSM), and as a consequence can be considered as a complement to steric and electronic parameters in description of molecular properties and biological responses of organic compounds. The present strategy considers the molecule as a whole, thus there is no need to employ contributions of isolated fragments as in many calculations of molecular descriptors, like log P or the Free–Wilson analysis. The procedure has been tested in a widespread set of molecules: alcohols, alkanamides, indole derivatives and 1-alkylimidazoles. Molecular properties, as well as toxicity, are correlated using $\langle V_{ee} \rangle$ as a parameter, and extensions to the method are given for handling difficult systems. In almost all studied cases, satisfactory linear relationships were finally obtained.

Introduction

Due to continuous development of experimental chemistry, a great amount of new compounds are synthesised every year. The major part of these compounds are not tested for fundamental or relevant thermodynamic and physicochemical properties or biological activities, which are to remain unknown due to unavailability or costs of experimental methods. A procedure able to predict, within a reasonable error margin, molecular properties and biological activities for untested compounds is required to evaluate these molecular features in a fast and inexpensive way.

In order to provide a solution to the above mentioned problem, quantitative structure–activity relationships (QSAR) were defined. These models evaluate the dependence of any biological property under study with known physicochemical magnitudes, like in Hansch or the Free–Wilson analysis, based on purely molecular structural features [1–15]. Present

applications are due to the linear free-energy model proposed by Hansch and Fujita [1] in 1964, based on the octanol/water partition coefficient, namely log P , or alternatively on the lipophilicity, π , parameters [1, 3, 6, 9]. Independently, Free and Wilson [2] developed a model based on addition of molecular substructures, and further improvements to QSAR came from combinations of both methods [3, 4, 9]. The major assumption in this kind of QSAR models is the capacity to relate biological activities, commonly on a logarithmic scale, to some addition of molecular features yielding to free-energy models. In this framework, several descriptors, briefly discussed below, satisfying this condition have been developed in addition to log P , π , and group contribution, such as topological [16–19], geometrical, polar [3–6, 9], electronic [20–22] or electrostatic [23] and steric [3–6, 9] parameters.

Topological parameters are intrinsically connected to molecular structures, since they are based on molecular connection matrices (see, for example, [16–19, 24–29]). Geometrical descriptors lead to the idea that molecular activity can be related to molecular

*To whom correspondence should be addressed. E-mail: director@iqc.udg.es

shape and size. Some of the most commonly used geometrical descriptors are molecular volume [30, 31], molecular surface area [32–37] or charged partial surface area [38, 39]. Polar parameters result from computing interactions that arise from dispersion forces, and usually are described by molar refractivity [3–6, 9]. Electronic parameters [20] are based on electronic interactions, as obtained from experimental data, semiempirical or quantum chemistry computations. Electronic descriptors, reflecting, for example, the electrophilic/nucleophilic molecular character (see [20–22] for a more extensive review) can be related to isolated molecules, like Hammett sigma (σ) [40–42]; or to intermolecular interactions, as HOMO/LUMO energy gaps, which are computed from molecular (MO) calculations [43]. Steric parameters can be used as a measure of molecular bulkiness, for example, to evidence the effect of variation of substituents on a common molecular skeleton, like Taft's E_s [44–46], or to describe the importance of substituent width or length in biological responses.

As a final review in QSAR/QSPR technique description, the recent appearance of 3D-QSAR models should be pointed out. These new methods, like CoMFA [47], MS-WHIM [48], E-State [49] and similarity methods, like Molecular Quantum Similarity Measures (MQSM) [50–62], include three-dimensional parameters which embody information about molecular conformation and stereochemistry.

MQSM can be used as an ideal computational tool to solve current chemical problems; several examples of these practical applications are related to the use of MQSM for QSAR and QSPR [63–71]. Such a proposal may be based on the recent description that QSAR/QSPR can be derived from MQSM theory [72]. From this reference, results can be inferred that MQSM provide QSAR models of a well defined theoretical foundation, and in this way, the usual statistical models obtained within MQSM represent QSAR associated to an approximate quantum mechanical relationship. The present work considers the possibility of characterising molecular properties using electron-electron repulsion energy (V_{ee}) as a new kind of MQSM, allowing in this way the use of MQSM even if no specific software is available to compute regular MQSM.

Theoretical background

(A) Molecular quantum similarity measures

Arising from the theoretical foundation of MQSM [50], a general definition may be written as:

$$Z_{AB}(\Omega) = \iint \rho_A(\mathbf{r}_1) \Omega(\mathbf{r}_1, \mathbf{r}_2) \rho_B(\mathbf{r}_2) d\mathbf{r}_1 d\mathbf{r}_2, \quad (1)$$

where $\rho_A(\mathbf{r}_1)$ and $\rho_B(\mathbf{r}_2)$ are density functions of molecules A and B , $\Omega(\mathbf{r}_1, \mathbf{r}_2)$ is a positive definite operator and $Z_{AB}(\Omega)$ is the resulting MQSM. Equation 1 may also be reordered:

$$Z_{AB}(\Omega) = \iint \Omega(\mathbf{r}_1, \mathbf{r}_2) \rho_A(\mathbf{r}_1) \rho_B(\mathbf{r}_2) d\mathbf{r}_1 d\mathbf{r}_2, \quad (2)$$

for any non-differential operator. From here, a further step can be performed considering a molecular quantum self-similarity measure (MQS-SM), constructed as:

$$Z_{AA}(\Omega) = \int \Omega(\mathbf{r}) \rho_A^2(\mathbf{r}) d\mathbf{r}, \quad (3)$$

which in turn, can be alternatively written in the form:

$$Z_{AA}^{(2)}(\omega) = \int \omega(\mathbf{R}) \rho_A^2(\mathbf{R}) d\mathbf{R}, \quad (4)$$

where $\mathbf{R}$ represents a set of particle co-ordinates and $\omega(\mathbf{R})$ some function of them. On the other hand, definition (4) can be easily generated to a multiple MQS-SM:

$$Z_{AA}^{(n)}(\omega) = \int \omega(\mathbf{R}) \rho_A^n(\mathbf{R}) d\mathbf{R}, \quad (5)$$

which in the most simple case, $n = 1$, becomes:

$$Z_{AA}^{(1)}(\omega) = \int \omega(\mathbf{R}) \rho_A(\mathbf{R}) d\mathbf{R} \quad (6)$$

Equation 6 shows that quantum mechanical expectation values for non-differential operators can be considered as particular cases of MQS-SM, that is:

$$\langle \omega(\mathbf{R}) \rangle = Z_{AA}^{(1)}(\omega). \quad (7)$$

(B) Coulomb and exchange operators expectation values: Electron-electron repulsion energy

At the monoconfigurational ground closed shell discrete Hartree–Fock level, the Coulomb operator expectation value for an n -electron system can be written [73] as:

$$\langle V_{ee} \rangle = \sum_i \sum_j [2 \cdot (ii|jj) - (ij|ij)], \quad (8)$$

where $\{(ii|jj)\}$ and $\{(ij|ij)\}$ are the respective Coulomb and exchange integrals over the MO basis set.

Also, $\langle V_{ee} \rangle$ can be interpreted as a sum of similarity measures over the occupied MO set [51]. Starting from this previous point of view, $\langle V_{ee} \rangle$ can be considered an n -body self-similarity measure involving the positive definite Coulomb operator, $C(\mathbf{R})$, defined as follows:

$$C(\mathbf{R}) = \sum_{i>j} |\mathbf{r}_i - \mathbf{r}_j|^{-1} \wedge \mathbf{R} = (\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_n), \quad (9)$$

using a system density function, constructed in the usual way:

$$\rho_A(\mathbf{R}) = |\Psi_A(\mathbf{R})|^2, \quad (10)$$

where $\Psi_A(\mathbf{R})$ is the system's approximate wavefunction. So, an equivalent expression of Equation 8 appears to be written as:

$$\langle V_{ee} \rangle = \int C(\mathbf{R}) \rho_A(\mathbf{R}) d\mathbf{R}. \quad (11)$$

Equation 11 has the same form as Equation 6, which, in turn, was deduced within an MQS-SM framework. In this manner, it can be said that the evaluation of Coulomb operator expectation values constitutes a calculation of a first-order MQS-SM class, and thus opens a new field for the potential use of this measure as a molecular descriptor.

Simple linear model using $\langle V_{ee} \rangle$

Given a set of molecules, a molecular property can be associated to the expectation value of some unknown operator [72]. This general quantum mechanical definition can be approximated in a discrete framework constructing an $(n \times n)$ symmetric similarity matrix defined as: $\mathbf{Z} = \{Z_{AB}\}$, whose elements are built up using the MQSM between pairs of molecules as described in Equation 1. When homogeneous molecular series are studied, off-diagonal terms can be considered as constants, and a single vector that contains MQS-SM is used as the unique molecular descriptor [67, 68] for approximate molecular expectation values evaluation.

In this work, it is intended that $\langle V_{ee} \rangle$ be used as an alternative MQS-SM, and so to find a linear relationship between this parameter and the molecular properties, such as:

$$\pi_I \cong a \cdot \langle V_{ee} \rangle_I + b. \quad (12)$$

As presented in a previous work, MQS-SM have been proposed as an alternative to $\log P$ values in QSAR studies [68]. Here it is also suggested that MQS-SM based on $\langle V_{ee} \rangle$ can complement empirical parameters where steric and electronic effects play an important role. $\langle V_{ee} \rangle$ is strongly connected with the number of atoms and their type, hence $\langle V_{ee} \rangle$ is related to molecular volume. As it is based on quantum mechanical ideas, $\langle V_{ee} \rangle$ reflects the part of intramolecular interactions related to electron-electron repulsion energy.

Results and discussion

In this section it is intended that the usefulness of this new proposed methodology be proved. First of all, molecular modelling and the statistical considerations are explained. Then, an example that connects this new procedure with a previous use of MQS-SM in QSPR is presented. Finally, an assorted set of examples, involving molecular properties, biological activity and anomalous examples, is given.

(A) Molecular modelling and statistical considerations

The geometries of all compounds in this study have been considered in the gas phase and optimised according to the AM1 Hamiltonian using Ampac 6.01 [74]. $\langle V_{ee} \rangle$ values have been calculated with Gaussian 94 [75] within the HF/3-21G* computational level over the AM1 optimised geometry.

In addition, all values of $\langle V_{ee} \rangle$, $\{\langle V_{ee} \rangle_I\}$ have been standardised in the statistical sense and then symbolised by: $\langle V_{ee} \rangle_I^{(N)}$. That is, $\langle V_{ee} \rangle_I^{(N)} = (\langle V_{ee} \rangle_I - \langle \bar{V}_{ee} \rangle) \cdot s^{-1}$, where $\langle \bar{V}_{ee} \rangle$ is the arithmetic mean and s is the standard deviation of the various $\{\langle V_{ee} \rangle_I\}$ values associated to the molecular set.

(B) Connection with MQS-SM

In order to illustrate the validity of the hypothesis formulated in Equation 12, a relationship between $\langle V_{ee} \rangle_I^{(N)}$ and $\log P$ is presented for a set of amides. Compounds, $\{\langle V_{ee} \rangle_I\}$ and $\log P$ are listed in Table 1. In addition, a cross-validation study is performed, whose results are also included in Table 1, to test the predictive capacity of the obtained relationship. In general terms, this process consists of leaving each element out of the model construction for later prediction of

Table 1. $\langle V_{ee} \rangle$ and $\log P$ values for a set of amides

Molecule	$\langle V_{ee} \rangle$	$\log P^a$	Cross-validated $\log P$
Formamide	131.33	-1.51	-1.46
Acetamide	194.51	-1.26	-0.99
Propylamide	264.41	-0.66	-0.55
Butylamide	337.14	-0.21	-0.05
Pentylamide	413.94	0.33	0.49
Hexylamide	493.73	0.87	1.10
N-methylformamide	191.63	-0.97	-1.09
N-ethylformamide	258.85	-0.43	-0.63
N-propylformamide	330.24	0.11	-0.14
N-butylformamide	405.91	0.65	0.34

^aFrom Reference 68.

its value [76]. The equation of the obtained model, as well as relevant statistical parameters: r^2 (coefficient of determination), s (root mean square for error between experimental and predicted values) and q^2 (predictive capacity parameter) are presented below.

$$\log P = 0.746 \cdot \langle V_{ee} \rangle^{(N)} - 0.308,$$

$$n = 10, r^2 = 0.955, s = 0.197, q^2 = 0.931$$

The good correlation observed, $r^2 = 0.955$, confirms the actual presence of a sound relationship between $\log P$ and $\langle V_{ee} \rangle$. This example also shows the close relation between $\langle V_{ee} \rangle$ and molecular size. As can be seen in Table 1, $\langle V_{ee} \rangle$ presents an additive behaviour, as $\log P$ does for homologous series, and thus increasing their respective values with the augmenting molecular size. In the previous study, a good correlation was also found using the overlap octanol-water MQS-SM, defined in [68], as a molecular descriptor, $r^2 = 0.954$, pointing out that both parameters contain almost the same information.

(C) Molecular examples

As mentioned above, three series of examples are presented. The first case consists of a relationship between $\langle V_{ee} \rangle$ and molecular properties. For this case, a set of 15 alcohols are tested for correlation with melting ($T_{m.p.}$) and boiling points ($T_{b.p.}$). The next example is related to toxicology, as a particular example of biological activity of large interest in the scientific community. Here, the relationship between $\langle V_{ee} \rangle$ and 50% inhibitory growth concentration ($-\log GC_{50}$) of *Tetrahymena pyriformis*, a ciliate often used to test ecotoxicology, is presented for a set of alkanamines.

Table 2. Values of $\langle V_{ee} \rangle$, $T_{m.p.}^2$ and $T_{b.p.}^2$ for a set of alcohols, all temperatures are given in Kelvin (K)

Alcohols	$\langle V_{ee} \rangle$	$T_{m.p.}^2$		$T_{b.p.}^2$	
		Exp. ^a	Pred.	Exp. ^a	Pred.
Methyl	81.07	30821.31	25459.32	114081.82	119433.95
Ethyl	135.21	25300.08	31081.85	123453.85	130076.48
Propyl	196.33	21626.64	36353.32	137166.53	141513.47
Butyl	263.38	33620.89	39480.91	152771.54	154125.36
Pentyl	335.28	37736.95	44289.67	168970.32	167862.25
Hexyl	411.32	52239.67	48417.68	185554.18	182536.32
Heptyl	490.92	57197.51	53868.74	202104.19	198006.25
Octyl	573.71	66388.68	59213.94	219267.43	214098.48
Nonyl	659.30	71909.79	65015.46	236643.33	230739.12
Decyl	747.48	78433.60	70883.90	254278.15	247862.33
Undecyl	837.97	85357.47	76820.61	266421.15	265919.04
Dodecyl	930.62	88304.07	83268.95	283194.27	284113.39
Tetradecyl	1121.73	97756.28	96463.47	316023.87	322277.00
Hexadecyl	1319.77	103980.45	112079.30	368643.27	356904.69
Heptadecyl	1421.12	106902.84	121390.00	367429.95	385165.64

^aFrom Reference 77.

Finally, a number of examples where no linear correlation can be found are handled. The first anomalous example consists of displacement of flunitrazepam from binding to bovine brain membrane. In this case the studied molecular set is formed by indole derivatives and it has been found that expectation value scaling solves the problem. The second difficult case is related to a set of 1-alkylimidazoles, where two biological activities are studied: aldrin epoxidation in armyworm gut ($-\log I_{50}$) and binding affinity to rat liver cytochrome P-450 ($-\log K_s$). A quadratic relationship, which is theoretically justified, is found to solve the problem.

(C.1) $\langle V_{ee} \rangle$ vs. molecular properties

The first example presented, in order to demonstrate the applicability of $\{\langle V_{ee} \rangle\}$ as a descriptor of molecular properties, consists of a series of 15 aliphatic alcohols. The values of $\langle V_{ee} \rangle$, $T_{b.p.}^2$ and $T_{m.p.}^2$ for the set of alcohols [77], including the predicted values for both properties, are summarised in Table 2 and the equation of the obtained model, including its statistical parameters, is presented below for $T_{m.p.}^2$:

$$T_{m.p.}^2 = 27634 \cdot \langle V_{ee} \rangle^{(N)} - 63838,$$

$$n = 15, r^2 = 0.947, s = 7689.40, q^2 = 0.924$$

Table 3. Compounds, $\langle V_{ee} \rangle$, experimental and predicted $-\log GC_{50}$ for a set of amines

Molecule	$\langle V_{ee} \rangle$	$-\log GC_{50}$	
		Exp. ^a	Pred.
Isopropylamine	190.98	-0.88	-1.22
t-Butylamine	266.65	-0.90	-0.81
Propylamine	188.56	-0.71	-1.25
s-Butylamine	260.68	-0.67	-0.86
Butylamine	255.43	-0.57	-0.89
t-Amylamine	345.85	-0.70	-0.42
1-Ethyl-propylamine	336.33	-0.81	-0.46
1-Methyl-butylamine	333.63	-0.69	-0.48
Isoamylamine	333.42	-0.58	-0.49
Amylamine	324.66	-0.48	-0.54
2-Methylbutylamine	337.08	-0.48	-0.47
3-Phenyl-1-propylamine	616.32	0.28	-0.94
Hexylamine	402.83	-0.22	-0.15
4-Phenyl-1-butylamine	709.94	0.62	1.47
Heptylamine	485.12	0.21	0.25
Octylamine	568.59	0.61	0.65
Nonylamine	657.13	1.70	0.99
Decylamine	746.31	2.06	1.39
Unidecylamine	840.19	2.33	1.83
1,1-Dimethylpropargylamine	308.29	-0.91	-0.60
N-Methylpropylamine	256.91	-0.81	-0.87
N-Methylbutylamine	328.94	-0.68	-0.51
N-methylpropargylamine	224.79	-0.98	-1.03
N,N-dimethylethylamine	263.29	-0.91	-0.83

^aFrom Reference 78.

and for $T_{b,p}^2$:

$$T_{b,p}^2 = 79849 \cdot \langle V_{ee} \rangle^{(N)} - 226.40 \times 10^3,$$

$$n = 15, r^2 = 0.995, s = 6846.88, q^2 = 0.993$$

As can be seen from the results, in both cases a fairly good linear relationship is obtained when using $\langle V_{ee} \rangle$ as a single descriptor. The values of r^2 for the two properties are quite high, 0.947 and 0.995, indicating that there is a correlation present. In addition, predicted values are also quite close to experimental ones.

In this case, the relationship between $\langle V_{ee} \rangle$ and molecular size is also evidenced. As is well known, boiling and melting points for homologous series are very much influenced by the effect of molecular weight, and $\langle V_{ee} \rangle$ simulates that effect.

(C.2) $\langle V_{ee} \rangle$ vs. toxicity

In this example, the utility of $\langle V_{ee} \rangle$ as a molecular descriptor for toxicity is demonstrated. In this case, the

studied molecular set consists of 24 alkanamides [78], from which 20 are primary, 3 are secondary and 1 tertiary. The values of $\langle V_{ee} \rangle$, experimental and predicted $-\log GC_{50}$ are presented in Table 3, and the correlation obtained for this system, as well as all relevant statistical parameters, are presented below:

$$-\log GC_{50} = 0.896 \cdot \langle V_{ee} \rangle^{(N)} - 0.174,$$

$$n = 24, r^2 = 0.885, s = 0.346, q^2 = 0.847$$

The model obtained presents an acceptable correlation between $\langle V_{ee} \rangle$ and $-\log GC_{50}$, and the predicted values are quite close to experimental ones. The previous study [78] yielded slightly improved results using $\log P$, $r^2 = 0.934$, however, it only involved a molecular subset of 20 primary amines.

This example also reveals the correspondence of $\langle V_{ee} \rangle$ with electronic effects. Several compounds form the studied molecular set with the same structural formula but different connectivity. If $\langle V_{ee} \rangle$ were only connected with molecular size, it would not be able to differentiate these compounds.

(C.3) Extensions to the method

In some cases, it is found that no correlation can be directly found between $\langle V_{ee} \rangle$ and any biological activity. This fact can be considered as common, since prediction of molecular activity with a single or small number of descriptors is still a hard task. However, some of these problems can be easily solved, as will be presented in this section.

(C.3.1) *Scaling of $\langle V_{ee} \rangle$.* When dealing with molecular sets presenting heavy atom substitutions, such as chlorines or bromines, the values of $\langle V_{ee} \rangle$ are highly increased with respect to the molecules that do not present such substitutions. In order to obtain a smooth distribution of $\langle V_{ee} \rangle$ values, a scaling factor is applied. This scaling factor has been chosen to be the square of the number of electrons (n_e^2). In this way, all values of $\langle V_{ee} \rangle$ are to be divided by n_e^2 , and the resulting value $\langle V_{ee}^Y \rangle$, which is also standardised to $\langle V_{ee}^Y \rangle^{(N)}$, is tested for correlation.

The example presented to test this scaling procedure consists of a set of 20 indole derivatives, for which it had been found that no direct correlation between $\langle V_{ee} \rangle$ and $-\log K_I$ is present, yielding $r^2 = 0.574$. When the normalised $\langle V_{ee}^Y \rangle$ descriptor is applied, the resulting model improves considerably, as can be seen from the statistical parameters below. All data concerning $\langle V_{ee} \rangle$, number of electrons, experimental [79] and predicted $-\log K_I$ and the basic

Table 4. Basic molecular structure, compounds, $\langle V_{ee} \rangle$, experimental and predicted biological activity for a set of indole derivatives

Compound				$\langle V_{ee} \rangle$	n_e	$-\log K_1$	
R_1	R_2	R_3	R_4			Exp. ^a	Pred.
Cl	H	H	H	2177.44	162	6.21	5.96
NO ₂	H	H	H	2209.08	168	6.93	6.72
H	OCH ₃	H	H	2057.37	162	6.78	6.71
Cl	OCH ₃	H	H	2465.27	178	6.68	6.81
NO ₂	OCH ₃	H	H	2500.32	184	7.27	7.44
H	H	OCH ₃	H	2056.47	162	6.54	6.73
Cl	H	OCH ₃	H	2464.26	178	6.79	6.81
NO ₂	H	OCH ₃	H	2499.60	184	7.42	7.42
H	OCH ₃	OCH ₃	H	2372.43	178	7.03	7.28
Cl	OCH ₃	OCH ₃	H	2790.32	194	7.52	7.35
NO ₂	OCH ₃	OCH ₃	H	2829.30	200	7.96	7.88
H	H	H	Cl	2207.63	162	5.59	5.89
H	OH	H	H	1924.67	154	6.37	6.27
Cl	OH	H	H	2327.40	170	6.82	6.34
H	H	OH	H	1925.67	154	6.09	6.30
Cl	H	OH	H	2328.55	170	6.24	6.39
NO ₂	H	OH	H	2362.13	176	7.19	7.03
H	OH	OH	H	2082.22	162	6.46	6.57
Cl	OH	OH	H	2490.17	178	6.75	6.68
NO ₂	OH	OH	H	2525.53	184	7.32	7.30

^aFrom Reference 79.

molecular structure are presented in Table 4.

$$-\log K_1 = -0.519 \cdot \langle V_{ee}^Y \rangle^{(N)} + 6.798,$$

$$n = 20, r^2 = 0.903, s = 0.179, q^2 = 0.879$$

The results indicate that, when $\langle V_{ee}^Y \rangle$ is applied, a fair correlation can be obtained. In this case, a correct relationship is obtained, $r^2 = 0.903$, with a considerable prediction capacity, as pointed out by the coefficient: $q^2 = 0.879$. However, the original set contained 23 molecules, obtaining a poor correlation, $r^2 = 0.603$, with a reasonable predictive capacity, $q^2 = 0.507$. The previous study [79], where Hammett's sigma was used as a single descriptor, obtained slightly poorer results; up to three molecules were also extracted from

the full set, yielding $r^2 = 0.498$. Later addition of discrete descriptors depending on fragment presence increased to $r^2 = 0.728$ for one extra descriptor and $r^2 = 0.810$ for two extra descriptors, keeping a set of 20 molecules and providing no information about the predictive capacity.

In this example, the electronic effects are clearly evidenced, as the molecular size is not relevant in this biological process, but the electronic effects provided by the different substitutions at different locations are. $\langle V_{ee}^Y \rangle$ reflects these electronic effects and it is able to characterise this biological activity within the subset chosen.

(C.3.2) *Polynomial correlation using $\langle V_{ee} \rangle$.* In some cases, it can be seen by direct observation of the data that there exists a correlation between $\langle V_{ee} \rangle$ and some biological property. However, usage of a single or a small number of parameters is not enough to construct a valid linear model. The theoretical origin of this situation could be the fact that the ordinate at the origin, b , in Equation 12, can no longer be considered a constant. In this case, the proposed linear model can be extended to a general polynomial formulation, in order to take into account somehow the fluctuations in the off-diagonal MQSM, affecting the parameter b . Under these circumstances, the following equation can be used:

$$\pi_A \cong \sum_{p=1}^N a_p \cdot \langle V_{ee} \rangle_A^p + \beta. \quad (13)$$

This methodology is very close to the one proposed by Hansch and Fujita when they published the linear free-energy related model [1], using not only a single polynomial descriptor formulation, but many parameters, such as $\log P$ and Hammett's sigma, resulting in the so-called parabolic model. Further descriptor combinations and mathematical transformations, not applied here, can be performed (see, for example, [80] for assorted examples).

The example used to illustrate such a proposed situation consists of a set of 13 1-alkylimidazoles, in which two biological activities have been tested: $-\log I_{50}$ and $-\log K_S$. The results obtained for quadratic relationships are presented below with all relevant statistical parameters. All data concerning compounds, $\langle V_{ee} \rangle$, biological activities [81] and predicted values are presented in Table 5.

Table 5. Compounds, $\langle V_{ee} \rangle$, biological activities and predicted values for a set of 1-alkylimidazoles

Alkyl group	$\langle V_{ee} \rangle$	$-\log I_{50}^a$		$-\log K_s^b$	
		Exp. ^c	Pred.	Exp. ^c	Pred.
H	238.99	2.53	2.01	3.68	3.84
CH ₃	314.30	2.90	3.20	3.70	4.25
CH ₂ CH ₃	394.39	3.27	4.11	4.68	4.38
(CH ₂) ₂ CH ₃	474.88	4.85	4.71	4.94	4.68
(CH ₂) ₃ CH ₃	558.46	5.58	5.38	5.03	4.96
(CH ₂) ₄ CH ₃	643.95	6.41	5.94	5.39	5.15
(CH ₂) ₅ CH ₃	731.99	6.87	6.43	5.47	5.35
(CH ₂) ₆ CH ₃	821.99	6.94	6.86	5.56	5.49
(CH ₂) ₇ CH ₃	914.17	7.02	7.17	5.49	5.62
(CH ₂) ₈ CH ₃	1008.18	7.09	7.32	5.41	5.69
(CH ₂) ₉ CH ₃	1104.06	6.99	7.35	5.46	5.67
(CH ₂) ₁₁ CH ₃	1300.79	6.75	6.93	5.35	5.49
(CH ₂) ₁₃ CH ₃	1503.63	6.08	5.29	5.24	4.50

^aArmyworm gut.

^bRat liver.

^cFrom Reference 81.

$$-\log I_{50} = -1.034 \cdot (\langle V_{ee} \rangle^{(N)})^2 + 1.624 \cdot \langle V_{ee} \rangle^{(N)} + 6.670,$$

$$n = 13, r^2 = 0.965, s = 0.231, q^2 = 0.931$$

$$-\log K_s = -0.399 \cdot (\langle V_{ee} \rangle^{(N)})^2 + 0.588 \cdot \langle V_{ee} \rangle^{(N)} + 5.430,$$

$$n = 13, r^2 = 0.900, s = 0.156, q^2 = 0.745$$

In the armyworm case, it can be seen that the relationship found describes the system well, yielding $r^2 = 0.965$, and as can be observed in Table 5, the predicted values fit the experimental ones with low error, also evidenced by the predictive capacity, $q^2 = 0.931$. However, although an acceptable relationship results for the rat liver case, $r^2 = 0.900$, the obtained results are poorer, the error increases slightly in the predicted property, and as a result, the predictive capacity of the model is lower, yielding $q^2 = 0.745$. The previous study [81] obtained slightly better results, $r^2 = 0.966$ for the armyworm gut case and $r^2 = 0.923$ for the rat liver model using also a quadratic regression analysis with π as a descriptor.

As can be seen in this example, $\langle V_{ee} \rangle$ acts mainly as a descriptor of molecular size, even if the biological property at first increases and later decreases. This fact can be easily monitored by $\langle V_{ee} \rangle$ because of the negative coefficient of the quadratic term present in the

correlation equation. At the same time, it can be seen that π and $\langle V_{ee} \rangle$ are fairly equivalent in this molecular set, as they achieve comparable results. Thus one can conclude that π can be safely substituted, in this case, by $\langle V_{ee} \rangle$.

Conclusions

Molecular sets composed of alcohols, alkanamides, indole derivatives and 1-alkylimidazoles have been analysed for a correlation search between experimental properties, biological activities, toxicology data and $\langle V_{ee} \rangle$, considered as an MQS-SM. Acceptable correlations have been obtained for all studied series, including those which have been handled by extensions of the method. The results indicate that MQSM, a general method based on quantum mechanical electron densities instead of the conventional atomic/fragment contribution approaches, may constitute a suitable alternative way of complementing or quantifying QSPR/QSAR in general.

Acknowledgements

This research has been partially supported by the Fundació Maria Francisca de Roviralta and an EU contract: #ENV4-CT97-0508. One of us, X.G., wishes to acknowledge the University of Girona for a predoctoral fellowship. The authors would also like to thank Ms. M. Crous for her bibliographical help supplies. Some comments by the evaluators have been helpful to ameliorate the structure of this paper, hence the authors wish to express their gratitude to them.

References

1. Hansch, C. and Fujita, T., *J. Am. Chem. Soc.*, 86 (1964) 1616.
2. Free, S.M., Jr and Wilson, J.W., *J. Med. Chem.*, 7 (1964) 395.
3. Ramsden, C.A. (Ed.), *Comprehensive Medicinal Chemistry*, Vol. 4, Pergamon, Oxford, 1990.
4. Kubinyi, H., *QSAR: Hansch Analysis and Related Approaches*, VCH, Weinheim, 1993.
5. Hansch, C. and Leo, A., *Exploring QSAR. Fundamentals and Applications in Chemistry and Biology*, American Chemical Society, Washington, DC, 1995.
6. Hansch, C., Leo, A. and Hoekman, D., *Exploring QSAR. Hydrophobic, Electronic and Steric Constants*, American Chemical Society, Washington, DC, 1995.
7. Van de Waterbeemd, H. (Ed.), *Chemometric Methods in Molecular Design*, VCH, Weinheim, 1995.
8. Van de Waterbeemd, H. (Ed.), *Advanced Computer-Assisted Techniques in Drug Discovery*, VCH, Weinheim, 1995.

9. Kubinyi, H., In Wolff, M.E. (Ed.) *Burger's Medicinal Chemistry*, 5th edn., Vol I, Wiley, New York, NY, 1995.
10. Sanz, F., Giraldo, J. and Manaut, F. (Eds), *QSAR and Molecular Modelling: Concepts, Computational Tools and Biological Applications*, Proc. 10th European Symp. on Quantitative Structure-Activity Relationships, Barcelona, 1994, Prous Science, Barcelona, 1995.
11. Pliska, V., Testa, B. and Van de Waterbeemd, H. (Eds), *Lipophilicity in Drug Action and Toxicology*, VCH, Weinheim, 1996.
12. Van de Waterbeemd, H. (Ed.), *Structure-Activity Correlations in Drug Research*, Academic, R.G. Landes Company, Austin, TX, 1996.
13. Van de Waterbeemd, H., In Wermuth, C.G. (Ed.) *The Practice of Medicinal Chemistry*, Academic, London, 1996.
14. Böhm, H.-J., Klebe, G. and Kubinyi, H., *Wirkstoffdesign*, Spektrum Akademischer Verlag, Heidelberg, 1996.
15. Van de Waterbeemd, H., Testa, B. and Folkers, G. (Eds), *Computer-Assisted Lead Finding and Optimization*, Proc. 11th European Symp. on Quantitative Structure-Activity Relationships, Lausanne, 1996, Verlag Helvetica Chimica Acta, Basel, and VCH, Weinheim, 1997.
16. Wiener, H., *J. Am. Chem. Soc.*, 69 (1947) 17.
17. Hosoya, H., *Bull. Chem. Soc. Jpn.*, 44 (1971) 2332.
18. Kier, L.B. and Hall, L.H., *Molecular Connectivity in Chemistry and Drug Research*, Academic, New York, NY, 1976.
19. Kier, L.B. and Hall, L.H., *Molecular Connectivity in Structure-Activity Analysis*, Research Studies Press, Letchworth, 1986.
20. Van de Waterbeemd, H. and Testa, B., *Adv. Drug. Res.*, 16 (1987) 85.
21. Purcell, W.P., Bass, G.E. and Clayton, J.M., *Strategy of Drug Design*, Wiley, New York, NY, 1973.
22. Karelson, M., Lobanov, V.S. and Katritzky, A.R., *Chem. Rev.*, 96 (1996) 1027.
23. Carbó, R., Martín, M. and Pons, V., *Afinidad*, 34 (1977) 348.
24. Brown, R.D. and Martin, Y.C., *J. Chem. Inf. Comput. Sci.*, 36 (1996) 572.
25. Brown, R.D. and Martin, Y.C., *J. Chem. Inf. Comput. Sci.*, 37 (1997) 1.
26. Doman, T.N., Cisulskis, J.M., Cisulskis, M.J., McCray, P.D. and Spangler, D.P., *J. Chem. Inf. Comput. Sci.*, 36 (1996) 1195.
27. Klopman, G., Balthasar, D.M. and Rosenkranz, H.S., *Environ. Toxicol. Chem.*, 12 (1990) 231.
28. Rosenkranz, H.S. and Klopman, G., *Mutation Res.*, 228 (1990) 105.
29. Randic, M., *J. Am. Chem. Soc.*, 97 (1975) 6609.
30. Bondi, A., *J. Phys. Chem.*, 68 (1964) 441.
31. Pearlman, R.S., In Yalkowsky, S.H., Sikula, A.A. and Valvani, S.C. (Eds) *Physical Chemical Properties of Drugs*, Vol. 10, Marcel Dekker, New York, NY, 1980.
32. Pearlman, R.S., *Quantum Chem. Prog. Exchange Bull.*, 1 (1981) 15.
33. Pearlman, R.S., In Dunn, W.J., Block J.H. and Pearlman, R.S. (Eds) *Partition Coefficient Determination and Estimation*, Pergamon, New York, NY, 1986.
34. Camilleri, P., Watts, A. and Boraston, J.A., *J. Chem. Soc., Perkin. Trans. 2*, (1988) 1699.
35. Lee, B. and Richards, F.M., *J. Mol. Biol.*, 55 (1971) 379.
36. Hermann, R.B., *J. Phys. Chem.*, 76 (1972) 2754.
37. Pearlman, R., SAREA, Quantum Chemistry Program Exchange, University of Indiana, Bloomington, IN, Program Number 432.
38. Grigoras, S., *J. Comput. Chem.*, 11 (1990) 493.
39. Howel, J., Rossi, A., Wallace, D., Hiraki, K. and Hoffman, R., FORTICON 8, Quantum Chemistry Program Exchange, University of Indiana, Bloomington, IN, Program Number 469.
40. Hammett, L.P., *J. Am. Chem. Soc.*, 59 (1937) 96.
41. Hammett, L.P., *Physical Organic Chemistry*, McGraw-Hill, New York, NY, 1940.
42. Unger, S.H. and Hansch, C., *J. Med. Chem.*, 16 (1973) 745.
43. Streitwieser, A., *Molecular Orbital Theory for Organic Chemists*, Wiley, New York, NY, 1961.
44. Taft, R.W., *J. Am. Chem. Soc.*, 86 (1968) 5175.
45. Taft, R.W., In Newman, M.S. (Ed.) *Steric Effects in Organic Chemistry*, Wiley, New York, NY, 1956.
46. Swain, C.G. and Lupton Jr., E.C., *J. Am. Chem. Soc.*, 90 (1968) 4328.
47. Cramer III, R.D., Patterson, D.E. and Bunce, J.D., *J. Am. Chem. Soc.*, 110 (1988) 5959.
48. Bravi, G., Gancia, E., Mascagni, P., Pegna, M., Todeschini, R. and Zalianni, A., *J. Comput.-Aided Mol. Design*, 11 (1997) 79.
49. Kellogg, G.E., Kier, L.B., Gaillard, P. and Hall, L.H., *J. Comput.-Aided Mol. Design*, 10 (1996) 513.
50. Carbó, R., Leyda, L. and Arnau, M., *Int. J. Quant. Chem.*, 17 (1980) 1185.
51. Carbó, R. and Domingo, L.I., *Int. J. Quant. Chem.*, 23 (1987) 517.
52. Besalú, E., Carbó, R., Mestres, J. and Solà, M., *Top. Curr. Chem.*, 173 (1995) 31.
53. Carbó, R., Calabuig, B., Vera, L. and Besalú, E., *Adv. Quant. Chem.*, 25 (1994) 253.
54. Carbó, R. (Ed.), *Molecular Similarity and Reactivity: From Quantum Chemical to Phenomenological Approaches*, Kluwer, Amsterdam, 1995.
55. Carbó-Dorca, R. and Mezey, P.G. (Eds), *Advances in Molecular Similarity*, Vol. 1, JAI Press, Greenwich, CT, 1996.
56. Carbó-Dorca, R. and Mezey, P.G. (Eds), *Advances in Molecular Similarity*, Vol. 2, JAI Press, Greenwich, CT, 1998.
57. Cioslowski, J. and Fleischmann, E.D., *J. Am. Chem. Soc.*, 113 (1991) 64.
58. Burt, C., Richards, W.G. and Huxley, P., *J. Comput. Chem.*, 10 (1990) 1139.
59. Mezey, P.G., *Top. Curr. Chem.*, 173 (1995) 63.
60. Allan, N.L. and Cooper, D.L., *Top. Curr. Chem.*, 173 (1995) 85.
61. Ponec, R., *Top. Curr. Chem.*, 174 (1995) 1.
62. Carbó-Dorca, R. and Besalú, E., *J. Mol. Struct.*, 451 (1998) 11.
63. Fradera, X., Amat, L., Besalú, E. and Carbó-Dorca, R., *Quant. Struct.-Act. Relat.*, 16 (1997) 25.
64. Lobato, M., Amat, L., Besalú, E. and Carbó-Dorca, R., *Quant. Struct.-Act. Relat.*, 16 (1997) 465.
65. Amat, L., Robert, D., Besalú, E. and Carbó-Dorca, R., *J. Chem. Inf. Comput. Sci.*, 38 (1998) 624.
66. Robert, D., Amat, L. and Carbó-Dorca, R., *J. Chem. Inf. Comput. Sci.*, 39 (1999) 333.
67. Ponec, R., Amat, L. and Carbó-Dorca, R., *J. Comput.-Aided Mol. Design*, 13 (1999) 259.
68. Amat, L., Carbó-Dorca, R. and Ponec, R., *J. Comput. Chem.*, 19 (1998) 1575.
69. Ponec, R., Amat, L. and Carbó-Dorca, R., *J. Phys. Org. Chem.*, 12 (1999) 447.

70. Amat, L., Carbó-Dorca, R. and Poncet, R., Simple Linear QSAR Models based on Quantum Similarity Measures, *J. Med. Chem.*, 42 (1999) 5169.
71. Robert, D. and Carbó-Dorca, R., SAR QSAR Environ. Res., 10 (1999) 401.
72. Carbó, R., Besalú, E., Amat, L. and Fradera, X., *J. Math. Chem.*, 18 (1995) 237.
73. Roothaan, C.C.J., *Rev. Mod. Phys.*, 23 (1951) 69.
74. Ampac 6.01, 1994, Semichem, Shawnee, KS.
75. Frisch, M.J., Trucks, G.W., Schlegel, H.B., Gill, P.M.W., Johnson, B.G., Robb, M.A., Cheeseman, J.R., Keith, T., Petersson, G.A., Montgomery, J.A., Raghavachari, K., Al-Laham, M.A., Zakrzewski, V.G., Ortiz, J.V., Foresman, J.B., Cioslowski, J., Stefanov, B.B., Nanayakkara, A., Challacombe, M., Peng, C.Y., Ayala, P.Y., Chen, W., Wong, M.W., Andres, J.L., Replogle, E.S., Gomperts, R., Martin, R.L., Fox, D.J., Binkley, J.S., Defrees, D.J., Baker, J., Stewart, J.P., Head-Gordon, M., Gonzalez, C. and Pople, J.A., Gaussian-94 (Revision E.2), Gaussian, Inc., Pittsburgh, PA, 1995.
76. Allen, D.M., *Technometrics*, 16 (1974) 125.
77. Lide, D.R., *Handbook of Chemistry and Physics*, 76th Edition, CRC Press, Boca Raton, FL, 1995.
78. Sinks, G.D., Carver, T.A. and Schultz, W., SAR QSAR Environ. Res., 9 (1998) 217.
79. Hadjipavlou-Litina, D. and Hansch, C., *Chem. Rev.*, 94 (1994) 1483.
80. Hansch, C., Kim, D., Leo, A.J., Novellino, E., Silipo, C. and Vittoria, A., *CRC Crit. Rev. Toxicol.*, 19 (1989) 185.
81. Wilkinson, W.C.F., Hetnarsky, K., Cantwell, P. and di Carlo, F. J., *Biochem. Pharmacol.*, 23 (1974) 2377.



Antimalarial activity of synthetic 1,2,4-trioxanes and cyclic peroxy ketals, a quantum similarity study

X. Gironés, A. Gallegos & R. Carbó-Dorca*

Institute of Computational Chemistry, University of Girona, Campus Montilivi, 17071 Girona, Catalonia, Spain

Key words: artemisinin, malaria, MQSM, 1,2,4-trioxanes, peroxy ketals, plasmodium falciparum, quantum similarity, QSAR

Summary

In this work, the antimalarial activity of two series of 20 and 7 synthetic 1,2,4-trioxanes and a set of 20 cyclic peroxy ketals are tested for correlation search by means of Molecular Quantum Similarity Measures (MQSM). QSAR models, dealing with different biological responses (IC_{90} , IC_{50} and ED_{90}) of the parasite *Plasmodium Falciparum*, are constructed using MQSM as molecular descriptors and are satisfactorily correlated. The statistical results of the 20 1,2,4-trioxanes are deeply analyzed to elucidate the relevant structural features in the biological activity, revealing the importance of phenyl substitutions.

Introduction

Approximately 300 million people worldwide are affected by malaria and between 1 and 1.5 million people die from it every year [1, 2]. Previously extremely widespread, the malaria is now mainly confined to Africa, Asia and Latin America. The problems of controlling malaria in these countries are aggravated by inadequate health structures and poor socioeconomic conditions, and the situation is becoming even more complex because over the last few years there has appeared an increasing resistance to the drugs used to combat the *Plasmodium* parasite, which causes the disease. Malaria is treated by an assorted set of therapies that act as inhibitors of synthesis and reduction of dihydrofolase in the parasites [3]. However, deteriorating social and economic conditions, along with inadequate sanitation and health care, lead to self-administration of drugs, often with incomplete treatment, which result in an increase in resistance of the parasites to commonly used drugs. Fortunately, two natural peroxides, artemisinin and yingzhaosu, which possess potent antimalarial activity, have opened a new door in the chemotherapy of malaria [4–6]. Due

to the complex structure of the natural products, structurally simpler 1,2,4-trioxanes have been synthesized and tested for antimalarial activity [7–9]. In this way, the design of new, both cheap and powerful, compounds able to treat malaria are of major importance. QSAR studies represent one of the best methodologies in computer-based drug design, offering valuable information about biological activity and providing a computationally inexpensive methodology to the design of potential bioactive drugs. Thus this work is devoted to this task.

In the present study, the QSAR models for two sets of 1,2,4-trioxanes and a series of cyclic peroxy ketals, which possess antimalarial activity, are built using molecular quantum similarity measures (MQSM) [10–16]. This methodology is based upon quantitative comparative measures between molecular density functions and has been applied successfully to QSAR studies within pharmacological [17–23], including also antimalarial compounds [24], and toxicological [25, 26] environments. Quantum similarity measures provide a suitable quantification of the resemblance between two molecular electron distributions, and the similarity matrices, obtained by collecting the MQSM among a molecular set, can then be manipulated to generate QSAR parameters. The mathematical foun-

*To whom correspondence should be addressed. E-mail: E-mail: director@iqc.udg.es

PDF-OUTPUT

DISK, CP Gr.: 201010336, JCAM 127 (jcamkap:bio2fam) v.1.1
jm401809.tex; 18/02/2002; 11:52; p.1

please correct and return this marked proof to
KLUWER ACADEMIC PUBLISHERS
Manufacturing department

P.O. BOX 990
3300 AZ DORDRECHT - THE NETHERLANDS

dation of MQSM, as well as many up-to-date applications into the QSAR field, is compiled in a recent volume by Carbó-Dorca et al. [27].

This paper is organized as follows: first, the basic methodology concerning MQSM and the statistical treatment of the resulting similarity matrices is exposed. Next, the molecular sets composed by 20 and 7 1,2,4-trioxanes, and 20 cyclic peroxy ketals are presented and the QSAR analysis discussed. Finally, the conclusions of this work are given.

Materials and methods

Molecular quantum similarity measures (MQSM) and Carbó indices

The foundation of MQSM lays on the comparison of quantum mechanical density functions. These density functions are, if chosen as first-order, quantum-mechanical observable elements that produce information on the electron distribution of the molecules. Within this framework, two molecular structures are considered to be similar if their electron distributions are. Thus, among other possibilities [12, 15, 16], quantitative measures of the similarity between two molecules can be defined as the direct volume integral between their density functions:

$$Z_{AB} = \int \rho_A(\mathbf{r}) \rho_B(\mathbf{r}) d\mathbf{r}, \quad (1)$$

or as the volume integral weighted by the Coulomb operator:

$$Z_{AB} = \iint \rho_A(\mathbf{r}_1) |\mathbf{r}_1 - \mathbf{r}_2|^{-1} \rho_B(\mathbf{r}_2) d\mathbf{r}_1 d\mathbf{r}_2, \quad (2)$$

where $\rho_A(\mathbf{r})$ and $\rho_B(\mathbf{r})$ are the electron density functions of molecule A and B, respectively. Z_{AB} is the resulting quantum similarity measure. As MQSM depend on the relative position of both compared objects in space, an alignment procedure is required. When studied molecular sets share common structural features, Topo-Geometrical Superposition Algorithm (TGSA) [28] is used as it performs pairwise superpositions according to the molecular backbones.

TGSA consists of a molecular alignment procedure based on comparisons of atom types and interatomic distances. TGSA finds the largest common molecular substructure to a series of molecules, and orients the molecular superposition towards the matching of such substructure. TGSA program can be obtained upon request [28].

Once superposed, the overall set of MQSM for a series of molecules are computed and collected in *similarity matrix* form, which can be used to extract information and build QSAR descriptors. When the similarity matrix has been computed, a possible scaling of the MQSM can be done by means of the transformation:

$$C_{AB} = Z_{AB}/[Z_{AA}Z_{BB}]^{1/2}, \quad (3)$$

giving the so-called Carbó-index [10]. Carbó indices are comprised within the interval [0, 1]. The closer to one the index is, the more similar to each other the compounds are.

To avoid expensive theoretical calculations, the promolecular atomic shell approximation (ASA) [29–31] has been used here to construct molecular electron density. Within ASA the molecular density is expressed as a sum of discrete atomic density contributions, which are taken as 1S Gaussian functions and fitted to atomic *ab initio* ones. Fitting ASA parameters can be freely downloaded from our web site [32]. Once the overall atomic densities are built, each molecular density function can be constructed by adding as appropriate these elementary building blocks. Since it has been proved that the MQSM from molecular densities built in this way differ by up to a 2% from the *ab initio* ones [30], their use is clearly justified.

Once the similarity matrix has been constructed and, if applicable, Carbó indices calculated, no further calculations are needed before the construction of QSAR models, which obviously requires some computation. This is the main advantage of this method due that only molecular density functions are required, which are straightforwardly obtained using ASA, and the calculation of cross-similarities, which can be easily accomplished. There is no need to obtain or compute further parameters, like log *P*, Hammett sigma or molecular fields. Given a molecular set, MQSM, which are readily obtain with present computational resources, will constitute the basis for the generation of molecular descriptors. In this way, molecular quantum similarity parameters are unbiased and universal molecular descriptors. obtained

Treatment of quantum similarity matrices

Usual chemometric tools are applied here to deal with similarity matrices. Principal Components Analysis (PCA) [33] is an appropriate technique to reduce the dimensionality of the similarity data. The method considers the objects as points in a multidimensional

space and then finds coordinates such that the inter-point distances match as well as possible the original similarities. Taking the Carbo index matrix C as a starting point, a coordinate matrix X can be derived in a new multidimensional space. The coordinates in this space are called the principal components (PCs) of the system and constitute the descriptors used in this work. Each PC is associated with an eigenvalue, indicating the explained variance of the axis. All those PCs accounting for less than 1% variance are a priori neglected, avoiding unnecessary background noise correlations. Furthermore, the optimal variables to describe the desired property are not necessarily those that account for maximal variance, so a simple selection technique is adopted. The most predictive variable method (MPVM) [34] quantifies the importance of the descriptors simply by projecting each PC on the external data.

Quantitative models are built by means of multilinear regression using the selected PCs as independent X -variables. The goodness-of-fit is assessed with the conventional r^2 and standard deviation σ_N coefficients. The model obtained is corroborated by leave-one-out cross validation [35]. The cross-validation technique consists of removing one object from the set and then recalculating the predictive model with the remaining objects. This new model is then used to predict the property value for the extracted element. This process is repeated for all molecules of the set, and then a coefficient of prediction q^2 can be defined from the squared cross-validation residuals (PRESS). This coefficient of prediction measures the robustness of the proposed model, and a $q^2 > 0.5$ is commonly accepted as satisfactory.

A note of caution should be given at this point concerning the cross-validation using PCA. When removing the rows which describe each object to be predicted, little part of the information regarding the removed object remains into the objects left to construct the model; as well as the excluded descriptors keep also part of information of the rest of the set. However, as the evaluated property is not present when generating the PCs, this fact is not relevant, as it would be in other statistical procedures, such as Partial least-squares [36]. So, when predicting new molecules, the PCA matrix should be recomputed and the model, with the evaluated property, recalculated to account for the perturbations induced by this new molecule. If the number of molecules to predict is relatively small compared to the training set, the new models generated are not likely to be substantially modified.

Finally, a randomization test [37] is adopted to detect possible chance correlations. In this statistical technique, the dependent Y -variables are randomly permuted in their positions, and new predictive models are built with the altered vectors. If a real structure-activity exists, the only satisfactory results should be obtained for the correctly ordered Y -variables. Otherwise, even if the obtained model seems to correctly describe the system, it cannot be considered significant, because it is built up with an excess of parameters, able to correlate any data set.

Results and discussion

The major goal in QSAR analysis is to establish a simple mathematical relation between the function, namely biological activity, and a few set of molecular descriptors related to the molecular structures involved. Two molecular sets composed by 20 and 7 1,2,4-trioxanes and one group of 20 cyclic peroxy ketals, presented in Tables 1, 2 and 3 respectively, are tested for correlation search between their antimalarial activity and descriptors derived from MQSM.

Molecular modeling

All molecular structures have been constructed using the PC Spartan software package [38] and optimized with the built-in Sybyl Molecular Mechanics force-field. This choice was made according to previous studies [39], where a comparison between Molecular Mechanics and *ab initio* procedures was carried out, proving the superiority of Molecular Mechanics in these molecular sets, and thus saving computational time. A similar study was carried out in a previous work on antimalarial activity [24], where Molecular Mechanics also provided the optimal conformation.

Activity data

For the set composed by 20 antimalarial compounds, the analyzed properties (Y -variables) consist of the concentration (in ng/ml) of the drug able to inhibit 90% of synthesis and reduction of hydrofolate (IC₉₀) in the parasite. *In vitro* experiments have been performed studying two specimens: the *P. falciparum* W2, from Indochina, and D6 clone, from Sierra Leone. Due to the wide range of values present in reference [39], a logarithmic scaling has been adopted to uniform the activity data. For the smaller set formed by 7 compounds, the biological activity is the same

Table 1. Structures and observed in vitro (log IC₉₀) activities of the set of 20 1,2,4-trioxanes.

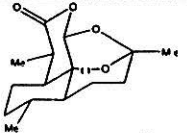
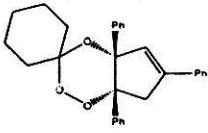
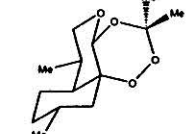
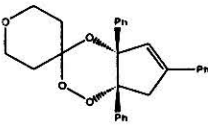
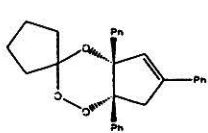
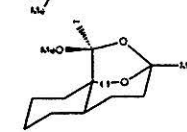
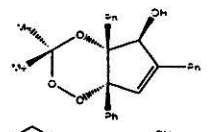
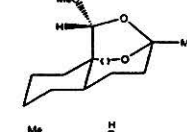
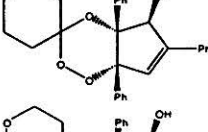
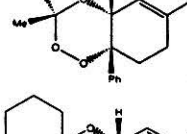
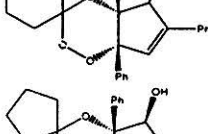
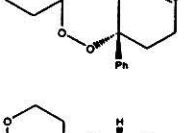
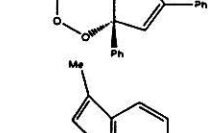
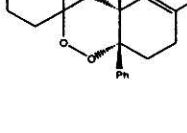
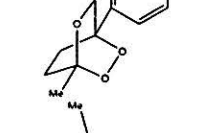
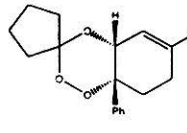
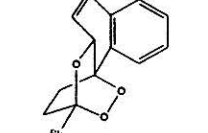
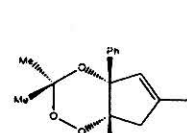
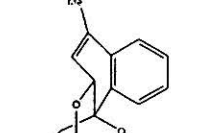
<i>n</i>	Structure	<i>P. falciparum</i> Indo-China W2 log IC ₉₀	<i>P. falciparum</i> Sierra Leone D6 log IC ₉₀	<i>n</i>	Structure	<i>P. falciparum</i> Indo-China W2 log IC ₉₀	<i>P. falciparum</i> Sierra Leone D6 log IC ₉₀
1		0.0414	0.3617	11		0.6990	1.0864
2		1.4116	1.7767	12		-0.3979	1.0969
3		0.5185	1.4829	13		0.0414	0.3617
4		0.5798	1.0607	14		0.9731	1.0086
5		1.2279	2.9420	15		0.3010	0.6990
6		3.0734	3.1004	16		1.0864	1.3054
7		2.8082	3.4125	17		-0.2218	1.0253
8		2.7482	3.1483	18		3.0550	2.8209
9		2.3483	2.3856	19		2.7543	2.6884
10		0.9243	1.2856	20		1.4771	1.5441

Table 2. Structures and observed in vivo (ED₉₀) activities of the set of 7 1,2,4-trioxanes

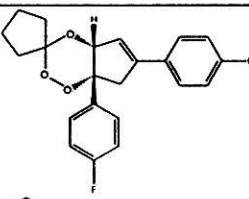
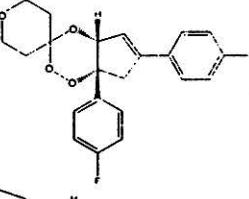
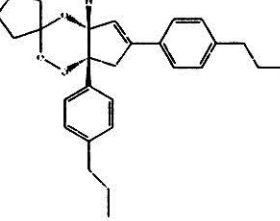
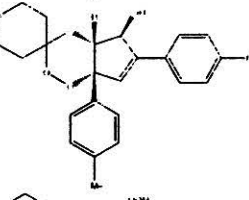
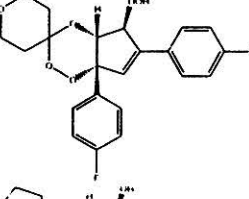
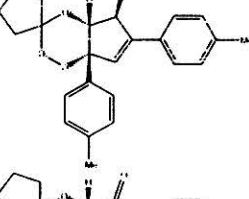
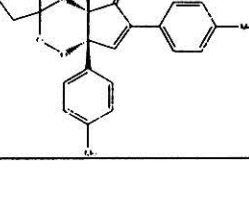
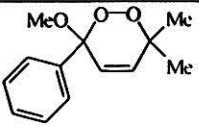
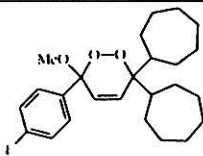
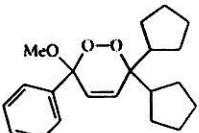
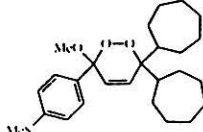
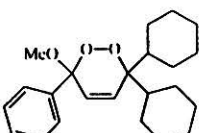
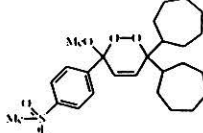
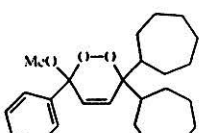
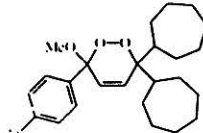
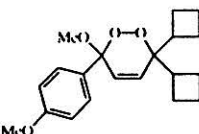
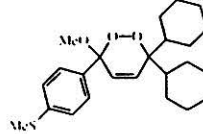
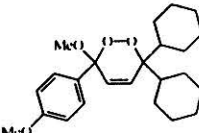
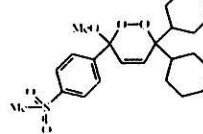
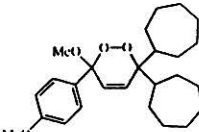
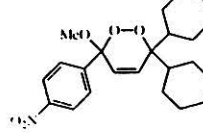
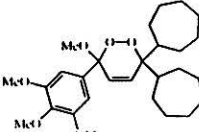
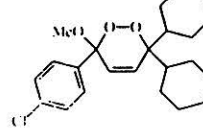
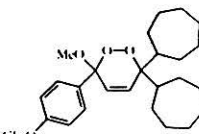
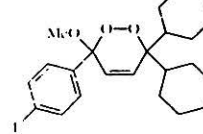
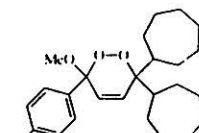
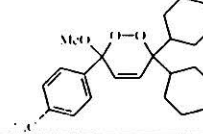
<i>n</i>	Structure	<i>P. beghei</i> ED ₉₀
1		6.8
2		19.5
3		22.5
4		17.0
5		13.2
6		15.0
7		16.0

Table 3. Structures and observed in vitro (log IC₅₀) activities of the set of 20 cyclic peroxy ketals

<i>n</i>	Structure	log IC ₅₀	<i>n</i>	Structure	log IC ₅₀
1		3.041	11		1.929
2		2.279	12		1.892
3		2.447	13		1.491
4		2.342	14		2.255
5		2.204	15		2.204
6		2.255	16		1.748
7		2.322	17		1.663
8		2.079	18		2.000
9		1.785	19		2.301
10		1.763	20		2.146

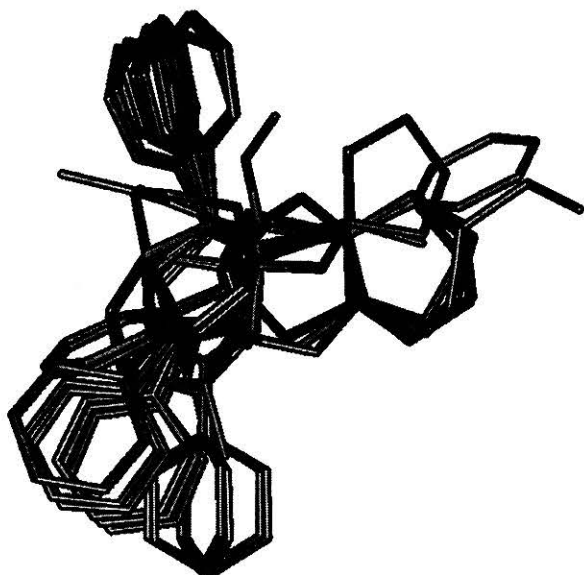


Figure 1. Graphical example of the TGSA alignment solution for artemisinin (molecule 1 from table 1). All molecules have been pairwise superposed to it. (Hydrogen atoms have been suppressed for clarity)

(but in mg/kg). However, it has been measured in the *P. berghei* *in vivo* (ED₉₀). In this case, as the range of values is narrower, no scaling has been performed. The biological property studied in the 20 cyclic peroxy ketals set [40] also consists of inhibition of the metabolism to hydrofolate, but 50% (IC₅₀) in nM concentration units. Similarly to the first set, activity values have been taken in logarithmic scale. All data is presented together with the molecular structures in Tables 1*, 2 and 3.

Molecular alignment

As commented before, MQSM depend on the relative position of the compared molecules in space. In this work, TGSA [28] has been used to perform the pairwise molecular alignments. This method overlays the molecules according the maximal common substructure shared by the analyzed molecules. An illustrative example of the application of TGSA in the 20 antimalarial molecules from the first set is presented in Figure 1.

As seen in Figure 1, the alignment solution provided by TGSA tends to superpose the trioxane groups, and, hence, the rest of the molecular structure is unequivocally overlaid, as it can be seen by

*IC₉₀ value for compound number 1 in Table 1 was not available and IC₅₀ is used instead.

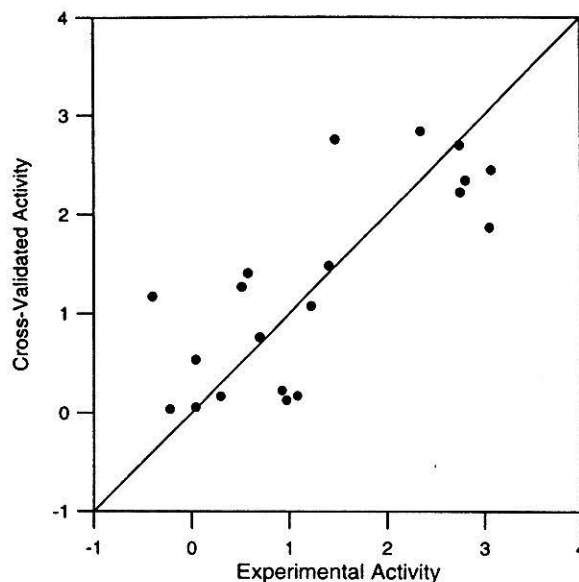


Figure 2. Cross-validated versus experimental antimalarial activity values over *P. falciparum* W2 (1,2,4-Trioxanes).

the phenyl groups, which are located in two clearly differentiated regions.

Quantitative structure-activity relationships

MQSM matrices have been constructed as follows: the sets composed by 1,2,4-Trioxanes have been calculated using the volume integral stated in Equation 1, whereas the Coulomb formulation presented in Equation 2 has been used in the cyclic peroxy ketals. The reason for a different MQSM calculation in these sets, beyond the statistical results, often resides in a different chemical behavior of the involved chemical species. Whereas Overlap operator accounts for simple steric effects, Coulomb describes more electrostatic interactions.

In order to quantify the influence on antimalarial activity of the diverse molecules according to their structure, the quantum similarity matrix deals with multivariate analysis techniques as detailed previously. Simple mathematical manipulations yield to reliable models that relate the molecular descriptors to the antimalarial activity.

A quantitative relationship between the structural descriptors and the property can be established by means of a multilinear regression, as pointed out in the theoretical section. Valuable models are achieved using few descriptors, as presented in Tables 4–7. Optimal values of $r^2 = 0.757$ and $q^2 = 0.589$ for the *P. falciparum* W2, and $r^2 = 0.789$ and $q^2 = 0.662$

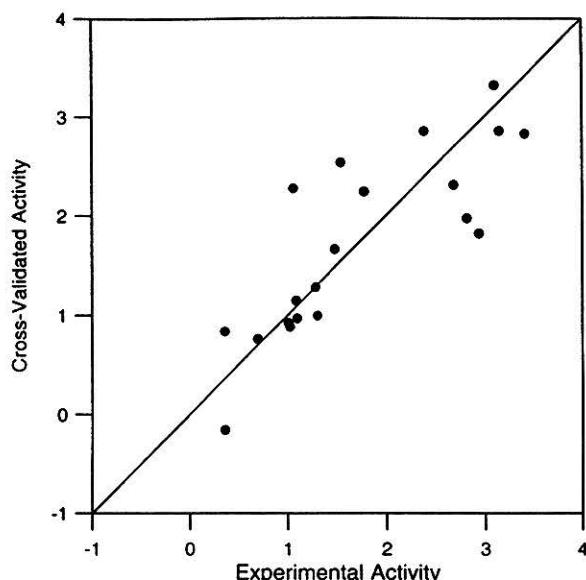


Figure 3. Cross-validated versus experimental antimalarial activity values over *P. falciparum* D6 (1,2,4-Trioxanes).

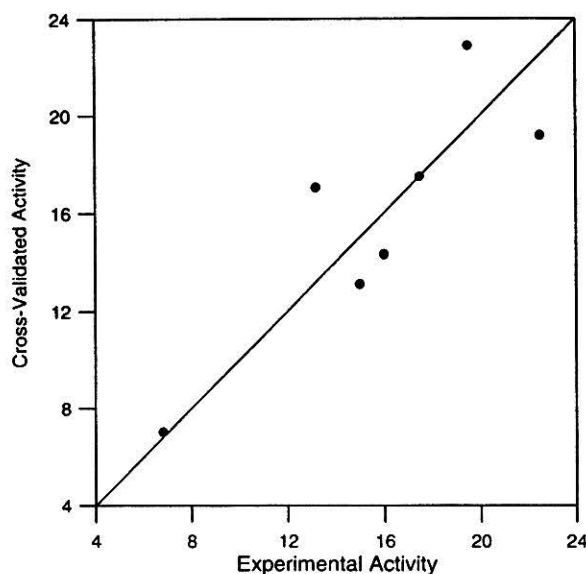


Figure 4. Cross-validated versus experimental antimalarial activity values over *P. berghei* (1,2,4-Trioxanes).

for the *P. falciparum* D6 are obtained when using 4 descriptors for the first molecular set. For the second series, $r^2 = 0.929$ and $q^2 = 0.708$ using three PCs. Finally, in the cyclic peroxy ketals set, $r^2 = 0.778$ and $q^2 = 0.691$ values have been achieved using 4 PCs. As evidenced in the following set of equations, the most descriptive PCs are not necessarily those accounting for maximal variance. The equations of the optimal

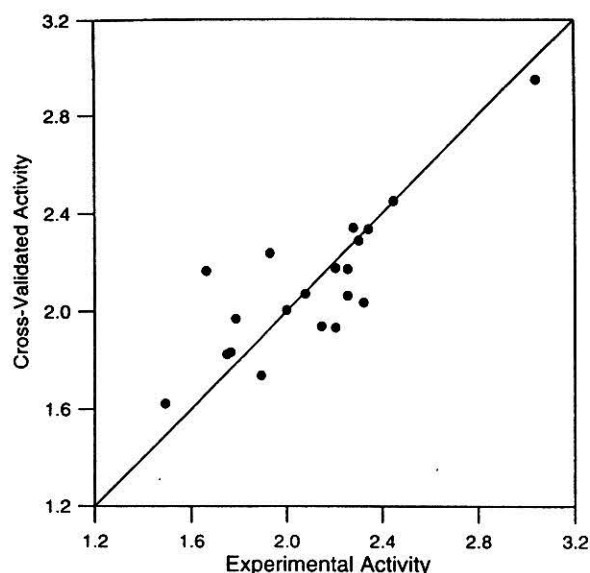


Figure 5. Cross-validated versus experimental antimalarial activity values over *P. falciparum* (cyclic peroxy ketals).

models are:

$$\log IC_{90}^{W2} = 2.912x_1 - 7.166x_2 - 2.083x_3 + 4.948x_7 + 1.272, \quad (4)$$

$$\log IC_{90}^{D6} = 1.806x_1 - 6.102x_2 + 5.966x_7 - 3.420x_{11} + 1.730, \quad (5)$$

$$ED_{90} = -16.53x_4 - 11.67x_5 - 37.52x_6 + 15.79 \quad (6)$$

$$\log IC_{90} = 1.739x_1 - 10.188x_2 - 4.042x_3 - 11.379x_6 + 2.107. \quad (7)$$

50

Figures 2–5 present the plots of the cross-validated activities of the two activities of the first set and the activities of the second and third sets, respectively. As seen, most of the predicted values are in good agreement to the experimental ones with low error.

Finally, the randomization test was performed to assess that the relationships found were not due to an overparameterization of the models or background noise correlations. Thus, in each case, a hundred new models were built in the same optimal conditions (MPVM and 4/3 PCs were applied), but the Y -variables were randomized permutations of the original activities. The results are compiled in the Table 8, where the average values, as well as the maximum values achieved, of r^2 and q^2 are quoted.

As seen, although many random models achieve high r^2 values, the vast majority of them lead to very

Table 4. Statistical parameters for *P. falciparum* Indo-china W2 clone (1,2,4-Trioxanes)

No. PCs	r^2	q^2	σ_N	PCs used
1	0.524	0.446	0.771	2
2	0.631	0.489	0.679	1, 2
3	0.722	0.578	0.589	1, 2, 7
4	0.757	0.589	0.550	1, 2, 3, 7
5	0.797	0.592	0.504	1, 2, 3, 5, 7
6	0.828	0.646	0.462	1, 2, 3, 5, 7, 11

Table 5. Statistical parameters for *P. falciparum* Sierra Leone D6 clone (1,2,4-Trioxanes)

No. PCs	r^2	q^2	σ_N	PCs used
1	0.519	0.423	0.663	2
2	0.575	0.427	0.623	1, 2
3	0.757	0.629	0.472	1, 2, 7
4	0.789	0.662	0.439	1, 2, 7, 11
5	0.795	0.621	0.433	1, 2, 4, 7, 11
6	0.801	0.540	0.426	1, 2, 4, 5, 7, 11

Table 6. Statistical parameters for *P. berghei* (1,2,4-Trioxanes).

No. PCs	r^2	q^2	σ_N	PCs used
1	0.643	0.270	2.76	6
2	0.852	0.428	1.78	4, 6
3	0.929	0.708	1.23	4, 5, 6
4	0.976	0.789	0.721	3, 4, 5, 6

Table 7. Statistical parameters for *P. falciparum* (cyclic peroxy ketals)

No. PCs	r^2	q^2	σ_N	PCs used
1	0.546	0.414	0.225	2
2	0.592	0.418	0.214	1, 2
3	0.738	0.607	0.171	1, 2, 6
4	0.778	0.691	0.158	1, 2, 3, 6
5	0.795	0.656	0.151	1, 2, 3, 4, 6

poor predictive capacity; hence the mean value of q^2 is negative in all cases. However, in a few cases greater q^2 values are obtained, but all of them below the limit of statistical significance ($q^2 > 0.5$), thus conclud-

Table 8. Results of the random test for 100 permutations of the activity vector

Studied set	Property	Mean r^2	Max r^2	Mean q^2	Max q^2
1	$\log IC_{90}^{W2}$	0.308	0.628	-0.267	0.358
	$\log IC_{90}^{D6}$	0.313	0.789	-0.256	0.478
2	ED_{90}	0.477	0.890	-4.746	0.397
3	$\log IC_{90}$	0.215	0.541	-0.435	0.090

ing that the obtained models are free from fortuitous correlations.

QSAR models interpretation

A further study can be performed to the obtained results by analyzing the obtained equations. The first set is ideal to carry out this analysis due that it is composed by assorted different substitutions. This molecular set, as studied, includes two different biological activities, which correlate fairly well ($r^2 = 0.796$), when applying the logarithmic transformation. This fact explains both the resemblance in the statistical results, and the elections and relevance of the optimal PCs, which were chosen in the same order, see Tables 4 and 5, except for the last one.

The subsequent step involves an analysis of the PCs chosen to see how the molecular point clouds are distributed in the lower dimensional space. As example, the first and second PCs are plotted in Figure 6.

As seen in Figure 6, a clear 4-cluster pattern is present, grouping molecules according to clear sub-structural features:

- Molecules 1–5, which present small aliphatic, but no aromatic, substitutions, present high activity.
- Molecules 6–9, which have an aromatic substitution and a phenyl group in the region where the two main ring fuse, present low activity.
- Molecular 10–17, which present phenyl groups at both sides of the fusion region, present high or very high activity.
- Molecular 18–20, which have a benzene ring fused to the main backbone, present low activity.

Thus, it can be seen the influence of these phenyl groups in the biological activity and how it is reflected in this simple 2-dimensional space. From this observation, it can be thought that any novel structure falling in the first, or better in the third, cluster would possess

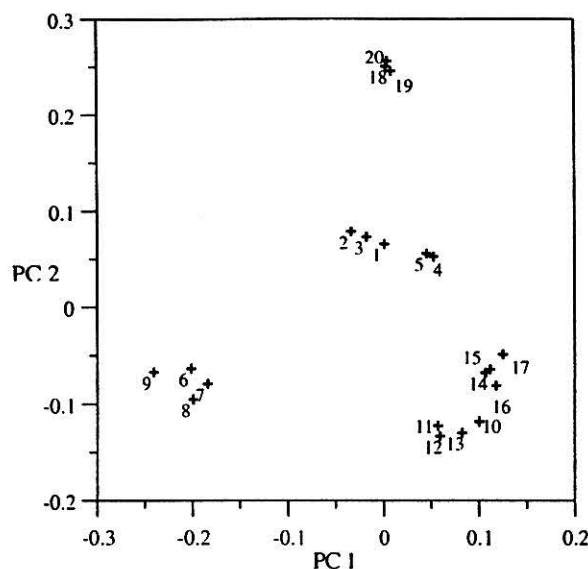


Figure 6. Plot of the first versus the second PC for the molecular set of Table 1.

a high activity against both breeds of *P. Falciparum*. In this way, computational design should be guided to those structures substituted with two phenyl groups at both sides of the ring fusion area and tested with different functional groups.

Conclusions

In the present study, molecular quantum similarity measures have been applied to correlate systematically the antimalarial activity of 1,2,4-trioxanes and cyclic peroxy ketals series. Satisfactory quantitative models have been obtained using little number of descriptors based on Principal Components regressions, achieving also good results in leave-one-out cross validations and random tests.

In addition, a qualitative analysis of the results for the 20 antimalarial set composed by 20 1,2,4-trioxanes has been carried out, revealing structural information about the data set. The molecules were clustered according to common structural features, which in turn explained the biological activity. When two phenyl substitutions are present in the molecule, it seems that the biological activity tends to increase. Hence, a pattern for computer-aided molecular design is given.

Acknowledgements

This research was partially supported by the European Commission Project number ENV4-CT97-0508 and the CICYT contract SAF2000-0223-C03-01. One of us, X.G., benefited from a predoctoral fellowship from the University of Girona. Financial support from the *Fundacio Maria Francisca de Roviralta* is also acknowledged. The authors wish to acknowledge the referee's efforts due to their comments have improved the original manuscript in many ways. Lively discussions with Dr Emili Besalú and Mr Lluís Amat are also acknowledged.

References

1. T.D.R. News (News from the WHO Division of Control of Tropical Diseases), 46 (1994) 5.
2. The Malaria Foundation International website, at <http://www.malaria.org>.
3. Bruce-Chwatt, L.J. (Ed.), *Chemotherapy of Malaria*, 2nd ed., W.H.O., Geneva, 1981.
4. Klayman, D.L., *Science*, 228 (1985) 1049.
5. Butler, A.R. and Wu, Y.L., *Chem. Soc. Rev.*, 21 (1992) 85.
6. Shen, C.C. and Zhuang, L.G., *Med. Res. Rev.*, 4 (1984) 47.
7. Jefford, C.W., Velarde, J.A., Bernadinelli, G., Bray, D.H., Warhurst, D.C. and Milhous, W.K., *Helv. Chim. Acta*, 76 (1993) 2775.
8. Jefford, C.W., Wang, Y. and Bernadinelli, G., *Helv. Chim. Acta*, 71 (1988) 2042.
9. Testa, B., Kyburz, E., Fuhrer, W. and Giger, R. (Eds.) *Perspectives in Medicinal Chemistry*, Vol. 25, VCH Publishers, Amsterdam, 1992, pp. 459-472.
10. Carbó, R., Arnau, J. and Leyda, L., *Int. J. Quantum Chem.*, 17 (1980) 1185.
11. Carbó-Dorca, R. and Besalú, E., *TEOCHEM*, 451 (1998) 11.
12. Carbó-Dorca, R. and Mezey, P.G. (Eds.) *Advances in Molecular Similarity*, Vol. 2, JAI Press, Greenwich, CT, 1998, pp. 1-42.
13. Carbó-Dorca, R. and Mezey, P.G. (Eds.) *Advances in Molecular Similarity*, Vol. 2, JAI Press, Greenwich, CT, 1998, pp. 43-72.
14. Besalú, E., Carbó, R., Mestres, J. and Solà, M., *Top. Curr. Chem.*, 173 (1995) 31.
15. Carbó, R. (Ed.) *Molecular Similarity and Reactivity: From Quantum Chemical to Phenomenological Approaches*, Kluwer, Amsterdam, 1995, pp. 3-30.
16. Carbó-Dorca, R. and Mezey, P.G. (Eds.) *Advances in Molecular Similarity*, Vol. 1, JAI Press, Greenwich, CT, 1996, pp. 1-42.
17. Carbó, R., Besalú, E., Amat, L. and Fradera, X., *J. Math. Chem.*, 18 (1995) 237.
18. Fradera, X., Amat, L., Besalú, E. and Carbó-Dorca, R., *Quant. Struct.-Act. Relat.*, 16 (1997) 25.
19. Lobato, M., Amat, L. and Carbó-Dorca, R., *Quant. Struct.-Act. Relat.*, 16 (1997) 465.
20. Amat, L., Robert, D., Besalú, E. and Carbó-Dorca, R., *J. Chem. Inf. Comput. Sci.*, 38 (1998) 624.

21. Robert, D., Amat, L. and Carbó-Dorca, R., *J. Chem. Inf. Comput. Sci.*, 39 (1999) 333.
22. Robert, D., Gironés, X. and Carbó-Dorca, R., *J. Comput. Aid. Mol. Des.*, 13 (1999) 597.
23. Gironés, X., Amat, L. and Robert, D. and Carbó-Dorca, R. *J. Comput. Aid. Mol. Des.*, 14 (2000) 477.
24. Gironés, X., Gallegos, A. and Carbó-Dorca, R. *J. Chem. Inf. Comput. Sci.*, 40 (2000) 1400.
25. Robert, D. and Carbó-Dorca, R., *SAR QSAR Environ. Res.*, 10 (1999) 401.
26. Gironés, X., Amat, L. and Carbó-Dorca, R., *SAR QSAR Environ. Res.*, 10 (1999) 545.
27. Carbó-Dorca, R., Robert, D., Amat, L., Gironés, X. And Besalú, E. (Ed.), *Molecular Quantum Similarity and Drug Design*, Springer, Lecture Notes in Chemistry 73, Berlin, 2000.
28. Gironés, X., Robert, D. and Carbó-Dorca, R., *J. Comp. Chem.*, 22 (2001) 255. An operative version can be obtained upon request at quantum@iqc.udg.es.
29. Constants, P. and Carbó, R., *J. Chem. Inf. Comput. Sci.*, 35 (1995) 1046.
30. Amat, L. and Carbó-Dorca, R., *J. Comput. Chem.*, 18 (1997) 2023.
31. Amat, L. and Carbó-Dorca, R., *J. Comput. Chem.*, 20 (1999) 911.
32. Fitting ASA parameters form an assorted collection of basis sets can be downloaded from our web site: <http://iqc.udg.es/cat/similarity/ASA/basisset.html>.
33. Cox, T.F. and Cox, M.A.A. (Eds), *Multidimensional Scaling*, Chapman & Hall, London, 1994.
34. Cuadras, C.M. and Arenas, C.A., *Commun. Stat. Theor. Method.*, 19 (1990) 2261.
35. Allen, D.M., *Technometrics*, 16 (1974) 125.
36. See for example this detailed tutorial by Geladi, P. and Kowalsky, B.R. *Anal. Chim. Acta*, 185 (1986) 1.
37. Van der Waterbeemd, H. (Ed.), *Chemometric Methods in Molecular Design*, VCH, New York, 1995, pp. 309–318.
38. PC Spartan Pro, version 1.4, 1997, Wavefunction Inc, Irvine.
39. Grigorov, M., Weber, J., Tronchet, J.M.J., Jefford, C.W., Milhous, W.K. and Maric, D., *J. Chem. Inf. Comput. Sci.*, 37 (1997) 124.
40. Posner, G.H., O'Dowd, H., Ploypradith, P., Cumming J.N., Xie, S. and Shapiro, T.A., *J. Med. Chem.*, 41 (1998) 2164.

Modeling Antimalarial Activity: Application of Kinetic Energy Density Quantum Similarity Measures as Descriptors in QSAR

Xavier Gironés, Ana Gallegos, and Ramon Carbó-Dorca*

Institute of Computational Chemistry, University of Girona, Campus Montilivi,
17071 Girona, Catalonia, Spain

Received May 30, 2000

In this work, is studied the application, within a quantum similarity framework, of the recently described Kinetic Energy Density Function in the evaluation of the antimalarial activity. First, this new type of Density Function is briefly presented from its theoretical foundations, and its inclusion in the molecular quantum similarity is discussed afterward. The application of Kinetic Energy-based Quantum Similarity Measures to QSAR is tested with 2 molecular sets composed of artemisinin derivatives, in which the 50% inhibition of synthesis and reduction of hidrofolate (IC₅₀) in different *Plasmodium falciparum* clones are analyzed. Satisfactory correlations are obtained for all antimalarial activities in all studied molecular sets. Molecular Quantum Similarity analysis provides a consistent, unbiased, and homogeneous set of molecular descriptors and is a feasible alternative to the use of classical physicochemical descriptors.

INTRODUCTION

Quantitative Structure–Activity Relationships (QSAR) find their origin in the XIX century when, in 1863, the toxicity of different alcohols in mammals was observed to increase with their decreasing solubility in water.¹ Since then, several attempts to relate the molecular structure to biological responses^{2–4} or to physicochemical features^{5–7} were postulated. However, present day formulation of QSAR is due to the work of Hansch and Fujita⁸ in 1964. The basis of this method, and consequently of all later inspired ones, consists of assuming that the biological activity is the result of the contribution of many factors behaving independently. A structural descriptor characterizes each contribution, and the biological activity of a set of compounds is adjusted to a multilinear model. The most used descriptors in this approach are those related to hydrophobicity (log P), Hammett's constant (σ), and lipophilicity (π). Other molecular features, also taken into account in this scope when deriving QSARs, are partial charges, orbital energies,⁹ polar interactions,¹⁰ electrostatic potentials,¹¹ and steric parameters.¹² Parallel to these techniques and avoiding the use of physicochemical parameters, molecular topology was developed. This theory considers the biological activity related to the molecular topological features, which are numerically represented through connectivity and distance indices. Even though the seminal works in this field are also dated in the XIX century,¹³ modern formulation is due to Wiener¹⁴ and later continued by Kier and Hall¹⁵ and Randic.¹⁶ Other contributions also involve Hosoya, Balaban and Platt.¹⁷ Several topological indices have also been successfully applied to QSAR.^{18,19}

A new trend in QSAR studies appeared in 1988 due to the introduction of three-dimensional parameters, which could account for the influence of conformation and stereochemistry. These new techniques, commonly known as 3D-QSAR, use to involve extensive calculations in the space

surrounding the molecules and thus, alignment of the molecules according to a common pharmacophore is required. The first model of this kind was the COMFA analysis²⁰ and it is one of the most used nowadays. Since then, several 3D-QSAR techniques have been presented^{21–26} including those based on similarity concepts.^{27–41}

Among these previously quoted methodologies dealing with similarity in QSAR studies, Molecular Quantum Similarity Measures (MQSM)^{42–48} stand as an efficient tool to solve actual chemical problems. MQSM methodology has been successfully applied within pharmacological^{49–54} and toxicological^{55–57} problems. MQSM are based on the concept of molecular similarity^{58,59} and are related to a self-evident molecular similarity principle: “the most similar two molecules are, the more similar properties will possess”. This last statement requires a procedure to compare the molecules, and MQSM easily establish the degree of similarity based on the electronic distribution, namely density functions (DF), of the molecular structures.

Traditionally, MQSM have been computed using electronic DF. However, nothing opposes the use of the recently described extended DF^{60–62} into the general formulation of MQSM. Among the assorted set of different extended DF, Kinetic Energy (KE) DF may constitute a suitable basis for the generation of molecular descriptors. In this work, it is intended to use Kinetic Energy-based MQSM to correlate the antimalarial activity of two series of compounds.

Malaria is an infectious disease endemic in many parts of the world.⁶³ Mostly located in tropical and subtropical areas, it is estimated that now between 300 and 500 million people have malaria. Therapies in these regions are being made difficult by the continuing spread of drug-resistant *Plasmodium falciparum* clones, due to the increasing immunity against traditional medications such as chloroquine,⁶⁴ and thus some reports indicate that this problem could perhaps be controlled by periodic introduction of new antimalarial drugs.^{65,66} In present times, a new class of nonalkaloidal

* Corresponding author phone: +34 972418367; fax: +34 972 418356; e-mail: director@iqc.udg.es.

trioxane compounds is becoming widespread: artemisinin and the semisynthetic ether and ester derivatives are reaching the general affected population. Artemisinin, a natural occurring peroxidic cadinane sesquiterpene possessing good antimalarial activity^{67,68} has been the subject of many total synthesis⁶⁹ and structure–activity relationships studies;^{70–73} however, up to date few QSAR studies involving antimalarial compounds have appeared in the literature (see for example ref 74).

As stated before, much of this intense interest stems from the need for new antimalarial drugs with unconventional structures and novel modes of action to be used for the treatment of pervasive strains of drug-resistant *P. falciparum*. QSAR studies represent one of the best methodologies in computer-based drug design, offering valuable information about biological activity and providing a computationally inexpensive methodology to the design of potential bioactive drugs. Thus, this paper is devoted to this task, being consequently structured as follows: first, the KE DF is defined. Next, its inclusion in the MQSM framework is discussed, and afterward the statistical treatment of the resulting similarity matrices is exposed. As computational examples, the molecular sets composed by 18 and 15 artemisinin derivatives are presented and the corresponding QSAR analysis performed. Finally, the conclusions of this work are given.

MATERIAL AND METHODS

Kinetic Energy MQSM and Carbó Indices. KE DF arise from the mathematical problem of the inclusion of the second-order differential operator ∇^2 , accounting for the expectation value of KE, within quantum mechanical theory. As quantum mechanical KE expectation values do not fulfill the usual expectation value statistical formalism, which other nondifferential operators do, the following expression (1) might be used in order to obtain a statistically consistent probability distribution

$$2\langle K \rangle = -\int \Psi^* \nabla^2 \Psi d\mathbf{r} = \int (\nabla \Psi)^* (\nabla \Psi) d\mathbf{r} = \int \kappa(\mathbf{r}) d\mathbf{r} \quad (1)$$

where $\langle K \rangle$ is the KE expectation value and Ψ any state wave function. The corresponding expression for the KE DF, distribution within Molecular Orbital (MO), theory can be written as

$$\kappa(\mathbf{r}) = \sum_i w_i |\nabla \varphi_i(\mathbf{r})|^2 \quad (2)$$

where $\{\varphi_i\}$ are the system's MOs and $\{w_i\}$ are the corresponding occupation numbers. This last equation provides the feasible inclusion of KE DF in the MQSM framework, resulting in the KE MQSM, between two Quantum Objects A and B

$$Z_{AB} = \int \kappa_A(\mathbf{r}) \kappa_B(\mathbf{r}) d\mathbf{r} \quad (3)$$

where Z_{AB} is the resulting KE MQSM and the *Overlap*^{43,44,46,47} operator has been assumed. Z_{AB} provides the quantitative measure of resemblance between two molecules based on their KE DF, however, as Z_{AB} is an absolute magnitude, a transformation by means of the Carbó Index⁴²

is applied in order to obtain a relative value of similarity between the interval [0,1]. That is the more similar both molecules are, the closer to 1 the value of the Carbó Index will be and vice versa. In this way, all measures computed as in eq 3 may be transformed into Carbó Index⁴² using the following expression:

$$C_{AB} = Z_{AB} (Z_{AA} Z_{BB})^{-1/2} \quad (4)$$

and, in the general case of QSAR studies, a set composed of n molecules can be submitted to this procedure.^{48,49} The resulting values of the Carbó Index for every molecular pair can be collected in matrix form, whose columns could be thereafter used as molecular descriptors.

Promolecular Atomic Shell Approximation (ASA). To avoid expensive theoretical computations, a promolecular ASA^{75–78} is used here in order to construct the electronic DF. This approach views molecular density as a sum of discrete atomic density contributions, which are taken as linear combinations of 1S Gaussian functions, previously fitted to atomic ab initio DF. ASA DF are mathematically expressed as

$$\rho_A^{\text{ASA}}(\mathbf{r}) = \frac{1}{N_e} \sum_{a \in A} Z_a \rho_a^{\text{ASA}}(\mathbf{r}) \wedge \rho_a^{\text{ASA}}(\mathbf{r}) = \sum_{i \in a} w_i |S_i(\mathbf{r} - \mathbf{R}_i)|^2 \quad (5)$$

where N_e is the number of electrons of molecule A, Z_a is the atomic number of atom a , w_i are the convex fitted coefficients of the linear expansion,^{76–78} and S_i are the 1S Gaussian functions, which are expressed in turn as

$$S_i(\mathbf{r} - \mathbf{R}_i) = N(\alpha_i) \exp(-\alpha_i |\mathbf{r} - \mathbf{R}_i|^2) \quad (6)$$

Once the overall atomic densities are built, each molecular density function can be constructed by adding as appropriate these elementary pieces together within a given geometry, obtained from experimental determinations or previous theoretical calculations. Since it has been proved that regular MQSM from molecular densities built in this way differ by less than 2% from ab initio densities,^{76,78} their use is clearly justified.

Starting from the promolecular ASA DF framework, the gradient vector appearing in eq 2 can be written as

$$\nabla \varphi_i(\mathbf{r}) = \nabla (N(\alpha_i) \exp(-\alpha_i |\mathbf{r} - \mathbf{R}_i|^2)) = -2\alpha_i |\mathbf{r} - \mathbf{R}_i| S_i(\mathbf{r} - \mathbf{R}_i) \quad (7)$$

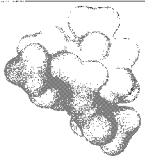
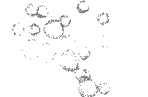
so a Promolecular ASA KE DF can be constructed as

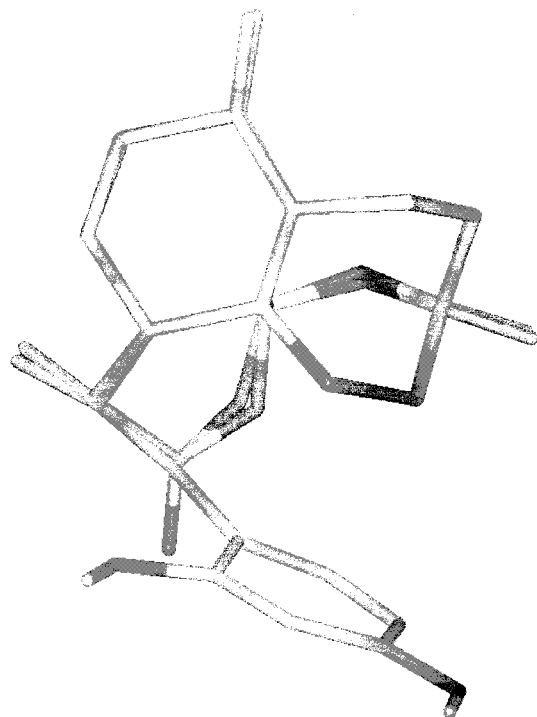
$$\kappa(\mathbf{r}) = 4 \sum_i w_i \alpha_i^2 |\mathbf{r} - \mathbf{R}_i|^2 |S_i(\mathbf{r} - \mathbf{R}_i)|^2 \quad (8)$$

To make a clearer picture of the behavior of KE DF, a comparative analysis of both electronic and KE DF is presented in Table 1.

As seen in Table 1, at low and high DF values both representations behave almost in the same way, reflecting the molecular shape and the atomic locations, respectively. However, at intermediate values some relevant differences become visible, like an oversize of the heavier atoms and the appearance of interatomic maxima. These features

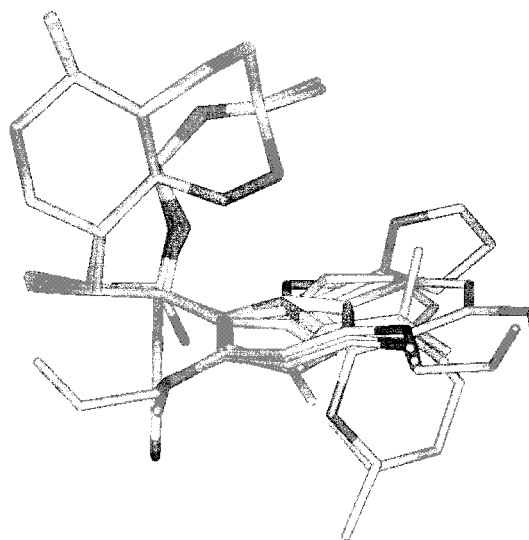
Table 1. Graphical Comparative Analysis of Electronic and KE DF over the Artemisinin Molecule

Density value	Electronic DF	KE DF
$1.25 \cdot 10^{-4}$		
$2.50 \cdot 10^{-4}$		
$7.05 \cdot 10^{-4}$		
$1.80 \cdot 10^{-3}$		

**Figure 1.** Pairwise molecular alignment between artemisinin and a derivative (compound number 2).

preclude a different description of the electronic distribution based on KE concepts.

Molecular Alignment. MQSM are dependent on the relative position of the molecules under comparison. Although there have been described superposition procedures based on MQSM, in the present work, the simpler Topo-Geometrical Approach (TGSA) superposition⁷⁹ has been used to perform the pairwise alignments. This molecular superposition method overlays the molecules according to the maximal common substructure shared by the analyzed compounds. As illustrative examples of TGSA, a pairwise superposition between artemisinin and a derivative (second com-

**Figure 2.** Molecular alignment solution for a set of 17 synthetic 1,2,4-trioxanes over artemisinin.**Table 2.** Upper Triangle of the Carbó Index Matrix for Artemisinin Used To Compare the Different Computational Optimization Methodologies

Sybyl	AM1	HF/3-21 g*
1.000	0.978	0.985
	1.000	0.980
		1.000

pound of the first set) as well as the global superposition based on artemisinin acting as a template are presented in Figures 1 and 2, respectively.

As observed in Figure 1, both structures are aligned around the peroxy bridge present in artemisinin and the rest of the molecular structures belonging to the set. This fact is again evidenced observing Figure 2, where the entire molecular set is overlaid according to this substructure. Hence, the molecular skeleton has been detected and superposed.

Molecular Modeling. All molecular structures have been constructed using the PC Spartan software package⁸⁰ and optimized with the built-in Sybyl Molecular Mechanics force field. This choice was made after performing a comparative analysis between different optimization methodologies over the artemisinin molecule. The comparison was carried out using the following methodologies: Sybyl Molecular Mechanics force field and AM1, both included in PC Spartan,⁸⁰ and using a direct ab initio method, which in this case corresponds to a restricted Hartree–Fock with the 3-21G* basis set, implemented in Gaussian 94.⁸¹ From the optimized molecular coordinates of artemisinin, the electronic DF has been constructed within the previously discussed promolecular ASA, and MQSM involving the Coulomb operator⁴⁴ have been carried out. The choice of Coulomb operator is based on a previous work⁸² where a comparative analysis was also performed to choose the best geometry optimization method. The quantum similarity results over the artemisinin molecule are given in terms of Carbó Indices, resulting from the comparison of the different geometries of artemisinin and are presented in Table 2.

From the results shown in Table 2, it is evidenced that in this case all methods lead to very close structures. However, there is a very important difference in computational time

required to complete the calculations: whereas the *Sybyl* process was completed in a few seconds, AM1 and ab initio methods lasted some minutes and several hours, respectively. In this way, the *Sybyl* methodology was chosen to optimize the geometry due to its accuracy and efficiency regarding these molecular sets.

Treatment of Quantum Similarity Matrices and Model Building. Common chemometric tools are applied to deal with the resulting quantum similarity matrices. Principal Component Analysis (PCA)⁸³ is an appropriate technique to transform (dis-)similarity data. Taking the Carbó Index matrix, C ($n \times n$) being n the number of molecules in the studied set, as a starting point, PCA allows deriving a new $n \times n$ coordinate matrix in a new multidimensional space. These coordinates are called principal components (PCs) of the system and constitute the molecular descriptors used in this work.

Quantitative models are built up by means of multilinear regression using these PCs as independent X -variables. If a maximum number of descriptors, k , when constructing a QSAR model has to be chosen, a Nested Do Loop (NDL)^{84–87} of dimension k over the n PC set is carried out. Each generated model is tested by means of its predictive capacity, evaluated by the squared leave-one-out cross-validation correlation coefficient (q^2), expressed as in eq 9

$$q^2 = \frac{SD - PRESS}{PRESS} \quad (9)$$

where SD is the quadratic deviation of the experimental values compared to their mean (expressed as $SD = \sum_i^n (y_i - \bar{y})^2$)

and $PRESS$ stands for the predictive residual sum of squares (formulated as $PRESS = \sum_i^n (y_i - \hat{y})^2$). The chosen model,

for k descriptors, is the one that maximizes the value of the q^2 parameter. The models obtained are evaluated by goodness-of-fit (r^2), standard deviation (σ_N) coefficients, and the previously mentioned q^2 , derived from the leave-one-out cross-validation procedure.⁸⁸ Cross-validation consists of removing one molecule from the set, recalculating the model with the remaining elements, and predicting the value of the removed compound. This procedure is repeated for each member of the set and allows testing the predictive capacity of the proposed models by means of the squared cross-validated residuals ($PRESS$). A value of $q^2 > 0.5$ is generally accepted as satisfactory.⁸⁹

Finally, the randomization test⁹⁰ is adopted to detect possible chance correlations. In this statistical technique, the dependent Y -variables are randomly permuted in their positions, and new predictive models are built with the altered vectors. If a real structure–activity exists, the only satisfactory results should be obtained for the correctly ordered Y -variables. Otherwise, even if the obtained model seems to correctly describe the system, it cannot be considered statistically significant because it is built up from an excess of parameters, able to correlate any data set, and thus it should be disavowed.

RESULTS AND DISCUSSION

Activity Data. For the set composed of 18 antimalarial compounds, the analyzed property is the nanomolar concen-

Table 3. Molecular Structures and Biological Activities of a Set of 18 Artemisinin Derivatives

n°	Structure	IC ₅₀	n°	Structure	IC ₅₀
1		9.908	10		5.105
2		4.198	11		4.609
3		6.607	12		15.996
4		7.798	13		9.397
5		8.995	14		9.099
6		1.400	15		4.000
7		5.200	16		10.990
8		8.590	17		8.299
9		10.000	18		8.395

tration of the drug able to inhibit 50% synthesis and reduction of hydrofolate in the NF54 strain (chloroquine sensitive) of *P. falciparum* *in vitro* (IC₅₀),⁹¹ as reported in a previous work.⁹² The molecular structures, as well as the activities,

Table 4. Molecular Structures and Biological Activities of a Set of 15 Artemisinin Derivatives

Molecular Structure			log _e IC ₅₀	
			Relative activity	
n	R ₁	R	D6	W2
1	H	CH ₃	2.000	2.000
2	CH ₃	H	1.944	2.049
3	CH ₃ CH ₂	H	3.323	2.828
4	CH ₃ (CH ₂) ₂	H	1.301	1.255
5	(CH ₃) ₂ CH	H	1.724	1.653
6	EtO ₂ CCH ₂	H	2.365	2.365
7	C ₆ H ₅ CH ₂	H	0.477	0.000
8	p-ClC ₆ H ₄ (CH ₂) ₂	H	2.057	2.104
9	C ₆ H ₄ (CH ₂) ₃	H	2.342	2.449
10	CH ₃	CH ₃ (CH ₂) ₃	2.265	2.410
11	CH ₃ (CH ₂) ₂	CH ₃ (CH ₂) ₃	1.447	1.519
12	C ₆ H ₅ CH ₂	CH ₃ (CH ₂) ₃	0.000	0.000
13	p-ClC ₆ H ₄ (CH ₂) ₂	CH ₃ (CH ₂) ₃	1.633	1.724
14	C ₆ H ₄ (CH ₂) ₃	CH ₃ (CH ₂) ₃	1.591	1.681
15	EtO ₂ CCH ₂	CH ₃ (CH ₂) ₃	3.141	3.359

are presented in Table 3. For the second series, the studied property is also IC₅₀; however, the activity has been measured relative to the artemisinin value when tested in vitro in human blood over *P. falciparum* Indochina (W2) and Sierra Leone (D6) clones,⁹³ following a procedure proposed by Desjardins.^{94,95} These clones present an interesting resistance to drugs: whereas the W2 strain is chloroquine-resistant and mefloquine-sensitive, the D6 breed is sensitive to chloroquine and resists mefloquine. This fact allows evaluating the effect and mechanism of action of a novel drug in this widely mutated parasite. The relative potency of these compounds has been adjusted according to IC₅₀ values and then multiplied by the ratio between molecular weight of the analogue and molecular weight of artemisinin. These activities and the corresponding molecular structures are compiled in Table 4. A logarithmic scaling has been applied to the last set of activities due to the wide range of values present in ref 93.

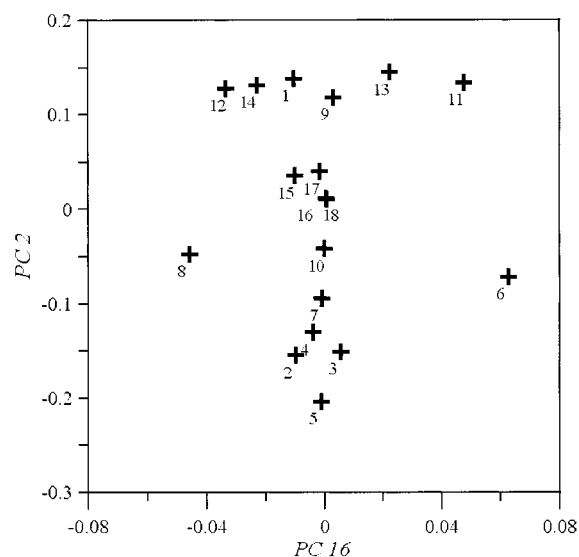
Quantitative Structure–Activity Relationships. MQSM matrices have been constructed using the formula present in eq 3 and scaled using the Carbó Index notation present in eq 4. To quantify the influence on antimalarial activity of the diverse molecules according to their structure, the quantum similarity matrix deals with multivariate analysis techniques as detailed previously. These simple mathematical manipulations yield to reliable models that relate the molecular descriptors to the antimalarial activity.

A quantitative relationship between the structural descriptors and the property is established by means of a multilinear regression, as pointed out in the theoretical section. Valuable models are achieved using few descriptors, as presented in Table 5. Optimal values of $r^2 = 0.754$ and $q^2 = 0.561$ for the NF54 strain are obtained when using four descriptors. As an illustrative example, Figure 3 presents how the first two most predictive PCs spread the molecules in the 2D space.

As seen in Figure 3, both PC 16 and 2 behave differently. PC 16 scatters the molecules according to the weight of the

Table 5. Statistical Parameters for the Proposed QSAR Models

<i>P. falciparum</i> strain	no. of PCs	PCs used	% variance explained	r^2	q^2	σ_N
NF54	4	2,4,9,16	27.775	0.754	0.561	2.140
D6	4	4,6,14,15	19.503	0.767	0.520	0.580
W2	4	1,4,14,15	37.693	0.821	0.576	0.575

**Figure 3.** 2D distribution of the antimalarial compounds according to the first two chosen PCs.

substitution, roughly collapsing the heavy substitutions nearby the origin and distributing the light ones along the sides. PC 2 mostly acts as a substituent discriminator. In the negative part of PC 2, most of the compounds presenting oxygen are present (2, 3, 4, 5, 6, 7, 8), those compounds having an aluminum atom are approximately located in the middle of the axis (16, 17, 18), and finally the nitrogenated substitutions are present in the positive part of PC 2 (11, 12, 13, 14). Exceptions to this rule are compound 1, which only contains a ketonic substitution, compound 9, which contains oxygen but is located with the nitrogen group, and compound 15, which contains a nitrogen and it is located too close to the aluminum containing substitutions. Compound 10 contains a sulfur and is located between the oxygen containing rings and the heavier substitutions with aluminum.

Both biological activities of the second series are also correctly predicted using four PCs, achieving values of $r^2 = 0.767$ and $q^2 = 0.520$ for the D6 clone and $r^2 = 0.821$ and $q^2 = 0.576$ for the W2 strain. The equations of the optimal models are presented in eqs 10–12 and the statistical parameters in Table 5.

$$\text{IC}_{50}^{\text{NF54}} = 11.956\mathbf{x}_2 - 12.179\mathbf{x}_4 - 16.280\mathbf{x}_9 - 74.446\mathbf{x}_{16} + 7.700 \quad (10)$$

$$\log \text{IC}_{50}^{\text{D6}} = 1.900\mathbf{x}_4 + 4.620\mathbf{x}_6 - 32.563\mathbf{x}_{14} + 5.965\mathbf{x}_{15} + 1.826 \quad (11)$$

$$\log \text{IC}_{50}^{\text{W2}} = 4.276\mathbf{x}_1 + 2.507\mathbf{x}_4 - 30.241\mathbf{x}_{14} + 5.822\mathbf{x}_{15} + 1.841 \quad (12)$$

As seen in Table 5, not always the PCs accounting for the maximal variance are those related to the activity. In the

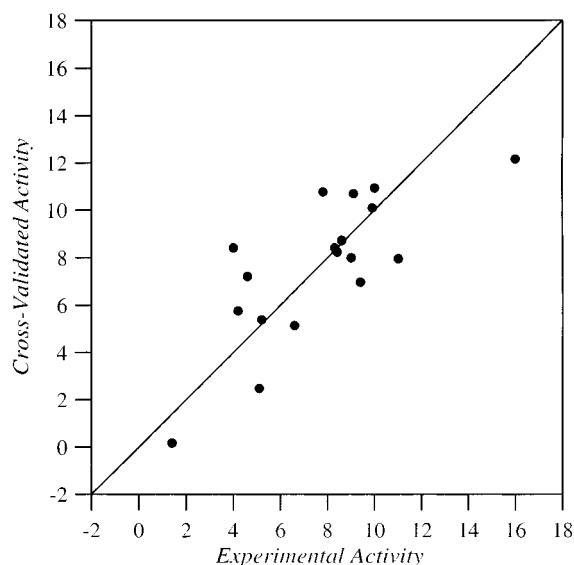


Figure 4. Cross-validated versus experimental antimalarial activity values over *P. falciparum* NF54.

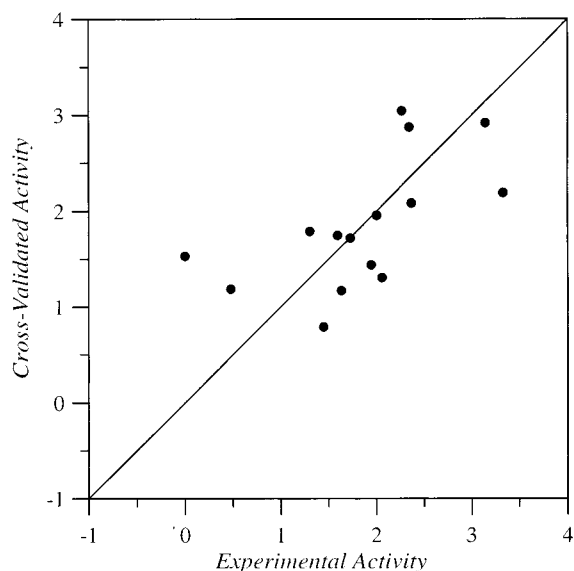


Figure 5. Cross-validated versus experimental antimalarial activity values over *P. falciparum* D6.

studied molecular sets, even the chosen PCs explain less than 40% of the whole variance, and they are able to provide an acceptable description.

Figures 4 to 6 present the plots of the cross-validated activities for the three models consecutively. As observed, most of the predicted values are in good agreement with the experimental ones with acceptable error.

Finally, the proposed models were subjected to a random test procedure to assess that they are not built up from an excess of parameters. Thus, in each molecular set, a hundred activity vectors were generated from randomized permutations of the original ordered one. All proposed models present a clear separation between the original solution and the randomized ones, which clearly achieve values of q^2 below 0.5 or furthermore being negative. As an example, Figure 7 presents the compilation of the results for the *P. falciparum* NF54 set. Similar results are obtained for the remaining studied systems.

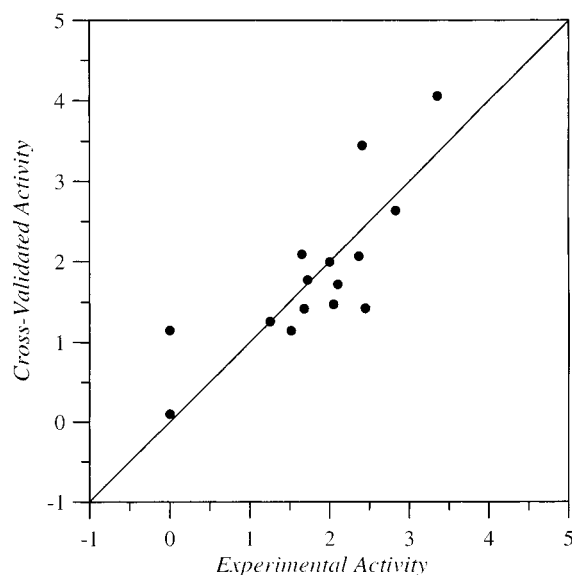


Figure 6. Cross-validated versus experimental antimalarial activity values over *P. falciparum* W2.

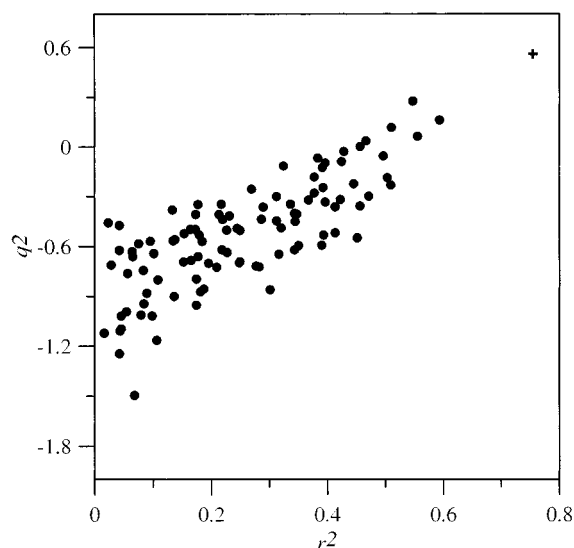


Figure 7. Randomization test for the optimal *P. falciparum* NF54 model. The randomized responses have been marked with dots, and the correctly ordered activity with a cross.

CONCLUSIONS

In the present study, Kinetic Energy based molecular quantum similarity measures have been applied to correlate systematically the antimalarial activity of various artemisinin derivatives. Quantitative structure–activity relationships provide valuable information about the biological behavior of potential drugs and might be of help in the development of new therapies in the fight against malaria. In addition, this approach carries low computational costs, which constitutes another interesting advantage to medical research.

ACKNOWLEDGMENT

This research has been partially supported by the European Commission Project number ENV4-CT97-0508. One of us (X. Gironés) benefits from a predoctoral fellowship from the University of Girona. Financial support from the *Fundació Maria Francisca de Roviralt* is also acknowledged. Some

comments by the evaluators have been helpful to ameliorate the structure of this paper, hence the authors wish to express their gratitude to them.

REFERENCES AND NOTES

- Quoted in: Borman, S. New QSAR Techniques Eyed for Environmental Assessments. *Chem. Eng. News* **1990**, 68, 20–23.
- Crum-Brown, A.; Fraser, T. R. On the connection between chemical constitution and physiological action. Part 1. On the physiological action of the ammonium bases, derived from Strychia, Brucia, Thebaia, Codeia, Morphia and Nicotia. *Trans. Royal Soc. Edinburgh* **1868–1869**, 25, 151–203.
- Meyer, H. Theorie der Alkoholnarkose; welche Eigenschaft die Anästhetica bedingt ihre narkotische Wirkung. *Arch. Exp. Pathol. Pharmacol.* **1899**, 42, 109–118.
- Ferguson, J. The use of chemical potentials as indicators of toxicity. *Proc. Royal Soc. London* **1939**, B127, 387.
- Hammett, L. P. The Effect of Structures upon the Reactions of Organic Compounds. Benzene Derivatives. *J. Am. Chem. Soc.* **1937**, 59, 96–103.
- Swain, C. G.; Lupton, E. C., Jr. Field and Resonance Effects Components of Substituent Effects. *J. Am. Chem. Soc.* **1968**, 90, 4328.
- Free, S. M.; Wilson, J. W. A mathematical contribution to structure–activity studies. *J. Med. Chem.* **1964**, 7, 395–399.
- Hansch, C.; Fujita, T. ρ – σ – π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *J. Am. Chem. Soc.* **1964**, 86, 148–152.
- Kubinyi, H. *QSAR: Hansch Analysis and Related Approaches*; VCH: Weinheim, 1993.
- Martin, Y. C. *Quantitative Drug Design: A Critical Introduction*; Marcel Dekker: New York, 1978.
- Chemical Application of Atomic and Molecular Electrostatic Potentials*; Politzer, P., Truhlar, D. G., Eds.; Plenum: New York, 1981.
- Verloop, A.; Hoogenstraaten, W.; Tipker, J. In *Drug Design*, Vol. VII; Ariens, E. J., Ed.; Academic Press: New York, 1976; p 165.
- Sylvester, J. J. *Am. J. Math.* **1874**, 1, 64.
- Wiener, H. Structural determination of boiling points. *J. Chem. Phys.* **1947**, 69, 17–20.
- Kier, L. B.; Hall, L. H.; Murray, W. J.; Randic, M. Molecular connectivity. I: Relationship to nonspecific local anaesthesia. *J. Pharm. Sci.* **1975**, 64, 1971–1974.
- Randic, M. On Characterization of Molecular Branching. *J. Am. Chem. Soc.* **1975**, 97, 6609–6615.
- Trinajstić, N. *Chemical graph theory*; CRC Press: Boca Raton, 1992.
- Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research*; Academic Press: New York, 1976.
- Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure–Activity Analysis*; Research Studies Press: Letchworth, 1986.
- Cramer, R. D., III; Paterson, D. E.; Brunce, J. D. Comparative Molecular Field Analysis (COMFA). 1 Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1998**, 110, 5959–5967.
- Klebe, G.; Abraham, U.; Mietzner, T. Molecular Similarity Indices in a Comparative Analysis (CoMSIA) of Drug Molecules to Correlate and Predict their Biological Activity. *J. Med. Chem.* **1994**, 37, 4130–4146.
- Silverman, B. D.; Platt, D. E. Comparative Molecular Moment Analysis (CoMMA): 3D-QSAR without molecular superposition. *J. Med. Chem.* **1996**, 39, 2129–2140.
- Jain, A. N.; Koile, K.; Chapman, D. Compass: Predicting Biological Activities from Molecular Surface Properties. Performance Comparisons on a Steroid Benchmark. *J. Med. Chem.* **1994**, 37, 2315–2327.
- Bravi, G.; Gancia, E.; Mascagni, P.; Pegna, M.; Todeschini, R.; Zaliani, A. MS–WHIM, a new 3D theoretical descriptors derived from molecular surface properties: A comparative 3D-QSAR study in a series of steroids. *J. Comput.-Aided Mol. Design* **1997**, 11, 79–92.
- Kellogg, G. E.; Kier, L. B.; Gaillard, P.; Hall, L. H. E-state fields: Applications to 3D QSAR. *J. Comput.-Aided Mol. Design* **1996**, 10, 513–520.
- Chen, H.; Zhou, J.; Xie, G. PARM: A Genetic Evolved Algorithm to Predict Bioactivity. *J. Chem. Inf. Comput. Sci.* **1998**, 38, 243–250.
- Martín, M.; Sanz, F.; Campillo, M.; Pardo, L.; Pérez, J.; Turmo, J.; Aulló, J. M. Quantum Chemical Structure–Activity Relationships on β -Carbolines as Natural Monoamine Oxidase Inhibitors. *Int. J. Quantum Chem.* **1983**, 23, 1643–1652.
- Sanz, F.; Martín, M.; Lapeña, F.; Manaut, F. Quantitative structure–activity relationships on MAO substrates by means of quantum chemical properties. *Quant. Struct.-Act. Relat.* **1986**, 5, 54–57.
- Sanz, F.; Manaut, F.; José, J.; Segura, J.; Carbó, M.; De la Torre, R. Automatic determination of MEP patterns of molecules and its application to caffeine metabolism inhibitors. *J. Mol. Struct. (THEOCHEM)* **1988**, 170, 171–180.
- Luque, F. J.; Sanz, F.; Illas, F.; Pouplana, R.; Smeyers, Y. G. Relationships between the activity of some H₂-receptor agonists of histamine and their ab initio molecular electrostatic potential (MEP) and electron density comparison coefficients. *Eur. J. Med. Chem.* **1988**, 23, 7–10.
- Sanz, F.; Manaut, F.; Dot, T.; López de Briñas, E. Complete or partial comparison of molecular electrostatic potential distributions? Some tests with 5-HT ligands. *J. Mol. Struct. (THEOCHEM)* **1992**, 256, 287–293.
- Sanz, F.; Manaut, F.; Rodríguez, J.; Lozoya, E.; López de Briñas, E. MEPSIM: A computational package for analysis and comparison of molecular electrostatic potentials. *J. Comput.-Aided Mol. Design* **1993**, 7, 337–347.
- Burt, C.; Huxley, P.; Richards, W. G. The Application of Molecular Similarity Calculations. *J. Comput. Chem.* **1990**, 11, 1139–1146.
- Richard, A. M. Quantitative Comparison of Molecular Electrostatic Potentials for Structure Activity Studies. *J. Comput. Chem.* **1991**, 12, 959–969.
- Rum, G.; Herndon, W. C. Molecular Similarity Concepts. 5. Analysis of Steroid-Protein Binding Constants. *J. Am. Chem. Soc.* **1991**, 113, 9055–9060.
- Herndon, W. C.; Rum, G. Three-Dimensional Topological Descriptors and Similarity of Molecular Structures: Binding Affinities of Corticosteroids. In *QSAR and Molecular Modeling: Concepts, Computational Tools and Biological Applications: Proceedings of the 10th European Symposium on Structure–Activity Relationships, QSAR and Molecular Modelling*; Sanz, F., Giraldo, J., Manaut, F., Eds.; Prous Science: Madrid, 1995; pp 380–384.
- Good, A. C.; So, S.-S.; Richards, W. G. Structure–Activity Relationships from Molecular Similarity Matrices. *J. Med. Chem.* **1993**, 36, 433–438.
- Good, A. C.; Peterson, S. J.; Richards, W. G. QSAR's from Similarity Matrices. Technique Validation and Application in the Comparison of Different Similarity Evaluation Methods. *J. Med. Chem.* **1993**, 36, 2929–2937.
- Good, A. C.; Richards, W. G. The extension and application of Molecular Similarity to Drug Design. *Drug Information J.* **1996**, 30, 371–388.
- Good, A. C.; Hodgkin, E. E.; Richards, W. G. the Utilization of Gaussian Functions for the Rapid Evaluation of Molecular Similarity. *J. Chem. Inf. Comput. Sci.* **1992**, 32, 188–191.
- Good, A. C.; Richards, W. G. Rapid Evaluation of Shape Similarity Using Gaussian Functions. *J. Chem. Inf. Comput. Sci.* **1993**, 33, 112–116.
- Carbó, R.; Arnau, J.; Leyda, L. How similar is a molecule to another? An electron density measure of similarity between two molecular structures. *Int. J. Quantum Chem.* **1980**, 17, 1185–1189.
- Carbó-Dorca, R.; Besalú, E. A general survey of molecular quantum similarity. *THEOCHEM* **1998**, 451, 11–23.
- Carbó-Dorca, R.; Amat, L.; Besalú, E.; Lobato, M. Quantum Similarity. In *Advances in Molecular Similarity*; Carbó-Dorca, R., Mezey, P. G., Eds.; JAI Press: Greenwich, CT, 1998; Vol. 2.
- Carbó-Dorca, R. Fuzzy sets and Boolean tagged sets; vector semispaces and convex sets; quantum similarity measures and ASA density functions; diagonal vector spaces and quantum chemistry. In *Advances in molecular similarity*; Carbó-Dorca, R., Mezey, P. G., Eds.; JAI Press: Greenwich, CT, 1998; Vol. 2, pp 1–42.
- Besalú, E.; Carbó, R.; Mestres, J.; Solà, M. Foundations and recent developments on molecular quantum similarity. *Top. Curr. Chem.* **1995**, 173, 31–62.
- Carbó, R.; Besalú, E.; Theoretical foundations of quantum molecular similarity. In *Molecular similarity and reactivity: from quantum chemical to phenomenological approaches*; Carbo, R., Ed.; Kluwer: Amsterdam, 1995; pp 3–30.
- Carbó-Dorca, R.; Besalú, E.; Amat, L.; Fradera, X. Quantum molecular similarity measures: concepts, definitions and applications to quantitative structure-properties relationships. In *Advances in molecular similarity*; Carbó-Dorca, R., Mezey, P. G., Eds.; JAI Press: Greenwich, CT, 1996; Vol. 1, pp 1–42.
- Carbó, R.; Besalú, E.; Amat, L.; Fradera, X. Quantum molecular similarity measures (QMSM) as a natural way leading towards a theoretical foundation of quantitative structure-properties relationships. *J. Math. Chem.* **1995**, 18, 237–246.
- Fradera, X.; Amat, L.; Besalú, E.; Carbó-Dorca, R. Application of molecular quantum similarity to QSAR. *Quant. Struct.-Act. Relat.* **1997**, 16, 25–32.
- Lobato, M.; Amat, L.; Carbó-Dorca, R. Structure–activity relationships of a steroid family using quantum similarity measures and topological quantum similarity indices. *Quant. Struct.-Act. Relat.* **1997**, 16, 465–472.

- (52) Amat, L.; Robert, D.; Besalú, E.; Carbó-Dorca, R. Molecular quantum similarity measures tuned 3D QSAR: an antitumoral family validation study. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 624–631.
- (53) Robert, D.; Amat, L.; Carbó-Dorca, R. Three-dimensional quantitative structure–activity relationships from tuned molecular quantum similarity measures. Prediction of the corticosteroid-binding globulin binding affinity for a steroid family. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 333–344.
- (54) Robert, D.; Gironés, X.; Carbó-Dorca, R. Facet diagrams for quantum similarity data. *J. Comput.-Aided Mol. Des.* **1999**, *13*, 597–610.
- (55) Gironés, X.; Amat, L.; Robert, D.; Carbó-Dorca, R. Use of electron–electron repulsion energy as a molecular descriptor in QSAR and QSPR studies. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 477–485.
- (56) Robert, D.; Carbó-Dorca, R., Aromatic compounds aquatic toxicity QSAR using quantum similarity measures. *SAR QSAR Environ. Res.* **1999**, *10*, 401–422.
- (57) Gironés, X.; Amat, L.; Carbó-Dorca, R. Using molecular quantum similarity measures as descriptors in quantitative structure–toxicity relationships. *SAR QSAR Environ. Res.* **1999**, *10*, 545–556.
- (58) Dean, P. M. Molecular Similarity. In *3D QSAR in Drug Design: Theory, Methods and Applications*; Kubinyi, H., Ed.; ESCOM Science Publishers B. V.: Leiden, The Netherlands, 1993; pp 150–172.
- (59) Richards, W. G. Molecular Similarity and Dissimilarity. In *Modeling of Biomolecular Structures and Mechanisms*; Pullman, A., Jortner, J., Pullman, B., Eds.; Kluwer Academic Publishers: Dordrecht, The Netherlands, 1995; pp 365–369.
- (60) Carbó-Dorca, R.; Amat, L.; Besalú, E.; Gironés, X.; Robert, D. Quantum Mechanical Origin of QSAR: Theory and Applications. *J. Mol. Struct. (THEOCHEM)* **2000**, *504*, 181–228.
- (61) Carbó-Dorca, R.; Besalú, E.; Gironés, X. Extended Density Functions and Quantum Chemistry. *Adv. Quan. Chem.* in press.
- (62) Carbó-Dorca, R.; Amat, L.; Besalú, E.; Gironés, X.; Robert, D. Quantum Molecular Similarity: Theory and Applications to the Evaluation of Molecular Properties, Biological Activities and Toxicity. Proceedings of the 4th Girona Seminar on Molecular Similarity; Kluwer: in press.
- (63) TDR News (News from the WHO Division of Control of Tropical Diseases) **1994**, *46*, 5.
- (64) Peters, W. *Chemotherapy and Drug Resistance in Malaria*, 2nd ed.; Academic Press: London, 1987.
- (65) Oaks, S. C.; Mitchell, V. S.; Pearson, G. W.; Carpenter, C. C. *MALARIA Obstacles and Opportunities*; National Academic Press: Washington, DC, 1991.
- (66) Oduola, A. M. J.; Omitowoju, G. O.; Gerena, L.; Kyle, D. E.; Milhous, M. K.; Sowunmi, A.; Salako, L. A. Reversal of Mefloquine Resistance with Penfluridol in Isolates of Plasmodium Falciparum from South-West Nigeria. *Trans. R. Soc. Trop. Med. Hyg.* **1993**, *87*, 81–83.
- (67) Klayman, D. L. Qinghaosu (Artemisinin): an antimalarial drug from China. *Science* **1985**, *228*, 1049–1055.
- (68) Luo, X.-D.; Shen, C.-C. The chemistry, pharmacology, and clinical applications of qinghaosu (artemisinin) and its derivatives. *Med. Res. Rev.* **1987**, *7*, 29–52.
- (69) Zhou, W.-S.; Xu, X.-X. Total Synthesis of the Antimalarial Sesquiterpene Peroxide Qinghaosu and Yingzhaosu A. *Acc. Chem. Res.* **1994**, *27*, 211–216.
- (70) Zaman, S. S.; Sharma, R. P. Some Aspects of the Chemistry and Biological Activity of Artemisinin and Related Antimalarials. *Heterocycles* **1991**, *32*, 1593–1638.
- (71) Jung, M. Current Developments in the Chemistry of Artemisinin and Related Compounds. *Curr. Med. Chem.* **1994**, *1*, 35–49.
- (72) Avery, M. A.; Gao, F.; Mehrotra, S.; Chong, W. K.; Jennings-White, C. *The Organic and Medicinal Chemistry of Artemisinin and Analogues. Research Trends Trivandrum*; India, 1993.
- (73) Posner, G. H.; Ploypradith, P.; Parker, M. H.; O'Dowd, H.; Woo, S.-H.; Northrop, J.; Krasavin, M.; Dolan, P.; Kensler, T. W.; Xie, S.; Shapiro, T. A. Antimalarial, Antiproliferative, and Antitumor Activities of Artemisinin-Derived, Chemically Robust, Trioxane Dimers. *J. Med. Chem.* **1999**, *42*, 4275–4280.
- (74) Grigorov, M.; Weber, J.; Tronchet, J. M. J.; Jefford, C. W.; Milhous, W. K.; Maric D. A QSAR Study of the Antimalarial Activity of Some Synthetic 1,2,4-Trioxanes. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 124–130.
- (75) Constants, P.; Carbó, R. Atomic shell approximation: electron density fitting algorithm restricting coefficients to positive values. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1046–1053.
- (76) Amat, L.; Carbó-Dorca, R. Quantum similarity measures under atomic shell approximation: first-order density fitting using elementary Jacobi rotations. *J. Comput. Chem.* **1997**, *18*, 2023–2039.
- (77) Amat, L.; Carbó-Dorca, R. Fitted electronic density functions from H to Rn for use in quantum similarity measures: cis-diamminedichloroplatinum(II) complex as an application example. *J. Comput. Chem.* **1999**, *20*, 911–920.
- (78) Amat, L.; Carbó-Dorca, R. Molecular Electronic Density Fitting Using Elementary Jacobi Rotations under Atomic Shell Approximation. *J. Chem. Inf. Comput. Sci.* in press.
- (79) Gironés, X.; Robert, D.; Carbó-Dorca, R. TGSA: a molecular superposition program based on Topo-Geometrical Considerations. IT-IQC-00-11. *J. Comput. Chem.* in press.
- (80) PC Spartan Pro, version 1.4; Wavefunction, Inc.: Irvine, 1997.
- (81) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Gill, P. M. W.; Johnson, B. G.; Robb, M. A.; Cheeseman, J. R.; Keith, T.; Petersson, G. A.; Montgomery, J. A.; Raghavachari, K.; Al-Laham, M. A.; Zakrzewski, V. G.; Ortiz, J. V.; Foresman, J. B.; Cioslowski, J.; Stefanov, B. B.; Nanayakkara, A.; Challacombe, M.; Peng, C. Y.; Ayala, P. Y.; Chen, W.; Wong, M. W.; Andres, J. L.; Replogle, E. S.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Binkley, J. S.; Defrees, D. J.; Baker, J.; Stewart, J. P.; Head-Gordon, M.; Gonzalez C.; Pople, J. A. Gaussian-94, Revision E.2; Gaussian, Inc.: Pittsburgh, PA, 1995.
- (82) Robert, D.; Gironés, X.; Carbó-Dorca, R. Quantification of the Influence of Single-Point Mutations on Haloalkane Dehalogenase Activity: A Molecular Quantum Similarity Study. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 839–46.
- (83) Cox, T. F.; Cox, M. A. A. *Multidimensional Scaling*; Chapman & Hall: London, 1994.
- (84) Carbó, R.; Besalú, E. Nested Summation Symbols and Perturbation Theory. *J. Math. Chem.* **1993**, *13*, 331–342.
- (85) Carbó, R.; Besalú, E. Definition, Mathematical Examples and Quantum Chemical Applications of Nested Summation Symbols and Logical Kronecker Deltas. *Comput. Chem.* **1994**, *18*(2), 117–126.
- (86) Carbó, R.; Besalú, E. Definition and Quantum Chemical Applications of Nested Summation Symbols and Logical Functions: Pedagogical Artificial Intelligence Devices for Formulae Writing, Sequential Programming and Parallel Implementation. *J. Math. Chem.* **1995**, *18*, 37–72.
- (87) Carbó, R.; Besalú, E. Application of Nested Summation Symbols to Quantum Chemistry: Formalism and Programming Techniques. In *Strategies and Applications in Quantum Chemistry: from Astrophysics to Molecular Engineering*; DeFranceschi, M., Ellinger, Y., Eds.; Kluwer Academic Publishers: Amsterdam, 1996; pp 229–248.
- (88) Allen, D. M. The relationship between variable selection and data augmentation and a method for prediction. *Technometrics* **1974**, *16*, 125–127.
- (89) Clementi, S.; Wold, S. How to choose the proper statistical method. In *Chemometric Methods in Molecular Design*; Van der Waterbeemd, H., Ed.; VCH: New York, 1995; pp 319–338.
- (90) Wold, S.; Erikson, L. Statistical validation of QSAR results. In *Chemometric Methods in Molecular Design*; Van der Waterbeemd, H., Ed.; VCH: New York, 1995; pp 309–318.
- (91) Posner, G. H.; Parker, M. H.; Northrop, J.; Elias, J. S.; Ploypradith, P.; Xie, S.; Shapiro, T. A. Orally Active, Hydrolytically Stable, Semisynthetic, Antimalarial Trioxanes in the Artemisinin Family. *J. Med. Chem.* **1999**, *42*, 300–304.
- (92) Posner, G. H.; González, L.; Cumming, J. N.; Klinedinst, D.; Shapiro, T. A. Synthesis and Antimalarial Activity of Heteroatom-Containing Bicyclic Endoperoxides. *Tetrahedron* **1997**, *53*, 37–50.
- (93) Avery, M. A.; Mehrotra, S.; Bonk, J. D.; Vroman J. A.; Goins, D. K.; Miller, R. Structure–Activity Relationships of the Antimalarial Agent Artemisinin. 4. Effect of Substitution at C-3. *J. Med. Chem.* **1996**, *39*, 2900–2906.
- (94) Desjardins, R. E.; Canfield, C. J.; Haynes, D. E.; Chulay, J. D. Quantitative Assessment of Antimalarial Activity In Vitro by a Semi-automated Microdilution Technique. *Antimicrob. Agents Chemother.* **1979**, *16*, 710–718.
- (95) Milhous, W. K.; Weatherly, N. F.; Bowdre, J. H.; Desjardins, R. E. In Vitro Activities of and Mechanisms of Resistance to Antifolate Antimalarial Drugs. *Antimicrob. Agents Chemother.* **1985**, *27*, 525–530.

Using Molecular Quantum Similarity Measures under Stochastic Transformation To Describe Physical Properties of Molecular Systems

Xavier Gironés and Ramon Carbó-Dorca*

Institute of Computational Chemistry, University of Girona, Campus Montilivi, 17071 Girona, Catalonia, Spain

Received September 26, 2001

The application of molecular quantum similarity measures (MQSM) to correlate physicochemical properties is reported. Satisfactory quantitative structure–property relationship (QSPR) models are obtained for three molecular sets, where boiling points and chromatographic retention times and indices are studied. In this work, MQSM are scaled using a stochastic transformation and related to molecular properties using the partial least-squares technique.

INTRODUCTION

One of the most promising subjects in present day computational chemistry is the characterization and quantification of molecular properties by means of theoretical descriptors. Given a set of molecules and a related molecular property, a quantitative structure–property or –activity relationship (QSPR/QSAR) model can be constructed by deriving descriptors from molecular structure, which will be related to the molecular feature by some statistical technique.

These disciplines were first started by Cross in 1863,¹ when he observed that toxicity of alcohols to mammals increased as water solubility of such alcohols decreased. Since then, QSPR/QSAR techniques have been largely developed using a variety of parameters as molecular descriptors of diverse structural characteristics. For example, Hammett σ values² are often used as electronic parameters, and other parameters have been devised to account for the shape, size, lipophilicity, polarizability, and many other molecular structural features. In addition, the recent and fast development of computer architectures has increased the computational power to stages that allow the application of the quantum theory to regular organic molecules at fairly accurate computational levels (semiempirical or ab initio methods with appropriate basis sets) within reasonable time limits and affordable costs. A number of reviews have been published^{3–5} concerning the historical development, generation of descriptors, and different methodologies in the QSPR/QSAR fields.

Arising from the conception of a molecule provided by quantum chemistry, and according to the postulate that its electronic density function contains all accessible information contained within this molecule, molecular quantum similarity measures (MQSM)^{6–12} stand as a general efficient tool to solve actual chemical problems. MQSM methodology has been successfully applied within pharmacological^{13–18} and toxicological^{19–21} problems. MQSM are based on the concept of molecular similarity^{22,23} and are related to a self-evident

molecular similarity principle: “the more similar two molecules are, the more similar properties they will possess”. This last statement requires a procedure to compare the molecules, and MQSM easily establish the degree of similarity on the basis of the electronic distribution, namely first-order density functions (DF), of the molecular structures.

In this work, it is intended to use the MQSM methodology to correlate boiling points and chromatographic parameters of simple organic molecules. For this purpose, this paper is structured as follows: a survey around the employed protocol is given first, next the presentation of the molecular sets, as well as the results achieved, are described, and the concluding remarks are finally given.

MATERIAL AND METHODS

MQSM and Stochastic Transformation. The practical application of MQSM relies on the use of first-order density functions. These functions are quantum-mechanical observable elements producing information on the molecular electron distribution. Within this framework, two molecular structures are considered to be similar if their electron distributions are similar. Thus, among other possibilities,⁸ quantitative measures of the similarity between two molecules can be defined as the direct volume integral between their density functions weighted by the *Coulomb* operator:

$$Z_{AB} = \int \int \rho_A(\mathbf{r}_1) |\mathbf{r}_1 - \mathbf{r}_2|^{-1} \rho_B(\mathbf{r}_2) d\mathbf{r}_1 d\mathbf{r}_2 \quad (1)$$

Here $\rho_A(\mathbf{r})$ and $\rho_B(\mathbf{r})$ are the first-order electron density functions of molecules A and B, respectively, and Z_{AB} is the resulting quantum similarity measure. As MQSM depend on the relative position of both compared objects in space, a molecular alignment procedure is required. When studied molecular sets share common structural features, the topogeometrical superposition algorithm (TGSA)²⁴ is used, as it performs pairwise superpositions according to the molecular backbones. Once superposed, the overall set of MQSM for a series of N molecules is computed and collected in the way of an $(N \times N)$ similarity matrix form, which can be

* Corresponding author phone: (+34)972418367; fax: (+34)972418356; e-mail: quantum@iqc.udg.es.

used to extract the appropriate correlation information and build QSPR/QSAR descriptors. Once the similarity matrix has been computed, a possible scaling of the MQSM can be done by means of eq 2, providing the recently described *stochastic transformation*²⁵ of MQSM. The construction of this stochastic MQSM matrix conceives that the sum of every row elements can be used as scale factors to obtain a new row. This procedure originates a new nonsymmetric square matrix, which will act as a set of input descriptors in any subsequent statistical procedure.

$$S_{ij} = Z_{ij} \left(\sum_j Z_{ij} \right)^{-1} \quad (2)$$

Promolecular Atomic Shell Approximation (ASA). To avoid expensive theoretical calculations, the promolecular atomic shell approximation (ASA)^{26–28} has been used here to construct molecular first-order electron density. Within ASA the molecular density is expressed as a sum of discrete atomic density contributions, which are taken as a superposition of 1S Gaussian functions and fitted to atomic ab initio ones. Once the overall atomic densities are built, each molecular density function can be constructed by adding as appropriate these elementary atomic building blocks. Since it has been proved that the MQSM from fitted densities built in this way differ by up to a 2% from the ab initio ones,²⁷ their use is clearly justified.

Molecular Alignment. MQSM are dependent on the relative position of the molecules under comparison. In this way, the topo-geometrical approach (TGSA)²⁴ has been used to perform the needed pairwise molecular alignments. This molecular superposition method overlays the involved molecules according to the maximal common substructure shared by the analyzed compounds.

Molecular Modeling. All molecules have been constructed using the Ampac 6.55²⁹ software package. The geometry optimization of the resulting structures has been carried out in gas phase with the same program at the AM1³⁰ computational level.

Treatment of Quantum Similarity Matrixes and Model Building. Common chemometric tools may be applied to deal with similarity matrices. Particularly, partial least squares (PLS)^{31,32} stands as an ideal technique to obtain a generalized regression to model the association between the matrices **X** (descriptors) and **Y** (responses). In computational chemistry, its main use is to model the relationship between computed variables, which together characterize the structural variation of a set of *N* compounds and a property of interest measured on those *N* substances.^{33–35} This variation of the molecular skeleton is condensed into the matrix **X**, whereas the analyzed properties are recorded into **Y**. In PLS, the matrix **X** is commonly built up from nonindependent data, as it uses more columns than rows; hence it is not called the *independent* matrix but a predictor or descriptor matrix. A good review, as well as its practical application in QSAR, is found in ref 36, and a detailed tutorial in ref 37.

Unlike regression, PLS is not based on the assumption of independent and precise **X** variables, but it is rather based on the more realistic assumption that **X** contains more or less collinear and noisy parameters. PLS summarizes these **X** variables by means of a few orthogonal score vectors (**t**_{*a*}

∈ **T**), and the matrix **Y** is also resumed in few score vectors (**u**_{*a*} ∈ **U**) that are not orthogonal. Plots of columns from **T** and **U** provide visual representation of the configuration of the observations in the **X** or **Y** spaces, respectively. To speed up the PLS routine, the dimensionality of the similarity matrix is a priori reduced in two ways:

(a) removal of columns with a variation coefficient below 2%, thus eliminating those descriptors that are roughly a vector of constants;^(b)

extraction of all columns that, after pairwise correlation, score a goodness-on-fit coefficient larger than 0.85, thus eliminating redundant descriptors.

The PLS procedure allows one to derive a number of *factors* and *weights*, which are used to describe the desired properties. QSPR models are built up from these factors and weights.

In this work, all obtained models are evaluated by commonly used statistical parameters: goodness-on fit (*r*²);³⁸ root-mean-square error between experimental and predicted values (*s*)³⁸ by *leave-one-out*;³⁹ predictive capacity (*q*²).³⁸ In addition, and in order to avoid chance correlations and excess of parameters, models are submitted to random tests, where the properties are randomly permuted in their positions and the entire modeling procedure is repeated a number of times. If satisfactory correlations are found within the random test, the model obtained should not be trusted, as the methodology used may be potentially capable to correlate any kind of data.

RESULTS AND DISCUSSION

In this section, it is intended to prove the usefulness of the proposed procedure with three different molecular sets. The molecular properties analyzed consist of boiling points, gas chromatographic retention times, and gas-liquid chromatographic retention indices, respectively. The boiling points are analyzed for a molecular set composed of 529 saturated hydrocarbons⁴⁰ (linear, branched, and cyclic). The retention times are contemplated for a set of 152 diverse chemical compounds.⁴¹ Finally, the retention indices are evaluated over a set of 50 phenol derivatives.⁴²

Boiling Point for 529 Saturated Hydrocarbons. Boiling point at normal pressure (bp) has become a benchmark property for the evaluation of novel QSPR descriptors.^{43,44} In this way, it is interesting to test the present proposed MQSM methodology for such a homologous molecular series. For this purpose, a set of 529 saturated hydrocarbons, whose data are presented in Table 1, was tested for correlation search with their bp using the previously described methodology. The initial 529 × 529 matrix was reduced to 529 × 513 after the redundant variable reduction. QSPR models have been constructed up to the usage of 15 PLS factors, in which both *r*² and *q*² have been inspected. The optimal number of parameters to be used in the final model, which in this case has been fixed to 6 PLS factors, has been chosen from the evolution of these magnitudes, as presented in Figure 1.

As observed in Figure 1, a sharp increase in the predictive capacity occurs up to the sixth PLS factor and a slow increase until the eighth one. A further increase in the number of parameters results, in despite of the expected increase of *r*², in a decrease of the predictive power of the generated model. In this way, the chosen number of PLS factors used was 6,

MOLECULAR QUANTUM SIMILARITY MEASURES

Table 1. SMILES Notation^{47,48} and Boiling Points (°C) of 529 Saturated Hydrocarbons⁴⁰

compd	bp	compd	bp	compd	bp
CCC	32.8	CCCCCCCCC	99.2	CCCC(C)(C)CCCC	172.5
C ₁ CC ₁	42.1	CC(C)CCCCC	118.2	CCC(CC)C(CC)CC	166
CCCC	8	CCC(C)CCCC	109.8	CC(C)(C)CC(C)CCC	164.5
CC(C)C	0.7	CCCC(C)CCCC	114.8	CC(C)C(C)C(C)CCC	179.5
C ₁ (C)CC ₁	12.6	CC(C)CCCC(C)C	106.5	CC(C)CCC(C)(C)CC	171.5
C ₁ CCC ₁	0.5	CCCCC(CC)CC	140.5	CC(C)C(CCC)C(C)C	175.8
C ₁₂ CC ₁ C ₂	11.7	CCCC(CC)CCC	168.3	CCC(C)C(C)C(C)CC	160.5
CCCCC	36	CC(C)CCC(C)CC	164.6	CC(C)(C)CC(CC)CC	165
CC(C)CC	46	CC(C)C(C)CCCC	162.5	CCC(C)(CC)CCCC	172.5
CC(C)(C)C	39	CCC(C)C(C)CCC	166	CC(C)C(C)C(C)CC	186.2
C ₁ (CC ₁)CC	33.5	CCC(C)CC(C)CC	161.4	CC(C)C(C)C(C)CC	174.5
C ₁ (C)CC ₁ C	35.9	CCC(CC)CC(C)C	142.5	CC(C)CC(C)(C)CCC	164
C ₁ (C)(C)CC ₁	32.6	CC(C)(C)CCCCC	147.8	CC(C)(C)C(C)CCCC	160.5
C ₁ (C)CCC ₁	20.6	CCC(C)C(C)CCC	140.7	CCC(C)(C)CC(C)CC	159.5
C ₁ CCCC ₁	36.3	CCCC(CC)C(C)C	161	CCCC(C)(CC)CCC	162
C ₁ C ₂ CC ₁ C ₂	49.3	CC(C)CC(C)CCC	146	CC(C)(C)CC(C)C(C)C	172.5
C ₁ CC ₂ CC ₁₂	36	CCC(C)CC(C)CC	137.5	CC(C)(C)CCC(C)(C)C	165.5
C ₁ CC ₁₂ CC ₂	27.8	CC(C)(C)CCC(C)C	137.5	CC(C)C(C)C(C)C(C)C	159.5
C ₁₂ (C)CC ₁ C ₂	9.5	CCC(C)(C)CCCC	131	CC(C)C(C)(C)CCCC	163
CCCCC	71	CCCCC(C)(C)CCC	128	CC(C)CC(C)(C)CCC	150
CC(C)CCC	76	CC(C)C(C)C(C)CC	152.5	CC(C)(C)C(CC)CCC	168
CCC(C)CC	83	CC(C)(C)CC(C)CC	169	CC(C)(C)C(C)CC(C)C	153
CC(C)C(C)C	81	CC(C)C(C)C(C)C	161.8	CCC(CC)(CC)CCC	168.5
CC(C)(C)CC	69.5	CCC(C)(CC)CCC	163	CC(C)C(C)C(C)C(C)C	160
C ₁ (CC ₁)CCC	60.5	CC(C)CC(C)(C)CC	163	CCC(C)C(C)C(C)CCC	152
C ₁ (CC ₁)C(C)C	55	CC(C)(C)C(C)CCC	164	CCC(C)C(C)(C)CCC	157.5
C ₁ (CC)CC ₁ C	69	CCC(CC)(CC)CC	157.8	CC(C)C(C)(C)CC(C)C	158.5
C ₁ (C)(CC ₁)CC	58.3	CC(C)C(C)C(C)CC	168	CCC(C)(C)C(C)CC	159.5
C ₁ (C)C(C)C ₁ C	63	CC(C)(C)C(CC)CC	155	CC(C)C(C)(C)CCC	152
C ₁ (C)(C)CC ₁ C	57	CCC(C)(C)C(C)C	154	CC(C)(C)CC(C)(C)CC	155.1
C ₁ (CC)CCC ₁	63	CC(C)(C)C(C)(CC)CC	152.5	CC(C)(C)C(C)C(C)CC	148.3
C ₁ (C)CC(C ₁)C	52.6	CC(C)(C)CC(C)(C)C	157	CCC(C)(CC)C(C)CC	153.5
C ₁ (C)CCC ₁ C	70.7	CC(C)(C)C(C)C(C)C	159	CC(C)(C)C(CC)C(C)C	146.7
C ₁ (C)(C)CCC ₁	59	CC(C)C(C)(C)C(C)C	149	CC(C)C(C)C(C)(C)CC	146
C ₁ CCC(C ₁)C	62	CC(C)(C)C(C)(C)CC	153	CC(C)C(CC)(CC)CC	158
C ₁ CCCCC ₁	53.6	C ₁ (CC ₁)CCCCC	145	CC(C)C(C)(C)C(C)CC	155.5
C ₁₂ CCC(C ₁)C ₂	71.8	C ₁ (CC ₁)C(C)CCCC	152	CC(C)C(C)(CC)C(C)C	154
C ₁ CC ₁ C ₂ CC ₂	80.7	C ₁ (CC ₁ C)CCCCC	148.2	CC(C)C(C)C(C)CC	145.5
C ₁ CC ₂ C ₁ CC ₂	68.7	C ₁ (CC)CC ₁ CCC	151	CC(C)(C)C(C)C(C)(C)C	180
C ₁₂ CCCC ₁ C ₂	60.3	C ₁ CC ₁ (CC)CCCC	146.7	CCC(C)C(C)C(C)CC	172
C ₁ CCC ₁₂ CC ₂	63.3	CC(C)(C)CC ₁ (C)CC ₁	151.5	CC(C)(C)C(C)(CC)CC	170.7
C ₁₂ (C)CCC ₁ C ₂	58	CC(C)CC ₁ CC ₁ (C)C	149	CC(C)(C)C(C)C(C)C	176.5
C ₁₂ (C)CC ₁ (C ₂)C	49.7	C ₁ (C)(C)CC ₁ C(C)(C)C	151	C ₁ (C)(CC ₁)CCCC(C)C	172
CCCCCCC	108.5	C ₁ (C)(CC)CC ₁ (C)CC	143	CC(C)CC ₁ CC ₁ C(C)C	153
CC(C)CCCC	104	C ₁ (C)(C)C(C)(C)C ₁ (C)C	138.8	C(C)CCC ₁ CC ₁ (C)C	175.7
CCC(C)CCC	105	C ₁ CCC ₁ C(CC)CC	146.5	CCCC ₁ CC ₁ (C)CCC	173.9
CCC(CC)CC	107	C ₁ CCCC ₁ CCCC	142	CC ₁ (C)CC ₁ C(C)(C)CC	173.5
CC(C)CC(C)C	110	C ₁ CCCC ₁ CC(C)C	148.2	C ₁ (C)C(C ₁)C(C)C(C)(C)C	160
CC(C)C(C)CC	107.5	C ₁ C(C)CCC ₁ CCC	150	C ₁ (CCC ₁ C(C)C)C(C)C	172
CC(C)(C)CCC	105.5	C ₁ CCCC ₁ C(C)CC	150.5	C ₁ (CC)C(C)C(C)C ₁ CC	173.9
CCC(C)(C)CC	110	C ₁ C(C)CCCC ₁ CC	132.2	C ₁ (C)(C)C(CC)CC ₁ CC	160.5
CC(C)(C)C(C)C	110.5	C ₁ (C)CCCC ₁ CCC	154.5	C ₁ (C)C(CC ₁ C(C)C)C(C)C	148.5
C ₁ CC ₁ CCCC	116	C ₁ (CC)CCCC ₁ CC	141	C ₁ CCCC ₁ CCCC	172
C ₁ CC ₁ C(C)CC	96.5	C ₁ C(C)CCC ₁ C(C)C	131	C ₁ CCCC ₁ CCCC(C)C	171
C ₁ (C)CC ₁ CCC	98.5	C ₁ (C)CCCC ₁ C(C)C	135	C ₁ (C)CCCC ₁ CCCC	163
C ₁ (CC)CC ₁ CC	100	C ₁ CCCC ₁ (C)CCC	127.2	C ₁ CCCC ₁ C(C)CCC	164
C ₁ (C)(CC ₁)CCC	103	C ₁ C(C)C(C)CC ₁ CC	149	C ₁ CCCC ₁ C(C)CCC	141.5
C ₁ (CC ₁)C(C)C	81.5	C ₁ C(C)CC(C)C ₁ CC	143	C ₁ CC(CC ₁ C)CC(C)C	180.9
C ₁ (CC ₁)C(C)(C)C	92	C ₁ CC(C)C(C)C ₁ CC	153	C ₁ CCCC ₁ C(CC)CC	171.3
C ₁ (CC ₁)(CC)CC	71.5	C ₁ CC(C)C(C)C ₁ CC	140.2	C ₁ CCCC ₁ CC(C)CC	173.4
C ₁ (CC)C(C)C ₁ C	74	C ₁ (CCCC ₁)C(C)C	125	C ₁ CCCC ₁ C(C)C(C)C	169
C ₁ (C)(CC ₁)C(C)C	78	C ₁ (CC)(CC)CCCC ₁	125.5	CC(C)C ₁ CCCC ₁ CC	179.3
C ₁ (C)(C)CC ₁ CC	84	C ₁ (CC)CCC(C ₁)(C)C	121	C ₁ (CC)CCC(C ₁)CC	175.5
C ₁ (C)(CC)CC ₁ C	98	C ₁ (C)CCC(C ₁)(C)CC	130.8	CC(C)(CC)C ₁ CCCC ₁	172
C ₁ (C)(C)C(C)C ₁ C	90.3	C ₁ CCC(C ₁)(C)C(C)C	124.5	CC(C)C ₁ CC(C)C(C)C ₁	170
C ₁ (C)(C)CC ₁ (C)C	93	C ₁ CC(CC)C(C ₁)(C)C	148.7	CC(C)C ₁ CCC(C)(C)C ₁	176
C ₁ CCC ₁ CCC	90	C ₁ C(C)C(C)C(C)C ₁ C	156.6	C ₁ (C)C(C)CC(C)C ₁ CC	167
C ₁ CCC ₁ C(C)C	84.9	C ₁ CCC(C)C ₁ C(CC	148	C ₁ (CC)C(C)CC(C)C ₁ C	168.5
C ₁ C(C)CC ₁ CC	81.1	C ₁ C(C)C(C)CC ₁ (C)C	148.3	C ₁ (CCCC ₁ C)C(C)C	171
C ₁ CC(C)C ₁ CC	80.5	C ₁ (C)CC(C)CC ₁ (C)C	154.3	C ₁ (C)C(C)C(C)C(C)C ₁ C	174.3
C ₁ CCCC ₁ CC	88.6	C ₁ (C)C(C)CCC ₁ (C)C	148.2	C ₁ C(C)(C)CC(C)C ₁ (C)C	175
C ₁ C(C)CCCC ₁ C	91	C ₁ C(C)(C)CCC ₁ (C)C	149.5	C ₁ CCCC(C ₁)CCCC	171.5
C ₁ (C)CCCC ₁ C	81.5	C ₁ (C)CCC(C)C ₁ (C)C	150.5	C ₁ CCCC(C ₁)CC(C)C	179.5
C ₁ CCCC ₁ (C)C	79.1	C ₁ (C)(C)CCCC ₁ (C)C	141	C ₁ CC(CCC)CCCC ₁ C	168

Table 1. (Continued)

compd	bp	compd	bp	compd	bp
C ₁ CCCC(C ₁)C	85.2	C ₁ CCCCC ₁ CCC	145	C ₁ (C)CCCC(C ₁)CCC	166.6
C ₁ CCCCCCC ₁	78	C ₁ CCCCC(C ₁)C(C)	145	C ₁ CCCC(C ₁)C(C)CC	166.5
C ₁ CC ₁ CCC ₂ CC ₂	76	C ₁ CC(C)CCCC ₁ CC	142.5	C ₁ CC(C)CCCC ₁ CC	177.5
C ₁ CC ₂ CCC ₁ C ₂	100.7	C ₁ C(C)CCCC ₁ CC	138	C ₁ CC(C)CC(C ₁)CC	172.5
C ₁ CC ₂ CC(C ₁)C ₂	92.7	C ₁ (C)CCCCC ₁ CC	147	C ₁ (C)CCCCC ₁ C	153
C ₁₂ CCCC ₁ CC ₂	89.5	C ₁ (C)CC(C)CC(C ₁)C	151	C ₁ CC(CCC ₁)C(C)C	160.3
C ₁₂ CCCCC ₁ C ₂	94	C ₁ (C)CCCC(C ₁)C(C ₁)C	145	C ₁ CCCC(C)C ₁ CC	158
C ₁ CCC ₁₂ CCC ₂	103.5	C ₁ CCC(C)C(C)C ₁ C	151	C ₁ C(C)CCCC ₁ C(C)C	153
C ₁ CCCC ₁₂ CC ₂	91.3	C ₁ CCCC(C ₁)CC	133	C ₁ C(C)CC(C)CC ₁ CC	155
C ₁₂ CCC(C ₁)C ₂	95.6	C ₁ (C)CCC(C ₁)C(C)CC ₁	135.5	C ₁ (C)CCCCC ₁ C(C)C	167
C ₁₂ CCCC ₁ C ₂ C	97.9	C ₁ CCCC(C ₁)CC ₁ C	143	C ₁ CCCCC ₁ (C)CCC	161.5
C ₁₂ CCC(C ₁)(C ₂)C	101	C ₁ CCC(C ₁)C(C ₁)C	138	C ₁ (C)C(C)CCCC ₁ CC	182.8
C ₁₂ CCCC ₁ (C ₂)C	118.4	C ₁ CCCCC(C ₁)CC	135.9	C ₁ (CCCCC ₁)C(C)(C)C	164.2
C ₁ C ₂ (C)CC ₁ (C)C ₂	98.5	C ₁ CC(C)CCCC(C ₁)C	142.5	C ₁ (CC)(CC)CCCC ₁	173
C ₁₂ (C)CCCC ₁ (C ₂)C	90	C ₁ CCC(C)CC(C ₁)C	127	C ₁ CC(C)CCC ₁ (C)CC	170.7
C ₁ CC ₁₂ CC ₂ (C)C	92	C ₁ CCCC(C ₁)C(C ₁)C	129.5	C ₁ C(C)C(C)CC(C ₁)C(C)C	177.5
C ₁₂ (C)CC ₁ C ₂ (C)C	93.5	C ₁ CCCCC(C ₁)(C)C	134	C ₁ C(C)CCCC ₁ (C)CC	161
C ₁₂ CC ₂ CC ₃ C ₁ C ₃	80.5	C ₁ CCCC(C)CCC ₁	118.2	C ₁ (C)C(C)CC(C)CC ₁ C	174
C ₁₂ CC ₁ C ₃ CC ₂ C ₃	89.8	C ₁ CCCCCCCC ₁	138	C ₁ CCCCC ₁ (C)C(C)C	191.4
C ₁₂ CCC ₃ C(C ₁)C ₃ C ₂	79.2	C ₁ CCC ₂ CC ₁ CCC ₂	135	C ₁ (C)C(C)CCC(C)C ₁ C	183.5
C ₁₂ CCCC ₃ C ₁ C ₂₃	86.1	C ₁ CCC ₁ CC ₂ CCC ₂	156.7	C ₁ (C)(C)CC(C)CC(C ₁)C	181.5
C ₁₂₃ CC ₁ CCC ₂ C ₃	80.9	C ₁ CCCC ₂ CC ₁ CC ₂	154.8	C ₁ (C)(C)CCC(C)C(C ₁)C	185.5
C ₁₂ CC ₃ C ₄ C ₂ C ₃₄	137.5	C ₁ CCCC ₂ CCCC ₁₂	150.8	C ₁ (C)CCC(C)CC ₁ (C)C	170.5
C ₁₂ CC ₁ C ₃ C ₄ C ₂ C ₃₄	153	C ₁ CCCC ₂ CCC ₂ C ₁	150	C ₁ (C)(C)CCC(C)(C)CC ₁	193.6
CCCCCCCC	106	C ₁ (CCCCC ₁)C ₂ CC ₂	154.3	C ₁ (C)(C)CCCC(C ₁)(C)C	142
CC(C)CCCC	142	C ₁ CCCCC ₂ CC ₂ C ₁	139.5	C ₁ (C)(C)CCCC(C)C ₁ C	174.1
CCC(C)CCCC	149	C ₁ C(C)C ₂ CCC ₁ C ₂	144.8	C ₁ (C)(C)CCCCC ₁ (C)C	166.8
CCCC(C)CCC	136	C ₁ CC ₂ CCC(C ₁)C ₂	149.4	C ₁ CCCCC(C ₁)CCC	167.8
C(C)(C)CCC(C)C	125	C ₁ CC ₂ CC(C ₁)C(C)C ₂	152	C ₁ C(C)CCC(C)CC ₁ C	165.7
C(C)CC(C)CC	120.5	C ₁ C(C)CC ₂ C ₁ CCC ₂	136	C ₁ CC(C)C(C)CCC ₁ C	160
C(C)(C)CC(C)CC	103	C ₁₂₃ CCCC ₁ C ₂ CCC ₃	136.6	C ₁ CC(C)CC(C)C(C ₁)C	165.1
C(C)(C)C(C)CCC	111	C ₁₂₃ CCCC ₁ C ₂ CCCC ₃	145.1	C ₁ CCC(C)C(C)C(C ₁)C	168
C(C)C(C)C(C)CC	115	C ₁ CCC ₂ CCC(C ₁)C ₁₂	163.7	C ₁ CCC(C)(C)CC(C ₁)C	158.5
CC(C)(C)CCCC	102	C ₁₂₃ CC ₁ C ₂ CCCCC ₃	153.8	C ₁ CCC(C)C(C)C(C ₁)C	164
CCC(C)C(C)C	129	C ₁ CC ₂ (C)CCC ₁ CC ₂	151	C ₁ CCCCC(C ₁)CC	157
CC(C)C(C)C(C)C	137	C ₁ CC ₂ CC ₁ CC ₂ CC	157	C ₁ CC(C)CCCCC ₁ C	161.8
CCC(C)(C)CCC	136	C ₁₂ CCCC(C ₂)(C)CC ₁	150	C ₁ (C)CCCCC(C ₁)C	153
CC(C)(C)CC(C)C	133	C ₁ CC ₂ CC ₁ C(C)C ₂ C	168.2	C ₁ CCCCC(C ₁)C ₁ C	159.5
CCC(C)(C)CC	141	C ₁ CC ₂ (CC)CC ₁ CC ₂	175	C ₁ CCCCC(C ₁)(C)C	164
CC(C)(C)C(C)CC	125	C ₁ CC ₂ CC(C ₂)C ₁ (C)C	157.8	C ₁ CCCCC(C)CCC ₁	159.7
CC(C)C(C)(C)CC	128	C ₁ CCCC(C ₁)C ₁₂ CC ₂	174.5	C ₁ CCCCCCCC ₁	159
C(C)(C)(C)C(C)C	128	C ₁ CCCC ₂ (C)CC ₂ C ₁	150.8	C ₁₂ CCCCC ₂ CCCC ₁	155
C(CCCC)C ₁ CC ₁	130.5	C ₁ C ₂ CCC(C ₂)C ₁ (C)C	142.8	C ₁ CCC(C ₁)C ₂ CCCC ₂	160
CC(CCC)C ₁ CC ₁	125	C ₁ C(C)C ₂ (C)CC ₁ CC ₂	144	C ₁ CCCCC ₂ CCCC ₁₂	145
C ₁ (CC ₁)CCCC	117	C ₁ CCCCC ₁₂ CC ₂ C	142.4	C ₁₂ CCCC(C ₂)CC(C ₁)C	166
C ₁ CC ₁ CC(C)(C)C	138	C ₁ CC ₂ (C)CCC ₁ C ₂ C	134	C ₁₂ CCCC(C ₂)C(C ₁)C	160
C ₁ CC ₁ C(C)C(C)C	125	C ₁ CCC ₂ CC ₁ C ₂ (C)C	143	C ₁₂ CCCC(C ₁)CCC ₂ C	155.7
C ₁ (CCC)CC ₁ CC	115	C ₁ (C)(C)C ₂ CCCC ₁ C ₂	142.1	C ₁ C(C)CCC ₂ C ₁ CCC ₂	166
CC ₁ CC ₁ CC(C)C	91	C ₁ CCCCC ₂ CC ₁₂ CC	136	C ₁ CCCCC ₂ C ₁ CC(C)C ₂	162.4
C ₁ (C)(C)CC ₁ CCC	126.1	C ₁ (C)CCC ₁₂ CCC ₂ C	133.5	C ₁ CCC ₂ CCCC ₁ C ₂ C	159.5
C ₁ (C)(CC)CC ₁ CC	104	C ₁ CCCCC ₂ C ₁ C ₂ (C)C	140.5	C ₁ CCCCC ₂ (C ₁)CCC ₂	148.2
C ₁ (C)(C)CC ₁ C(C)C	105	C ₁ (C)CCCC ₂ C ₁ C ₂ (C)C	136	C ₁ (C)CCCCC ₂ C ₁ CCC ₂	167
C ₁ (C)(C)CC ₁ (C)CC	159	C ₁₂ (C)CCC(C ₂)C ₁ (C)C	133.8	C ₁ CCCCC ₂ (C ₁)CCC ₂	159.7
C ₁ (C)(C)C(C)C ₁ (C)C	128	C ₁ CC ₂ (C)CC ₂ (C ₁)CC	132.7	C ₁₂ CCCC(C ₂)CC ₁ CC	167
C ₁ (CCC ₁)CC(C)C	117.7	C ₁ (C)CC ₂ (C)CC ₂ (C ₁)C	140.6	C ₁ CCCCC ₂ CCC(C ₁)C ₁₂	161.2
C ₁ (CCC)CC(C ₁)C	124	C ₁₂ CC ₃ CC(C ₁)C(C ₂)C ₃	138	C ₁₂ (C)CCCC(C ₁)CCC ₂	157
C ₁ (CCC ₁)C(C)CC	106	C ₁ C ₂ CC ₃ CCC ₁ C ₃ C ₂	131.3	C ₁ CCC ₂ CCC(C)C ₁₂	147
C ₁ (CC)CCC ₁ CC	115.5	C ₁₂ CCC(C ₂)CC ₃ C ₁ C ₃	140.4	C ₁₂ CCC(C ₂)CC(C)C ₁ C	157
C ₁ (C)C(C)C(C)C ₁ C	108	C ₁ CC ₂ C ₃ C ₁ CC ₂ CC ₃	124	C ₁₂ CCC(C ₂)CC ₁ C	160
C ₁ (C)(C)CC(C ₁)(C)C	110	C ₁ CC ₂ (CC ₂)CC ₁₃ CC ₃	137.3	CC ₁ CC ₂ CC(C)CC ₂ C ₁	162
C ₁ CCCC ₁ CCC	105.9	C ₁ (CC ₁)(C ₂ CC ₂)C ₃ CC ₃	135.2	C ₁₂ CC(C)C(C ₂)CCC ₁ C	147.7
C ₁ CCCC ₁ C(C)C	108.9	C ₁ CC ₂ CCC ₁ C ₂₃ CC ₃	139	C ₁₂ (CC)CCC(C ₁)CC ₂	163
C ₁ C(C)CCCC ₁ CC	94.4	C ₁₂ C(C)C ₂ C ₃ CC ₁ CC ₃	126.5	C ₁₂ CCCC(C ₂)C(C)C ₁ C	152.8
C ₁ (C)CCCC ₁ CC	104.5	C ₁ CCC ₂ (CC ₂)C ₁₃ CC ₃	136.7	C ₁ CCCCC ₂ CCCC ₁₂ C	163
C ₁ (C)CC(C)CC ₁ C	100.5	C ₁ C ₂ CC ₃ C ₁ C ₃ (CC)C ₂	140.6	C ₁ CC(C)C ₂ CCC(C)C ₁₂	164
C ₁ CCCC ₁ (C)CC	120.1	C ₁ C ₂ C ₃ CC ₁ C(C)(C)C ₂₃	130.7	C ₁ CCC ₂ C ₁ C(C)CC ₂ C	147
C ₁ (C)C(C)CCC ₁ C	117.4	C ₁ C ₂ CC ₃ C ₁ C ₃ (C)C ₂ C	133.6	C ₁₂ CCC(C ₂)CCC ₁ (C)C	163.8
C ₁ C(C)CCCC ₁ (C)C	123	C ₁ C ₂ CC ₃ C ₁ (C)C ₃ (C)C ₂	145	C ₁₂₃ CCCC ₁ C ₃ CCCC ₂ C	164
C ₁ (C)CCCC ₁ (C)C	119	C ₁₂ CC(C ₃ C ₁ C ₃)C ₄ C ₂ C ₄	137.7	C ₁₂ CCC(C ₂)CC ₁ (C)C	164
C ₁ CCCC(C ₁)CC	114.5	C ₁₂ CCCC ₃ C ₂ C ₄ C ₁ C ₃₄	133.8	C ₁₂ (C)CCCC(C ₂)CC ₁ CC ₁	152
C ₁ CC(C)CCC ₁ C	86	CCCCCCCCCCCC	122.3	C ₁ CCC ₂ CC(C)CC ₁₂ C	158
C ₁ CC(C)CC(C ₁)C	131	CCC(C)CCCCCCCC	133	C ₁₂ (C)CCCC(C ₂)(C)CC ₁	155.7
C ₁ CCC(C)C(C ₁)C	126.4	CCCC(C)CCCCCCC	141.5	C ₁₂ CC(C)C(C ₂)CC ₁ (C)C	166
C ₁ CCCC(C ₁)(C)C	121	CCCC(C)CCCCC	140.2	C ₁ CCCCC ₂ CC ₁₂ CCC	147.9

Table 1. (Continued)

compd	bp	compd	bp	compd	bp
C ₁ CCCC(C)CC ₁	124.7	CC(C)CCCC(C)C	160	C ₁₂ CCC(C ₂)C(C)C ₁ (C)C	137
C ₁ CCCCCCC ₁	115	CCCC(C)CCCC	174	C ₁₂ (C)CCCC(C ₂)C(C)C ₁ C	158
C ₁ (CC ₁)CCC ₂ CC ₂	121.5	CCC(CC)CCCC	155	C ₁₂ CCC(CC ₁)C ₂ (C)C	160
C ₁₂ CCCC ₁ CCC ₂	117	CC(C)CCCC(C)CC	175.5	C ₁₂ (C)CCC(C ₂)C(C ₁)(C)C	158
C ₁ (CCC ₁)C ₂ CCC ₂	104.9	CCCC(CC)CCCC	219	C ₁₂ CC(CCC ₁)C ₂ (C)C	159
C ₁ CCC ₂ CCC ₂ C ₁	114	CC(C)CCC(C)CCC	189.5	C ₁₂ (C)CCC(C)(CC ₁)C ₂ C	148.7
C ₁ CCCC ₂ C ₁ C ₂	131.8	CCCC(CCC)CCC	158.3	C ₁ (C)CCC ₂ C ₁ C ₂ (C)C	166.3
C ₁ CC ₂ CC ₁ (C)C ₂	121.8	CC(C)CC(C)CCCC	192.7	C ₁₂ (C)CCC(CC ₁)C ₂ (C)C	157
C ₁ CCCC ₁₂ CCC ₂	122.3	CCC(C)CCC(C)CC	160	C ₁ CCCC ₁₂ CC ₂ (C)C	164
C ₁₂ CCC(CC ₁)C ₂ C	126.6	CC(C)C(C)CCCC	151	C ₁ CC(C)C ₂ CC ₁₂ (C)C	164
C ₁ (C)CCC ₂ C ₁ CC ₂	119.5	CCC(C)CCC(C)C	143.5	C ₁ CCC ₂ (C)CC ₁₂ CCC	153
C ₁₂₃ CC ₁ C ₂ CCCC ₃	134	CCCC(C)CC(C)CC	153	C ₁ (C)CC ₂ (C)CC ₁₂ C	165
C ₁₂ (C)CCC(C ₂)CC ₁	149	CCCCCCC(C)(C)C	191.5	C ₁ CCC ₂ (C)CC ₁₂ (C)C	169
C ₁ CCCC ₂ C ₁ C ₂ C	142	CC(C)CC(CC)CCC	190	C ₁₂ CC ₁ CCC(C ₂)C ₃ CC ₃	153
C ₁ CCCC ₂ CC ₁₂ C	202	CC(C)CC(C)CC(C)C	193	C ₁ CC ₂ C ₁ C ₃ CCC ₂ CC ₃	155
C ₁₂ CC ₁ CC(C ₂)(C)C	171	CCC(C)C(C)CCCC	182	C ₁₂ CCC(CC ₂)C ₃ C ₁ CCC ₃	170
C ₁ CC ₂ (C)CC ₁ (C)C ₂	125.7	CCC(C)CC(CC)CC	190	C ₁ (C)CC ₂ CCC ₁ C ₃ CC ₂₃	155.3
C ₁₂ CCCC ₁ C ₂ (C)C	117.6	CC(C)CCC(C)C(C)C	187	C ₁ C(C ₁ (C)C ₂ CC ₂)C ₃ CC ₃	162
C ₁₂ C(C ₂ (C)C)C ₁ (C)C	118.9	CC(C)C(CC)CCCC	178	C ₁ CCCC ₁₂ C ₃ CCCC ₂₃	174
C ₁₂ (C)CC ₁ (C)C ₂ (C)C	117.7	CCCC(C)C(C)CCC	173.5	C ₁ CCCC ₂ (CC ₂)C ₁₃ CC ₃	164
C ₁₂ CC ₁ CC ₃ CC ₃ C ₂	109.1	CCCC(CCC)C(C)C	189.9	C ₁₂ CCC(C ₂)C ₃ C ₁ C ₃	169.4
C ₁₂ CC ₁ CCC ₃ C ₂ C ₃	118.5	CC(C)(C)CCCC(C)C	185	C ₁ (C)(C)C ₂ C ₃ CC ₁ (C)CC ₂₃	159.3
C ₁₂ CCC(C ₂)C ₃ C ₁ C ₃	109.4	CCC(CC)C(C)CCC	182	C ₁ C ₂ CC ₃ C ₁ (C)C ₃ C ₂ (C)C	170.5
C ₁ CC ₂ C ₃ CCC ₂ C ₁₃	115.6	CC(C)C(C)CC(C)CC	183	C ₁₂ (C)C ₃ C ₂ CC(C ₃)C ₁ (C)C	168
C ₁ (C)C ₂ CC ₃ C ₁ C ₃ C ₂	117.7	CCC(C)C(C)CCC	183	C ₁₂ CC ₂ C ₃ CC ₁ C ₄ (CC ₄)C ₃	166
C ₁₂₃ CC ₁ C ₂ C ₄ (C ₃)CC ₄	106.8	CCC(C)(C)CCCC	182	C ₁ CC ₂ C ₃ CC ₄ (C ₃)C ₂₄ C ₁	188
C ₁ C ₂ CC ₃ C ₁ (C)C ₃ C ₂	115.6	CC(C)CC(C)C(C)CC	178	C ₁₂ CC ₁ C ₃ C ₄ C ₂ C ₃ C ₄₅	167
C ₁ CC ₂ (CC ₂)C ₁₃ CC ₃	113.5	CC(C)(C)CCC(C)CC	174		
C ₁₂ CC ₃ C ₄ CC ₁ C ₂ C ₃₄	112	CC(C)CC(CC)C(C)C	174.5		

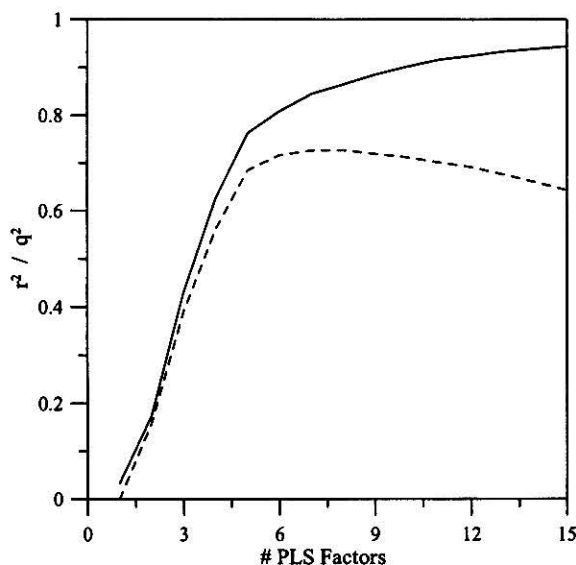


Figure 1. Evolution of r^2 (continuous line) and q^2 (dashed line) versus the number of PLS factors involved in the construction of the QSPR model for a set of 529 saturated hydrocarbons.

as the results provide the optimal balance between predictivity and descriptors applied, and a further increase does not lead to noticeable improved results and could lead to overfitted models. Thus, the final equation and statistical results for this molecular set are

$$\text{bp} = 3147.79f_1 + 15269.10f_2 + 34944.14f_3 + 43295.57f_4 + 47121.10f_5 + 28552.70f_6$$

$$r^2 = 0.808 \quad q^2 = 0.716 \quad s = 20.575$$

The results are also graphically presented by plotting the experimental bp versus the predicted ones from the PLS

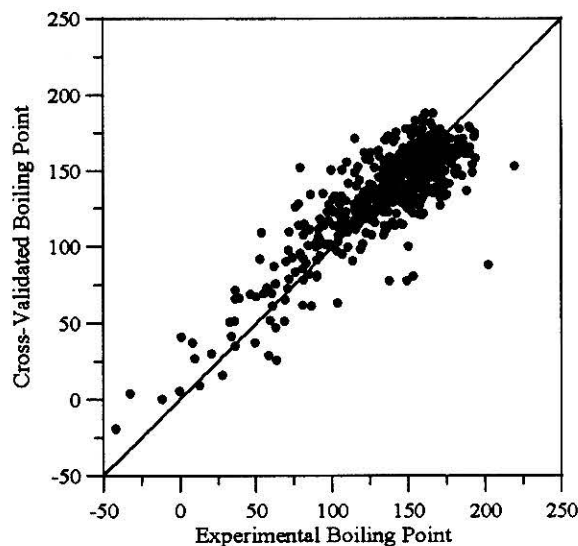


Figure 2. Experimental versus predicted boiling points for a set of 529 saturated hydrocarbons.

procedure, as shown in Figure 2. In addition, the random test was carried out permuting 100 times the bp values and reconstructing the QSPR model each time. The results are graphically compiled in Figure 3, where each point represents the values of r^2 and q^2 achieved in each permutation.

As seen from the statistical and graphical results, satisfactory results are obtained for this molecular set. Most of the compounds are correctly predicted, within a close margin, even few deviations are present, as displayed in Figure 2. The random test shows a clear separation between the original point (+) and the permuted ones (●). Among the permuted models, the highest values of r^2 and q^2 achieved are 0.475 and 0.029, respectively, whereas the average values

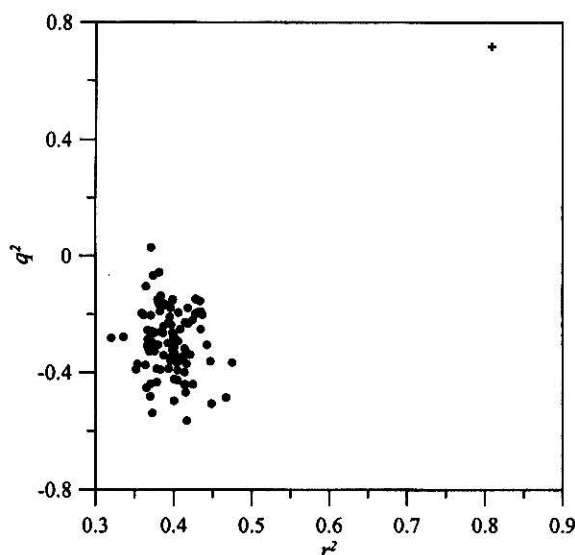


Figure 3. Random test results after permuting the boiling point values for a set of 529 saturated hydrocarbons 100 times. The real value is labeled with a cross.

are 0.395 and -0.293 . The results of the previously published work⁴⁰ are slightly better, yielding to $r^2 = 0.961$ using 6 topological indices but providing no information about the predictive capacity of the proposed models.

Gas Chromatographic Retention Times for 152 Diverse Chemical Compounds. The identification of single organic compounds in chemical mixtures can be made by means of comparing chromatographic peaks with samples of each compound. Due to the costs, or simply by unavailability, of pure samples, sometimes this comparison cannot be done. In this way, it is interesting to develop QSPR models to describe and to predict retention parameters from molecular structures. The present example deals with retention times (t_R), namely the mean time needed for a molecule to elute from a chromatographic column. A molecular set composed by 152 structurally different compounds has been studied together with their retention times, presented in Table 2, with the expressed procedure. The original 152×152 descriptor matrix was reduced to 152×135 and submitted to the PLS routine. The optimal model was found to be built up from five PLS factors, and the obtained results are

$$t_R = 75.83f_1 + 345.17f_2 + 869.53f_3 + 345.55f_4 + 492.72f_5$$

$$r^2 = 0.753 \quad q^2 = 0.541 \quad s = 1.657$$

As seen from the statistical results, an acceptable relationship with valuable predictive capacity is obtained. The predictive power is also graphically represented in Figure 4, where experimental versus predicted t_R values are plotted.

As seen in Figure 4, all compounds are accurately predicted in a narrow margin with the exception of a few molecules, which are mostly overestimated. The results of the random test for this molecular set are, similarly to the previous examples, very satisfactory, yielding to average r^2 and q^2 values of 0.229 and -0.234 , whereas the maximum values achieved were 0.373 and 0.180, respectively. The previous work⁴¹ obtained better results using a large set of 296 classical, topological, and quantum parameters computed

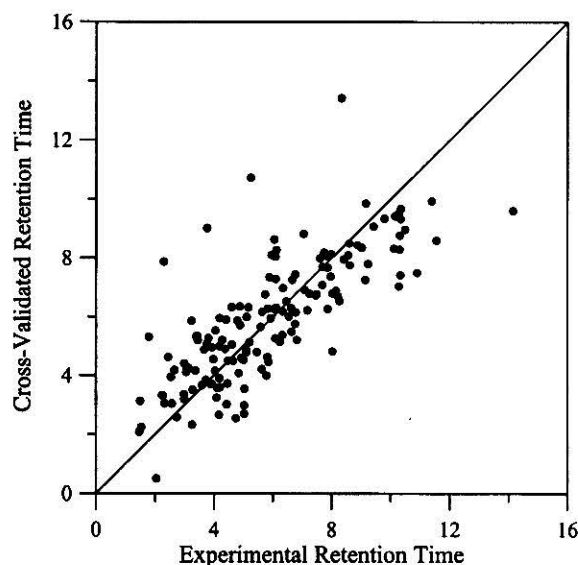


Figure 4. Experimental versus predicted retention times for a set of 152 compounds.

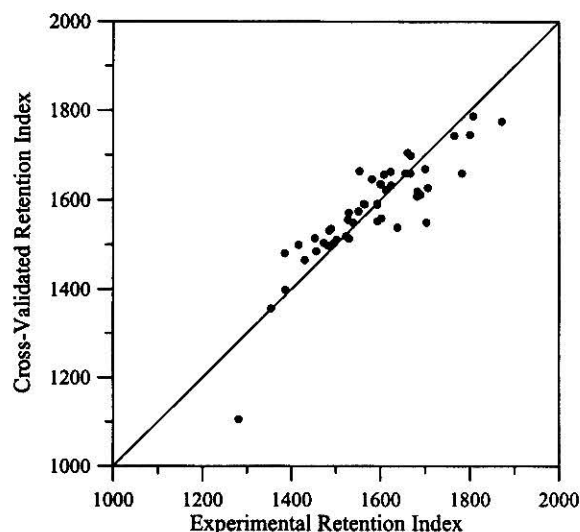


Figure 5. Experimental versus predicted retention indices for a set of 50 phenol derivatives.

with the CODESSA⁴⁵ software. Their models were calculated using linear and nonlinear multilinear regressions yielding to cross-validation regression coefficients (r^2_{cv}) of 0.961 and 0.967, respectively. Excellent results in the random tests were also reported.

Gas-Liquid Chromatography Retention Indices for 50 Phenol Derivatives. This example, similarly to the previous one, deals with chromatography. However, instead of t_R , Kováts retention indices (RI)⁴⁶ are studied. The analyzed molecular set consists of 50 phenol derivatives. The molecular structures, as well as the RI, are presented in Table 3. The original 50×50 descriptor matrix was reduced to 50×38 after the redundancies elimination, and the optimal QSPR model, built up from 6 PLS factors, results in

$$RI = 12938.48f_1 + 31183.31f_2 + 54224.34f_3 + 60530.94f_4 + 25264.86f_5 + 45970.82f_6$$

$$r^2 = 0.897 \quad q^2 = 0.762 \quad s = 59.294$$

Table 2. SMILES Notation^{47,48} and Retention Times of 152 Diverse Chemical Compounds⁴¹

compd	retention time	compd	retention time	compd	retention time
CCCCCCCCCCCC	6.34	c1cm3cc(c1)Oc2cm3ccc2	7.74	c1cm3cc(O)c1OC(C)C	5.95
CCCCCCCCCCCC CCC	8.56	c1cm3cc(c1)C#N	4.1	c1cnccc1C#N	3.89
CCCCCCCC=C	2.53	c1cm3cc(c1)Sc2cm3ccc2	9.02	c1(CC)ccccn1	3.92
CCCCCCCCC=C	4.45	c12CCCCc1c(O)ccc2	8.02	c1c(C)cc(C)nc1C	4.37
C1CCC(O)CC1	3.27	c1cm3ccc1C(=O)C2CCCCC2	9.14	c1c(C)cc(C)nc1N	5.85
C1CCC(=O)CC1	3.25	c1cm3cc(c1)NC	4.94	c1cm3cc(c1)C=CC=O	6.63
CCCCCCCCCN	4.88	c12cm3cc(O)c2cm3c(c1)O	10.33	c12cm3ccc1OC(=O)CC2	7.46
CCCCCCCCCCCN	6.62	C1=CC=CS1	1.49	c1cm3cc(c1)C=CC(=O)O	7.86
CCCCCCC#N	3.06	C1=CC=CN1	2.04	c12cm3ccc1OC(=O)C=C2	7.87
C1CCC(C1)=O	2.26	c1(CN)cccc1	4.47	c1(CC)ccccc1	3.97
CCCCCCCCSCCCCCC	10.9	c1(C)ccc(cc1)n(o)o	6.24	c1(CCN)ccccc1	6.08
CCOC(=O)OCC	2.29	c1cm3c(cc1)C(=O)Oc2cm3ccc2	9.42	c1CCN(CC1)C(=O)OCC	6.11
C1CCC(C1)O	2.32	c1(C)ccc(N)cc1	5.02	c1cm3nc(c1)CC	3.44
CCCCCCCCCO	5.1	c12cm3ccc1Cc2c2cm3cc3	10.28	c1cm3nc(c1)CCO	5.63
CCCCCCCCCCCO	6.77	c1(O)ccc(cc1)CC	5.81	c1cm3c(cc1)C(=O)C(=O)c2cm3ccc2	10.31
CCCCCCCCCCCCCO	8.26	c1(N)cccc1	4.08	c1(ccccc1)C(=O)OCC	6.04
C1CCCC(N)CC1	4.4	c1(N)cccc1C	5	c1cm3nc(c1)C(C)=O	4.6
C1CCC(N)CC1	3.05	c1cm3ccc1OCCCC	6.08	c1cncc(c1)C(C)=O	5.2
CCCNCCCC	4.18	c1cm3ccc1Br	3.75	c1cm3c2CCOCc2c1	6.04
CCCCCCCCCCCCCN	8.14	c1cm3ccc1I	4.74	C1(CN)=CC=CO1	2.73
CCCCCCCCCCCCCBr	7.42	c1cm3ccc1C=O	3.92	C1=CC=CN(C)C1=O	5.78
N1(C)CCNCC1	2.98	c1cm3ccc1CO	4.64	c1cm3c2c(c1)nc3cm3ccc23	10.49
C(O)CCO	2.65	c1cm3(C)ccc1C=O	5.1	c1cm3c(CC)c(c1)O	5.58
C(O)CCCO	3.72	C1(=O)C=CC(=O)c2c1cm3ccc2	7.68	c1cm3c(c1)C=O	4.17
C1CCC(C1)CC1	3.36	c12cm3ccc1C=CN2	6.81	c1cm3c2cm33cm3ccc3nc2c1	10.35
C1CCCC(=O)CC1	4.42	c1(ccccc1)n(o)o	5.09	c1cm3c(cc1)Nc2cm3ccc2	9.16
C1CCC(CC1)C(=O)O	5.62	c12cm3ccc1N=CS2	6.33	c1(C)ccc(O)cc1	5.02
c1cm3cc2cm3ccc12	6.09	c12cm3ccc1NC=N2	8.02	c1(C)cccc1	2.56
c1cm3cc(c1)CCc2cm3ccc2	7.6	c12cm3ccc1cm3c(n2)C	7.04	c1cm3cc(c1)Cc2cm3ccc2	7.96
c1cm3cc(c1)CCC	4.04	c1(C)ncccc1C	3.8	c1cm3cc(c1)CCc2cm3ccc2	8.61
c1cm3cc(c1)CC=C	3.39	c1(C)cccc(n1)C	3.23	c1cm3c(cc1)CC(C)(C)Cc2cm3ccc2	9.79
c1cm3cc2CCc12	4.79	c1(ccccc1)SSc2cm3ccc2	10.15	c1cm3(C)ccc1Cl	4.03
c1cm3cc(c1)C(C)C	3.75	c1(C)cccc(c1)C	3.66	c1cm3cc(c1)CBr	5.24
c1cm3cc2c1cm3c3cm3ccc23	10.27	c1(C)cccc1O	4.84	CC(C)C(=O)C(C)C	2.43
c1cm3c2CCCCc2c1	5.92	c1(OC)cc(OC)cc(c1)OC	7.68	c1cm3ccc1	1.46
c1cm3c2CCCNc2c1	7.17	c1(O)ccc(cc1)C(C)C	6.32	c1cm3ccc1C	2.23
c1cm3c2cm3cnc2c1	6.45	c1cm3cc(c1)C(=O)c2cm3ccc2	9.23	c1cm3cc(C)c1C	3.43
c1c(C)cc(C)cc1C	4.18	CC(=O)c1cm3ccc1	4.93	C(CC)CCCCCCC	4.59
c1cm3c2CNCCc2c1	6.75	CCCCCCCCC(=O)O	6.76	C(CC)CCCCCCCCCCCCCCCC	11.55
c1cm3c2cm33cm3ccc3cm32c1	10.34	c1cm3ccc1OC(=O)C	5.73	C1CCCCC1	1.53
c1cm3c(cc1)C(=C)C	4.27	c1(O)cccc1C(C)C	6.11	CCCCCCC	1.78
c1cm3c2c(c1)Oc3cm3ccc23	8.6	C(C#N)c1cm3ccc1	5.45	c1cm3ccc1CC	3.12
c1cm3c2c(c1)Sc3cm3ccc23	10.11	C(C(=O)O)c1cm3ccc1	6.54	c1cm3ccc1C(=O)CC	5.82
c1(ccccc2cm3ccc12)c3cm3cc4cm3ccc34	14.14	c1cm3cc(c1)CCCC(=O)O	7.86	c1cm3ccc1OC	3.59
c1cm3c2CCCC(=O)c12	7.48	c1cm3cc(c1)SC	5.16	c1(O)ccc2cm3cc(O)c2c1	10.3
c1cm3cc2cm3cc(Br)c12	8.33	c1cm3cc(c1)CSC	5.88	c1cm3c2cm3ncc2c1	6.65
c1cm3cc2cm3cc(O)c12	8.41	c1nccc(c1)C	2.98	c1cm3cnc1c2ncccc2	7.98
c1cm3ccc1C2CCCCC2	7.24	c1ncccc1C	2.97	C1(C)cc(C)cc(C)c1C(O)=O	7.82
c1cm3c2cm3cc(Cc3cm3ccc3)c2c1	11.39	c1c(C)cccc1O	5.03	c1cm3ccc1S(C)=O	7.07
c1cm3cc(c1)O	4.16	c1(C)c(C)c(C)c(C)c(C)c1C	8.21	c1cm3ccc1CCC=C	4.87
c1cm3cc(c1)OCc2cm3ccc2	8.87	n1cm3cc(c1)C#N	4.19		

As seen from the previous statistical results, a satisfactory QSPR model is found with a remarkable predictive capacity, as judged from the q^2 value achieved. A clearer picture of the obtained results can be seen by browsing Figure 5, where the experimental RI are plotted versus the cross-validated ones. As seen in Figure 5, all RI are correctly predicted except the one referred to the unsubstituted benzene, which is underestimated. In addition, the random test is satisfactorily surpassed, yielding to averaged values of 0.437 and -0.352 for r^2 and q^2 , respectively, whereas the maximum values were 0.675 and 0.377.

The previous work⁴⁶ achieved valuable results with a few number of molecular descriptors, consisting of vertex- and edge-weighted molecular graphs. Results ranging from 0.978 to 0.982 for r^2_{cv} values are obtained with up to 4 topological descriptors.

CONCLUSIONS

A number of molecular sets have been tested for correlation search with molecular properties, such as boiling points and chromatographic parameters. Satisfactory correlations are obtained using a small number of descriptors for molecular sets making use of quantum similarity measures.

Even if other methodologies may provide better results, it must be emphasized that the procedure used along this work, as well as the generation of molecular descriptors, has been kept unmanipulated. In other words, the exposed QSPR protocol consists of a general methodology made of unbiased and universal MQSM descriptors able to characterize different molecular properties without introducing further information. Additional refinements may be applied to the procedure to improve the results according to each

Table 3. Smiles Notation^{47,48} and Retention Indices of 50 Phenol Derivatives⁴²

compd	retention index	compd	retention index
c1(O)ccccc1	1281	c1(O)c(C)ccc(C)c1C	1551
c1(O)ccccc1C	1354	c1(O)cc(C)c(C)cc1C	1593
c1(O)cccc(c1)C	1386	c1(O)cc(C)c(C)c(c1)C	1667
c1(O)ccc(C)cc1	1385	c1(O)ccc(cc1)CC(C)C	1612
c1(O)ccccc1CC	1430	c1(O)ccccc1CCCC	1600
c1(O)cccc(c1)CC	1483	c1(O)cccc(c1)CCCC	1668
c1(O)ccc(cc1)CC	1473	c1(O)ccc(cc1)CCCC	1661
c1(O)cccc(C)c1C	1495	c1(O)ccc(cc1C)CCC	1623
c1(O)ccc(C)cc1C	1456	c1(O)c(C)ccccc1CCC	1553
c1(O)cc(C)ccc1C	1453	c1(O)cc(C)ccc1CCC	1602
c1(O)c(C)ccccc1C	1416	c1(O)ccc(C)cc1CCC	1593
c1(O)cc(C)cc(c1)C	1489	c1(O)ccc(CC)cc1CC	1602
c1(O)ccc(C)c(c1)C	1530	c1(O)cc(CC)ccc1CC	1624
c1(O)ccc(cc1C)C(C)C	1527	c1(O)ccc(CC)c(c1)CC	1682
c1(O)ccccc1CCC	1502	c1(O)cc(C)c(C)c(c1)C	1782
c1(O)cccc(c1)CCC	1565	c1(O)c(C)ccc(C)c(c1)C	1690
c1(O)ccc(cc1)CCC	1563	c1(O)c(C)c(C)c(c1)C	1683
c1(O)ccc(C)cc1CC	1523	c1(O)cc(C)c(C)cc1CC	1656
c1(O)ccc(C)ccc1CC	1529	c1(O)ccccc1CCCCC	1700
c1(O)c(C)ccc1CC	1485	c1(O)ccc(cc1)CCCCC	1765
c1(O)ccc(C)cc(c1)CC	1581	c1(O)ccc(cc1)CC(C)(C)C	1703
c1(O)ccc(CC)cc1C	1539	c1(O)c(CC)ccccc1CCC	1706
c1(O)ccc(CC)c(c1)C	1608	c1(O)ccccc1CCCCC	1800
c1(O)ccc(C)c(C)c1C	1638	c1(O)ccc(cc1)CCCCC	1871
c1(O)ccc(C)cc(C)c1C	1593	c1(O)cc(CCCC)ccc1CC	1807

molecular set under study; however, the exposed protocol may constitute an excellent starting point for subsequent research.

ACKNOWLEDGMENT

This research has been partially supported by Project No. SAF2000-0223-C03-01 from the Spanish Ministerio de Ciencia y Tecnología. X.G. benefits from a predoctoral fellowship from the University of Girona. Financial support from the Fundació Maria Francisca de Roviralta is also acknowledged.

REFERENCES AND NOTES

- Quoted in: Borman, S. New QSAR Techniques Eyed for Environmental Assessments. *Chem. Eng. News* **1990**, 68, 20–23.
- Hammett, L. P. The Effect of Structures upon the Reactions of Organic Compounds. Benzene Derivatives. *J. Am. Chem. Soc.* **1937**, 59, 96–103.
- Jurs, P. C. Quantitative Structure–Property Relationships (QSPR). In *Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, H. F., III; Schreiner, P. R., Eds.; John Wiley and Sons Ltd.: Chichester, U.K., 1998; Vol. 4, pp 2320–2330.
- Waterbeemd, H. v. d., Ed. *Structure–Property Correlations in Drug Research*; Academic R. G. Landes Co.: Austin, TX, 1996.
- Kubinyi, H. Quantitative Structure–Activity Relationships in Drug Design. In *Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, H. F., III; Schreiner, P. R., Eds.; John Wiley and Sons Ltd.: Chichester, U.K., 1998; Vol. 4, pp 2309–2319.
- Carbó, R.; Arnau, J.; Leyda, L. How similar is a molecule to another? An electron density measure of similarity between two molecular structures. *Int. J. Quantum Chem.* **1980**, 17, 1185–1189.
- Carbó-Dorca, R.; Besalú, E. A general survey of molecular quantum similarity. *THEOCHEM* **1998**, 451, 11–23.
- Carbó-Dorca, R.; Amat, L.; Besalú, E.; Lobato, M. Quantum Similarity. In *Advances in Molecular Similarity*; Carbó-Dorca, R., Mezey, P. G., Eds.; JAI Press: Greenwich, CT, 1998; Vol. 2, pp 1–42.
- Carbó-Dorca, R. Fuzzy sets and Boolean tagged sets; vector semispaces and convex sets; quantum similarity measures and ASA density functions; diagonal vector spaces and quantum chemistry. In *Advances in molecular similarity*; Carbó-Dorca, R., Mezey, P. G., Eds.; JAI Press: Greenwich, CT, 1998; Vol. 2, pp 43–72.
- Besalú, E.; Carbó, R.; Mestres, J.; Solà, M. Foundations and recent developments on molecular quantum similarity. *Top. Curr. Chem.* **1995**, 173, 31–62.
- Carbó, R.; Besalú, E. Theoretical foundations of quantum molecular similarity. In *Molecular similarity and reactivity: from quantum chemical to phenomenological approaches*; Carbo, R., Ed.; Kluwer: Amsterdam, 1995; pp 3–30.
- Carbó-Dorca R.; Besalú, E.; Amat, L.; Fradera, X. Quantum molecular similarity measures: concepts, definitions and applications to quantitative structure-properties relationships. In *Advances in molecular similarity*; Carbó-Dorca, R., Mezey, P. G., Eds.; JAI Press: Greenwich, CT, 1996; Vol. 1, pp 1–42.
- Carbó, R.; Besalú, E.; Amat, L.; Fradera, X. Quantum molecular similarity measures (QMSM) as a natural way leading towards a theoretical foundation of quantitative structure-properties relationships. *J. Math. Chem.* **1995**, 18, 237–246.
- Fradera, X.; Amat, L.; Besalú, E.; Carbó-Dorca, R. Application of molecular quantum similarity to QSAR. *Quant. Struct.-Act. Relat.* **1997**, 16, 25–32.
- Lobato, M.; Amat, L.; Carbó-Dorca, R. Structure–activity relationships of a steroid family using quantum similarity measures and topological quantum similarity indices. *Quant. Struct.-Act. Relat.* **1997**, 16, 465–472.
- Amat, L.; Robert, D.; Besalú, E.; Carbó-Dorca, R. Molecular quantum similarity measures tuned 3D QSAR: an antitumoral family validation study. *J. Chem. Inf. Comput. Sci.* **1998**, 38, 624–631.
- Robert, D.; Amat, L.; Carbó-Dorca, R. Three-dimensional quantitative structure–activity relationships from tuned molecular quantum similarity measures. Prediction of the corticosteroid-binding globulin binding affinity for a steroid family. *J. Chem. Inf. Comput. Sci.* **1999**, 39, 333–344.
- Robert, D.; Gironés, X.; Carbó-Dorca, R. Facet diagrams for quantum similarity data. *J. Comput.-Aided Mol. Des.* **1999**, 13, 597–610.
- Gironés, X.; Amat, L.; Robert, D.; Carbó-Dorca, R. Use of electron–electron repulsion energy as a molecular descriptor in QSAR and QSPR studies. *J. Comput.-Aided Mol. Des.* **2000**, 14, 477–485.
- Robert, D.; Carbó-Dorca, R., Aromatic compounds aquatic toxicity QSAR using quantum similarity measures. *SAR QSAR Environ. Res.* **1999**, 10, 401–422.
- Gironés, X.; Amat, L.; Carbó-Dorca, R. Using molecular quantum similarity measures as descriptors in quantitative structure-toxicity relationships. *SAR QSAR Environ. Res.* **1999**, 10, 545–556.
- Dean, P. M. Molecular Similarity. In *3D QSAR in Drug Design: Theory, Methods and Applications*; Kubinyi, H., Ed.; ESCOM Science Publishers BV.: Leiden, The Netherlands, 1993; pp 150–172.
- Richards, W. G., Molecular Similarity and Dissimilarity. In *Modeling of Biomolecular Structures and Mechanisms*; Pullman, A., Jortner, J., Pullman, B., Eds.; Kluwer Academic Publishers: Dordrecht, The Netherlands, 1995; pp 365–369.
- Gironés, X.; Robert, D.; Carbó-Dorca, R. TGSA: a molecular superposition program based on Topo-Geometrical Considerations. *J. Comput. Chem.* **2001**, 22, 255–263.
- Carbó-Dorca, R. Stochastic Transformation of Quantum Similarity Matrixes and Their Use in Quantum QSAR (QQSAR) Models. *Int. J. Quantum Chem.* **2000**, 79, 163–177.
- Constants, P.; Carbó, R. Atomic shell approximation: electron density fitting algorithm restricting coefficients to positive values. *J. Chem. Inf. Comput. Sci.* **1995**, 35, 1046–1053.
- Amat, L.; Carbó-Dorca, R. Quantum similarity measures under atomic shell approximation: first-order density fitting using elementary Jacobi rotations. *J. Comput. Chem.* **1997**, 18, 2023–2039.
- Amat, L.; Carbó-Dorca, R. Fitted electronic density functions from H to Rn for use in quantum similarity measures: cis-diamminedichloro-platinum(II) complex as an application example. *J. Comput. Chem.* **1999**, 20, 911–920.
- AMPAC 6.55; Semichem: Shawnee, KS, 1999.
- Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. AM1: A New General Purpose Quantum Chemical Molecular Model. *J. Am. Chem. Soc.* **1985**, 107, 3902–3909.
- Höskuldsson A. *Prediction Methods in Science and Technology*; Thor: Copenhagen, 1996.
- Tenenhaus, M. *Regression de PLS*; Editions Technip: Paris, 1997.
- Wold, S.; Johansson, E.; Cocchi, M. PLS–Partial Least-Squares Projections to Latent Structures. In *3D QSAR in Drug Design, Theory, Methods and Applications*; Kubinyi, H., Ed.; ESCOM Science: Leiden, The Netherlands, 1993; pp 523–550.
- Wold, S. PLS for multivariate linear modelling. In *Methods and Principles in Medicinal Chemistry. Vol. 2: Chemometric Methods in Molecular Design*; van de Waterbeemd, H., Ed.; VCH: Weinheim, Germany, 1995; pp 195–218.

- (35) Cramer, R. D., III; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). 1. Effect on Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959–5967.
- (36) Wold, S.; Sjöström, M.; Eriksson, L. Partial Least Squares Projections to Latent Structures (PLS) in Chemistry. In *Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, H. F., III, Screiner, P. R., Eds.; John Wiley and Sons Ltd.: Chichester, U.K., 1994; Vol. 4, pp 2006–2021.
- (37) Geladi, P.; Kowalski, B. R. Partial Least-Squares Regression: A Tutorial. *Anal. Chim. Acta* **1986**, *185*, 1–17.
- (38) Montgomery, D. C.; Peck, E. A. *Introduction to linear regression analysis*; Wiley: New York, 1992.
- (39) Wold, S. Cross-validatory estimation of a number of components in factor and principal component models. *Technometrics* **1978**, *20*, 397–405.
- (40) Rücker, G.; Rücker, C. On Topological Indices, Boiling Points, and Cycloalkanes. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 788–802.
- (41) Lucic, B.; Trinajstić, N.; Sild, S.; Karelson, M.; Katritzky, A. R. A New Efficient Approach for Variable Selection Based on Multi-regression: Prediction of Gas Chromatographic Retention Times and Response Factors. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 610–621.
- (42) Ivanciuc, O.; Ivanciuc, T.; Cabrol-Bass, D.; Balaban, A. T. Comparison of Weighting Schemes for Molecular Graph Descriptors: Application in Quantitative Structure-Retention Relationship Models for Alkylphenols in Gas–Liquid Chromatography. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 732–743.
- (43) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. Normal Boiling Points for Organic Compounds: Correlation and Prediction by a Quantitative Structure–Property Relationship. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 28–41.
- (44) Ivanciuc, T.; Balaban, A. T. Quantitative Structure–Property Relationship Study of Normal Boiling Points for Halogen-/Oxygen-/Sulphur-Containing Organic Compounds Using the CODESSA Program. *Tetrahedron* **1998**, *54*, 9129–9142.
- (45) Katritzky, A. R.; Ignatchenko, E. S.; Barcock, R. A.; Lobanov, V. S.; Karelson, M. Prediction of Gas Chromatographic Retention Times and Response Factors Using a General Quantitative Structure–Property Relationship Treatment. *Anal. Chem.* **1994**, *66*, 1799–1807.
- (46) Buryan, P.; Nabivach, V. M.; Dimitrikov, V. P. Structure-Retention Correlations of Isomeric Alkylphenols in Gas–Liquid Chromatography. *J. Chromatogr.* **1990**, *509*, 3–14.
- (47) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.
- (48) SMILES stands for simplified molecular input line entry system. More information can be found at the following: <http://esc.syrres.com/interkow/docsmile.html>.

CI0103370

SOBRE LES RELACIONS LINEALS D'ENERGIA LLIURE: MESURES DE SEMBLANÇA MOLECULAR QUÀNTICA SOBRE FUNCIONS DE DENSITAT ELECTRONICA MODIFICADES

Xavier Gironés i Ramon Carbó-Dorca*

Institut de Química Computacional, Universitat de Girona, Catalunya, 17071 Espanya

* Autor per a correspondència. A/e: quantum@iqc.udg.es. Telèfon: 972 41 83 67 Fax: 972 41 83 56

RESUM

En aquest treball es presenta l'ús de funcions de densitat electrònica de Forats de Fermi per incrementar el paper que pren una regió molecular concreta, considerada com a responsable de la reactivitat molecular, tot i mantenir la mida de la funció de densitat original. Aquestes densitats s'utilitzen per fer mesures d'autosemblança molecular quàntica i es presenten com una alternativa a l'ús de fragments moleculars aïllats en estudis de relació entre estructura i propietat.

El treball es complementa amb un exemple pràctic, on es correlaciona l'autosemblança molecular a partir de densitats modificades amb l'energia d'una reacció isodèsmica.

RESUMEN

En este trabajo se presenta el uso de funciones de densidad electrónica de Agujero de Fermi para incrementar el papel que toma una región molecular concreta, considerada como responsable de la reactividad molecular, manteniendo el tamaño de la función de densidad original. Estas densidades se utilizan para realizar medidas de autosemejanza molecular cuántica, presentándose como una alternativa al uso de fragmentos moleculares aislados en estudios de relación entre estructura y propiedad.

El trabajo se complementa con un ejemplo práctico, donde se relaciona la autosemejanza molecular a partir de densidades moleculares con la energía de una reacción isodésmica.

ABSTRACT

In this work, we present the use of electronic Fermi Hole density functions in order to increase the role of a precise molecular region, considered responsible for the molecular reactivity, maintaining the size of the original density function. These modified densities are used to calculate molecular quantum self-similarity measures. This option is presented as an alternative to the use of isolated molecular fragments in structure-property relation studies.

This work is complemented with a practical example, where self-similarity values from modified densities are related to the energy of an isodesmic reaction.

INTRODUCCIÓ

Al final de la dècada dels anys trenta, Hammett [1,2] va realitzar estudis sobre la ionització en sèries d'àcids benzènics substituïts. Observant aquesta reactivitat i el tipus de substitució present a l'anell, Hammett inferí un paràmetre capaç de descriu-

re l'acidesa del benzoàcid a partir del substituent present i de la seva posició. Aquest paràmetre, conegut com a sigma de Hammett, σ , es deriva a partir de la reactivitat de la molècula amb el substituent concret i de la reactivitat de l'estructura no substituïda. Si s'estudia un sistema molecular amb diferents substitucions, se'n pot relacionar l'acidesa amb la següent equació:

$$\log \frac{K}{K_0} = \rho \sigma, \quad (1)$$

on K i K_0 són respectivament les constants d'acidesa de les molècules substituïdes i de referència, ρ és una constant de proporcionalitat i σ és el descriptor dels substituents. El paràmetre ρ pertany a un equilibri en concret i relaciona l'efecte dels substituents amb l'efecte que tenen en l'equilibri de la molècula de referència (normalment l'àcid benzoic). Llavors, si l'efecte de les substitucions és proporcionalment major que en la reacció original, s'obté: $\rho > 1$; en cas contrari, $\rho < 1$. La magnitud de σ dona la força relativa respecte de l'acceptació o la donació electrònica dels substituents. El paràmetre σ és positiu si els substituents són electroacceptors, i negatiu si són electrodonadors. A causa de la natura del treball de Hammett, amb ell s'inicià el camp dels *efectes electrònics* sobre l'activitat molecular.

Aquestes relacions desenvolupades per Hammett reberen el nom de *relacions lineals d'energia lliure* (de l'anglès *LFER, linear free energy relationships*) per la seva semblança amb l'equació termodinàmica que relaciona l'energia lliure amb el logaritme neperià de la constant d'equilibri. També són anomenades *relacions extratermodinàmiques* perquè es poden formular a partir de paràmetres termodinàmics (tot i que cap principi n'estableix la veritable natura).

Aquest estudi de Hammett fou el primer a *partionar* la molècula i a explicar-ne l'activitat a partir de tan sols fragments seus, en lloc de referir-la a la totalitat de l'estructura. Altres contribucions d'aquest estil, on es consideren les substitucions, són degudes al treball de Taft sobre els efectes estèrics [3] o els estudis de camp i ressonància de Swain [4], entre d'altres [5-13].

Dins del marc de les mesures de semblança molecular quàntica (MQSM, de l'anglès *molecular quantum similarity measures*), la qual dona compte de la similitud entre dues estructures a partir de les seves respectives distribucions electròniques, s'han dut a terme diversos estudis [14,15] on s'enfoca la mesura sobre un fragment en concret de les molècules en lloc de la molècula sencera, fent així una partió de la *funció de densitat electrònica* molecular. El present treball pretén mostrar una metodologia diferent de partionament molecular i de realització de MQSM sobre aquests fragments resultants a partir de la modificació de la densitat electrònica per la inclusió de forats de fermi en la regió molecular d'interès.

METODOLOGIA

Mesures de semblança molecular quàntica

Les MQSM [16-30] donen una mesura quantitativa de la semblança entre dues molècules a partir de les seves distribucions electròniques, les quals matemàticament es representen per funcions de densitat de primer ordre. D'aquesta manera, les MQSM s'expressen matemàticament com a:

$$Z_{AB}(\Omega) = \iint \rho_A(\mathbf{r}_1) \Omega(\mathbf{r}_1, \mathbf{r}_2) \rho_B(\mathbf{r}_2) d\mathbf{r}_1 d\mathbf{r}_2, \quad (2)$$

on $\rho_A(\mathbf{r}_1)$ i $\rho_B(\mathbf{r}_2)$ són les funcions de densitat de les molècules A i B; $\Omega(\mathbf{r}_1, \mathbf{r}_2)$ és un operador definit positiu, i $Z_{AB}(\Omega)$ és la mesura de semblança final, ponderada per l'operador escollit.

L'ús de la funció de densitat de la totalitat de la molècula pot portar a valors sobredimensionats quan es troben presents a l'estructura àtoms pesats, tipus brom o iode. Aquests elements constitueixen *pous de densitat*, els quals condensen la major part de densitat electrònica, donant com a resultat magnituds molt elevades i la inclusió de soroll de fons dins la mesura final, ja que l'estructura bàsica estudiada gairebé no contribueix a la mesura. Per pal·liar aquest efecte contraproduent, les mesures de semblança de fragments tan sols consideren una part de la molècula, redefinint la funció de densitat, on $D_{\mu\nu}$ és la matriu densitat i $\{\chi_\mu\}$ el conjunt de funcions de base, de

$$\rho(\mathbf{r}) = \sum_{\mu} \sum_{\nu} D_{\mu\nu} \chi_{\mu}^*(\mathbf{r}) \chi_{\nu}(\mathbf{r}) \quad (3)$$

a l'expressió

$$\rho^F(\mathbf{r}) = \sum_{\mu \in F} \sum_{\nu \in M} D_{\mu\nu} \chi_{\mu}^*(\mathbf{r}) \chi_{\nu}(\mathbf{r}), \quad (4)$$

on F representa el fragment d'interès i M la molècula sencera. L'expressió (4) permet rescriure la funció de densitat (3) de forma natural:

$$\rho(\mathbf{r}) = \sum_{F \in M} \rho^F(\mathbf{r}). \quad (5)$$

D'aquesta manera, i emprant aquesta formulació a l'equació (2), es poden extreure valors d'*autosemblança molecular quàntica* (MQS-SM, de l'anglès *molecular*

quantum self-similarity measures) referits només al fragment que s'estudia, i donant així compte de la concentració de càrrega en aquesta regió. D'aquesta manera es pot estudiar l'evolució del valor del paràmetre prenent com a base la inclusió de diferents substitucions en una estructura mare.

Generació de densitats de Forats de Fermi

Tot i que aquest mètode de partionament molecular proporciona una mesura restringida de la regió sotmesa a estudi, és d'alguna manera artificial si es té en compte que la funció de densitat és quelcom difús, d'acord amb els principis de la mecànica quàntica. Un possible procediment per continuar conservant les propietats de la funció de densitat consistiria a mantenir-ne la dimensió, però magnificant-ne la regió d'estudi. D'aquesta manera s'obtingria una nova *funció de densitat modificada* de tota la molècula, la qual estaria definida majoritàriament sobre aquells àtoms que es considerin responsables de l'activitat d'interès. Aquesta modificació es pot realitzar per la via de l'ús de funcions densitat de Forats de Fermi mitjanats en l'espai [31,32], formulats d'acord amb la següent equació:

$$h_{\theta}(\mathbf{r}_1) = N_{\theta} \rho(\mathbf{r}_1) - 2 \int_{\theta} \rho(\mathbf{r}_1, \mathbf{r}_2) d\mathbf{r}_2 \quad (6)$$

on $h_{\theta}(\mathbf{r}_1)$ és la nova densitat modificada per la presència d'un electró mitjanat en la regió definida pel volum θ , $\rho(\mathbf{r}_1)$ és la ja descrita funció de densitat de primer ordre, N_{θ} és el nombre mitjà d'electrons a la regió θ , i $\rho(\mathbf{r}_1, \mathbf{r}_2)$ és la densitat bielectrònica.

Les diferències entre ambdues densitats es poden veure gràficament a la següent figura 1. Les densitats modificades per generar la figura han estat obtingudes a partir del programa WinFermi [33], en el qual s'ha entrat la densitat normal obtinguda amb el programa Gaussian 98 [34], emprant el nivell Hartree-Fock amb la base 3-21g*.

Tal com es pot veure a la figura 1, en valors baixos de les funcions, la densitat electrònica normal es distribueix uniformement al llarg de tota la molècula, mentre que la modificada es troba ja esbiaixada cap a la zona del grup carboxílic. A mesura que augmenta el valor de la funció, la densitat normal va perfilant més la forma molecular i els enllaços, fins a definir-se tan sols al voltant de les posicions atòmiques. Aquesta tendència no es reflecteix en la densitat modificada, ja que, en augmentar el valor de la funció, aquesta es concentra cada cop més en el grup carboxílic. Tanmateix, com es pot veure a la figura i en la seva formulació, la densitat modificada es troba definida en tot l'espai de la molècula i no està restringida a una regió.

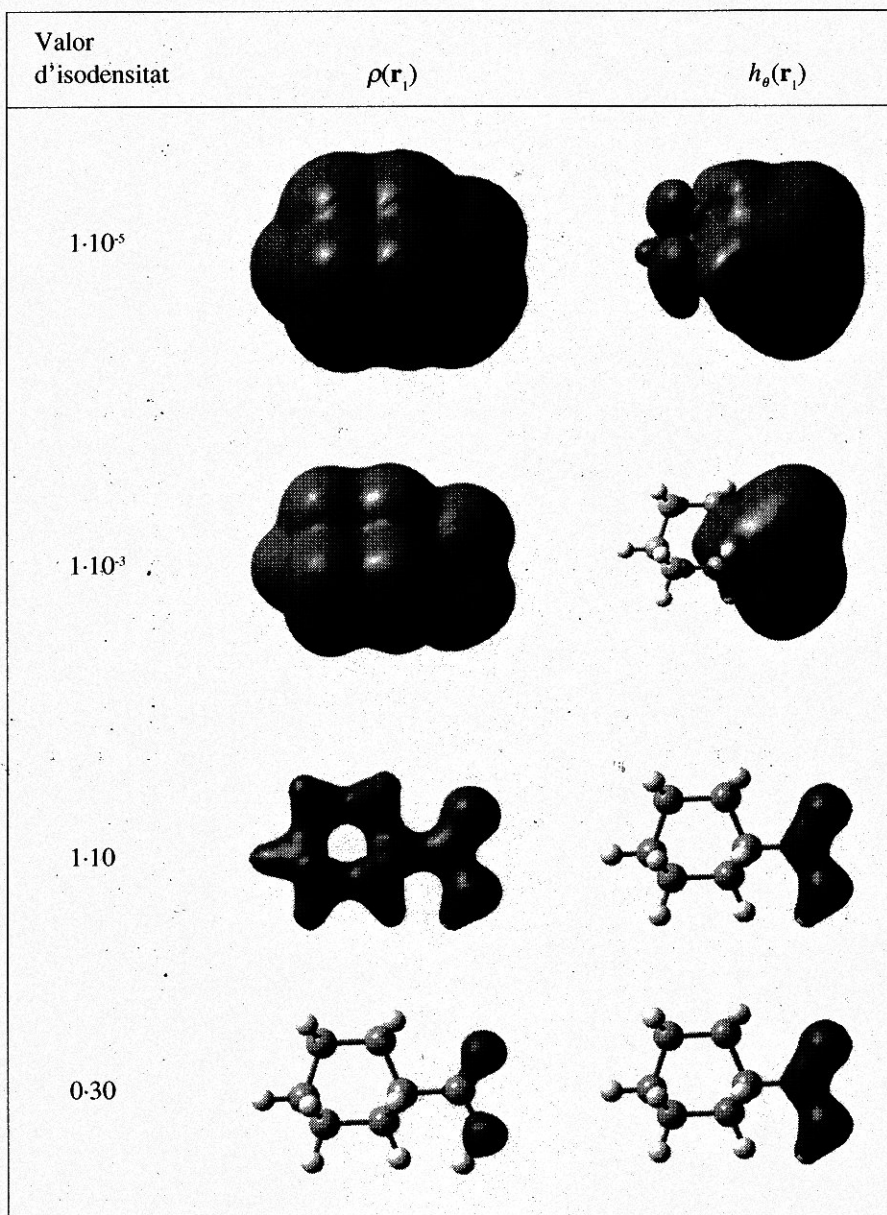


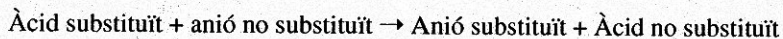
Figura 1. Comparació visual entre funcions de densitat per a la molècula d'àcid biciclo[2,2,2]octà carboxílic. ($\rho(\mathbf{r}_1)$ = funció de densitat de primer ordre, $h_\theta(\mathbf{r}_1)$ funció de densitat de primer ordre modificada per la presència amitjanada d'un electró a la regió definida pel grup carboxílic.

MQSM sobre densitats modificades: autosemblances de fragments

Així doncs, si s'utilitza $h_\theta(\mathbf{r}_i)$ en lloc de la densitat normal en la definició general de les MQSM, tal com s'ha formulat a (2), es poden obtenir mesures d'autosemblança quàntica, on el fragment d'interès, definit per la regió θ , té molta més rellevància sobre la totalitat de la molècula, però amb l'avantatge que la resta de la molècula s'hi troba encara present, per les definicions (3) i (6).

RESULTATS I DISCUSSIO

A fi d'aplicar la metodologia exposada anteriorment, s'han construït una sèrie de molècules derivades de l'àcid biciclo[2,2,2]octà carboxílic, on s'han dut a terme MQSM restringides a la regió definida pel grup COOH. L'activitat estudiada ha estat l'energia de la reacció isodèsmica:



A aquest efecte, s'han pres una sèrie de substituents i s'han construït les estructures corresponents als àcids i als anions, i s'ha optimitzat la geometria al nivell Hartree-Fock amb la base 3-21g* amb el programa Gaussian 98 [34]. Les energies corresponents a cada fragment s'han emmagatzemat. També s'ha dut a terme la transformació de les densitats i el càlcul de MQSM restringits a la regió carboxílica emprant l'operador de recobriment ($\Omega = \delta(\mathbf{r}_1 - \mathbf{r}_2)$), amb els quals s'ha intentat correlacionar l'energia de la reacció. Les dades obtingudes es recopilen a la taula 1.

Substituent	$Z_{\text{M}}(\theta = \text{COOH})$	σ d'inducció	ΔE reacció
Br	208.88903	0.47	7.22
CF ₃	208.88870	0.40	8.01
Cl	208.88718	0.47	7.86
CN	208.89395	0.57	9.67
F	208.88719	0.54	5.94
H	208.88397	0.00	0.00
Me	208.88253	-0.01	0.25
NH ₂	208.88731	0.17	1.02
NO ₂	208.88900	0.67	11.95
OH	208.88351	0.24	3.01

Taula 1. Dades referents a una sèrie de derivats del l'àcid biciclo[2,2,2]octà carboxílic (ΔE en KJ·mol⁻¹).

Els resultats de fer les regressions lineals són a la taula 2.

X	Y	Equació	r
$Z_M(\theta = \text{COOH})$	ΔE reacció	$\Delta E = 973.39Z_M - 203324$	0.775
$Z_M(\theta = \text{COOH})$	σ d'inducció	$\sigma = 55.341Z_M - 11560$	0.772
σ d'inducció	ΔE reacció	$\Delta E = 16.661\sigma - 0.3726$	0.951

Taula 2. Resultats de les regressions lineals.

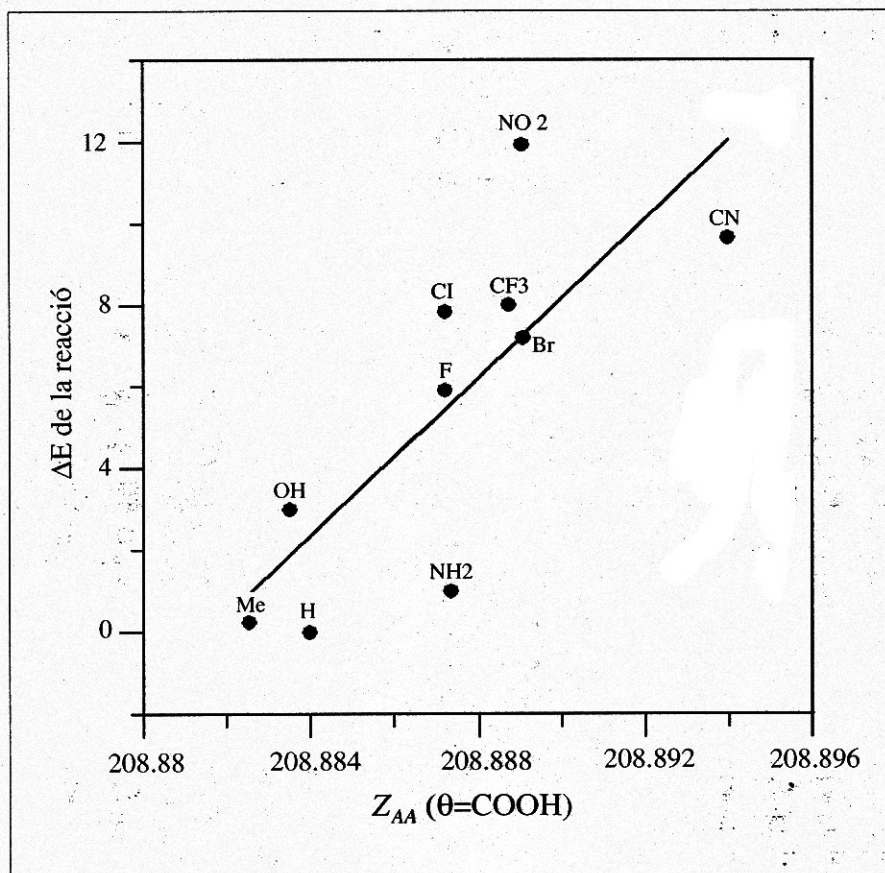


Figura 2. Recta de regressió entre $Z_M(\theta = \text{COOH})$ i la ΔE de la reacció isodèsmica.

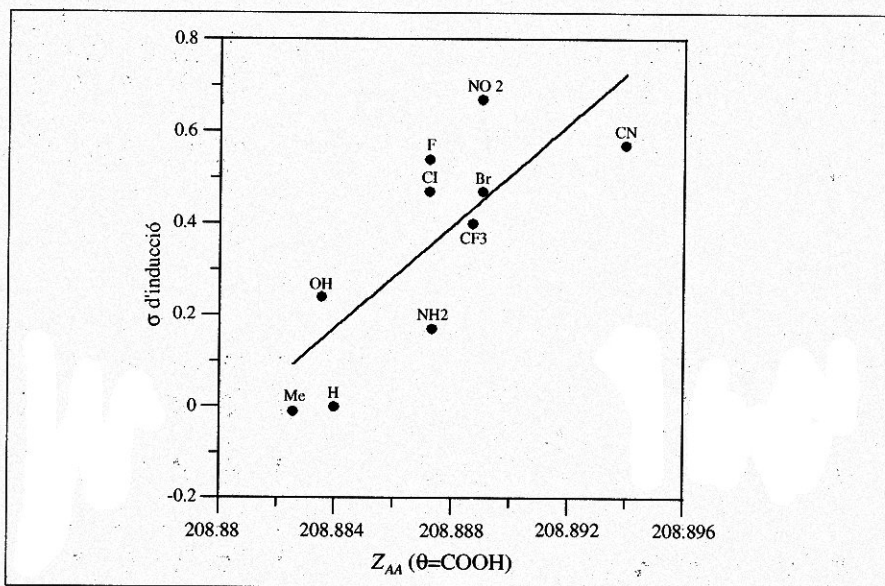


Figura 3. Recta de regressió entre $Z_{AA}(\theta = \text{COOH})$ i la σ de Hammett d'inducció.

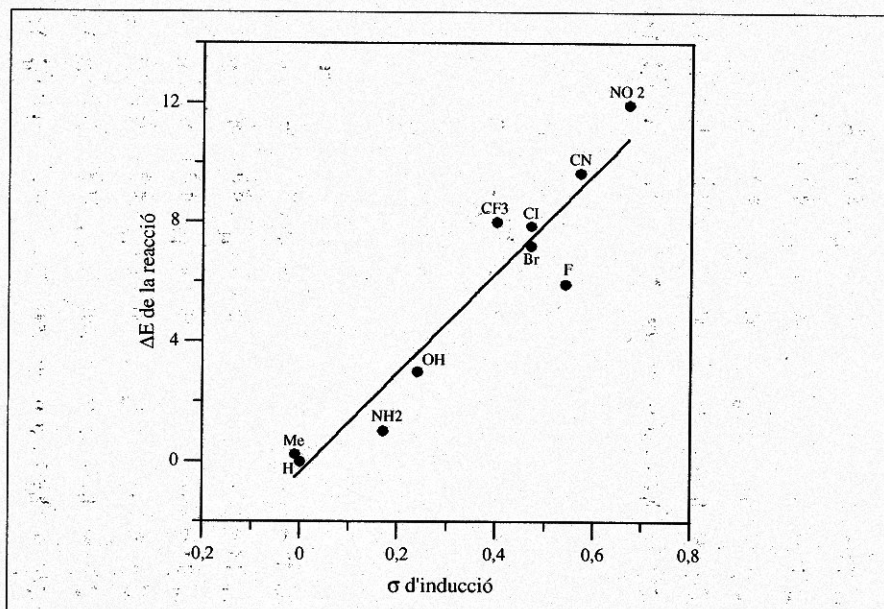


Figura 4. Recta de regressió entre la σ de Hammett d'inducció i la ΔE de la reacció isodèsmica.

Tal com s'observa a la taula 2, existeix una bona relació entre l'autosemblança i l'energia de la reacció (representada gràficament a la figura 2), així com amb la σ d'inducció (figura 3). Tanmateix, l'ús de la σ d'inducció amb l'energia de reacció porta als millors resultats (figura 4).

Aquest últim fet no és descoratjador, ja que, tal com s'ha comentat a la introducció, la σ d'inducció es deriva de dades experimentals, per la qual cosa s'ha de realitzar l'experiment a fi de poder obtenir un valor. En el cas que es tracta en aquest treball, les dades es deriven totalment de la teoria, per la qual cosa el fet de poder reproduir la tendència que provoquen les reaccions ja és satisfactori. Tanmateix, encara cal analitzar la bondat de les diferents etapes del càlcul, sobretot les parts de nivell teòric emprat i l'optimització de la geometria, ja que possiblement l'ús d'una base més completa podria portar a millors resultats, tot i que suposaria un encariment dels costos computacionals; de la mateixa manera, cal verificar que el procés d'optimització porti a un mínim realístic i analitzar l'ús de diferents operadors en la definició general de la semblança a l'equació (2).

CONCLUSIONS

S'ha presentat una metodologia general de càlcul de funcions de densitat modificades per incrementar la presència d'una regió molecular concreta dins d'una molècula. Aquestes densitats electròniques de Forats de Fermi que s'han inclòs en el marc de la semblança quàntica permeten el càlcul de mesures d'autosemblança molecular quàntica sense haver de recórrer a l'ús de fragments moleculars aïllats per realitzar estudis de la relació entre estructura i propietat.

Tot i que els resultats obtinguts en l'exemple presentat no arriben al nivell aconseguit a partir de dades experimentals, s'ha de considerar que la total procedència teòrica de les dades porta a un estalvi en temps i recursos experimentals, per la qual cosa és interessant recercar millores en el protocol seguit.

AGRAIMENTS

El desenvolupament d'aquesta recerca s'ha dut a terme gràcies a l'ajuda del prof. Robert Ponec, de l'Acadèmia de Ciències Txeca (Institute of Chemical Processes Fundamentals, Praga, República Txeca) durant una estada de X. Gironés al centre durant el període gener-març 2001, finançada per la Universitat de Girona. Així mateix, X. Gironés gaudeix d'una beca predoctoral de la mateixa Universitat.

També es reconeixen com a fonts de finançament el Ministeri de Ciència i Tecnologia (projecte núm. SAF2000-0223-C03-01) i la Fundació Maria Francisca de Roviralta.

Referències

1. HAMMETT, L.P. *J. Am. Chem. Soc.*, 59, 1937, 96.
2. HAMMETT, L.P. "Physical Organic Chemistry". McGraw-Hill, Nova York, 1940.
3. TAFT, R.W. *J. Am. Chem. Soc.*, 74, 1952, 3120.
4. SWAIN, C. G., LUPTON JR., E. C. *J. Am. Chem. Soc.*, 90, 1968, 4328.
5. BRUCE, T. C., KHARASCH, N., WINZLER, R.J. *Arch. Biochem. Biophys.*, 62, 1956, 305.
6. ZAHRADNÍK, R., CHVAPIL, M. *Experientia*, 16, 1960, 511.
7. ZAHRADNÍK, R. *Experientia*, 18, 1962, 534.
8. ZAHRADNÍK, R. *Arch. Int. Pharmacodyn.*, 125, 1962, 311.
9. CHVAPIL, M., ZAHRADNÍK, R., CMUCHALOVA, B. *Arch. Int. Pharmacodyn.*, 125, 1962, 330.
10. HANSEN, E. R. *Acta Chem. Escand.*, 16, 1962, 1593.
11. FREE, S. M., WILSON, J. W. *J. Med. Chem.*, 7, 1964, 395.
12. BODEK, K., KOPECK_, J., KRIVUCOVA, M., VLACHOVA, D. *Experientia*, 20, 1964, 667.
13. KOPECK_, J., BOCEK, K., VLACHOVA, D. *Nature*, 207, 1965, 667.
14. PONEC, R., AMAT, L., CARBÓ-DORCA, R. *J. Comput.-Aided Mol. Des.*, 13, 1999, 259.
15. PONEC, R., AMAT, L., CARBÓ-DORCA, R. *J. Phys. Org. Chem.*, 12, 1999, 447.
16. CARBÓ, R., LEYDA, L., ARNAU, M. *Int. J. Quant. Chem.*, 17, 1980, 1185.
17. CARBÓ, R., DOMINGO, L. *Int. J. Quant. Chem.*, 23, 1987, 517.
18. CARBÓ, R., CALABUIG, B. *Comput. Phys. Commun.*, 55, 1989, 117.
19. CARBÓ, R., CALABUIG, B. *Int. J. Quant. Chem.*, 42, 1992, 1681.
20. CARBÓ, R., CALABUIG, B. *Int. J. Quant. Chem.*, 42, 1992, 1695.
21. CARBÓ, R., CALABUIG, B. *J. Mol. Struct. (Theochem)*, 254, 1992, 517.
22. CARBÓ, R., CALABUIG, B., BESALÚ, E., MARTÍNEZ, A. *Mol. Engineering*, 2, 1992, 43.
23. CARBÓ, R., CALABUIG, B., BESALÚ, E., VERA, L. *Adv. Quantum Chem.*, 25, 1994, 255.
24. BESALÚ, E., CARBÓ, R., MESTRES, J., SOLA, M. *Top. Curr. Chem.*, 173, 1995, 31.
25. CARBÓ-DORCA, R., BESALÚ, E. *J. Mol. Struct. (Theochem)*, 451, 1998, 11.
26. CARBÓ-DORCA, R., AMAT, L., BESALÚ, E., GIRONÉS, X., ROBERT, D. *J. Mol. Struct. (Theochem)*, 504, 2000, 181.
27. CARBÓ, R. (ed.) "Molecular similarity and reactivity: from quantum chemical to phenomenological approaches", Kluwer, Amsterdam, 1995.
28. CARBÓ-DORCA, R., MEZEY, P.G. (ed.) "Advances in Molecular Similarity", vol. 1, JAI Press, Greenwich, 1996.
29. CARBÓ-DORCA, R., MEZEY, P.G. (ed.) "Advances in Molecular Similarity", vol. 2, JAI Press, Greenwich, 1998.

30. CARBÓ-DORCA, R. GIRONÉS, X., MEZEY, P.G. (ed.) "The Fundamentals of Molecular Similarity", Plenum Pub. Corp., Nova York, 2001.
31. PONEC, R., STRNAD, M. *Int. J. Quantum. Chem.*, 50, 1994, 43.
32. PONEC, R., UHLIK, F. *Croat. Chem. Acta*, 69, 1996, 941.
33. GIRONÉS, X., PONEC, R., ROITHOVA, J. WinFermi, v1.0. Institute of Chemical Process Fundamentals, Academy of Sciences of Czech Republic, 2001, Praga, República Txeca.
34. FRISCH, M. J., TRUCKS, G. W., SCHLEGEL, H. B., SCUSERIA, G. E., ROBB, M. A., CHEESEMAN, J. R., ZAKRZEWSKI, V. G., MONTGOMERY, JR., J. A., STRATMANN, R. E., BURANT, J. C., DAPPRICH, S., MILLAM, J. M., DANIELS, D., KUDIN, K. N., STRAIN, M. C., FARKAS, O., TOMASI, J., BARONE, V., COSSI, M., CAMMI, R., MENNUCCI, B., POMELLI, C., ADAMO, C., CLIFFORD, S., OCHTERSKI, J., PETERSSON, G. A., AYALA, P. Y., CUI, Q., MOROKUMA, K., MALICK, D. K., RABUCK, A. D., RAGHAVACHARI, K., FORESMAN, J. B., CIOSLOWSKI, J., ORTIZ, J. V., STEFANOV, B. B., LIU, G., LIASHENKO, A., PISKORZ, P., KOMAROMI, I., GOMPERTS, R., MARTIN, R. L., FOX, D. J., KEITH, M. T., AL-LAHAM, A., PENG, C. Y., NANAYAKKARA, A., GONZALEZ, C., CHALLACOMBE, M., GILL, P. M. W., JOHNSON, B., CHEN, W., WONG, M. W., ANDRES, J. L., GONZALEZ, C., HEAD-GORDON, M., REPLOGLE, E. S., POPL, J. A. Gaussian 98, Revision A.6, Gaussian, Inc., Pittsburgh PA, 1998.

Molecular Quantum Similarity-based QSARs for Binding Affinities of several Steroid Sets

Xavier Gironés and Ramon Carbó-Dorca*
Institute of Computational Chemistry, University of Girona.
Campus Montilivi, 17071 Girona, Catalonia, Spain.

Steroid_paper.doc

ABSTRACT

The application of Molecular Quantum Similarity Measures (MQSM) to correlate biological activities for three different sets of steroids is reported. A general protocol for the generation of descriptors is detailed, thus covering molecular superposition, electronic density fitting and quantum similarity calculation issues. Satisfactory Quantitative Structure-Activity Relationship (QSAR) models ($r^2 \in [0.69, 0.94]$ and $q^2 \in [0.59, 0.73]$), comparable to previous studies, are obtained in all cases, where steroid binding affinities to different enzymes are studied. In this work, MQSM, properly scaled using Carbó Index, are related to activity using a Partial Least Squares routine.

KEYWORDS

Quantum Similarity, MQSM, Carbó Index, PLS, QSAR, Steroids, binding affinity.

* Corresponding author. Phone: +34 972418367. Fax: +34 972418356. E-mail: quantum@iqc.udg.es.

INTRODUCTION

One of the most promising subjects in present day computational chemistry is the characterization and quantification of molecular properties and biological activities by means of structure-based descriptors, generated from theoretical development and computationally-based methodologies. Given a molecular set and a relevant molecular property or biological activity, a quantitative structure–property or –activity relationship (QSPR/QSAR) [1,2] model can be constructed by derivation of molecular structure descriptors, which will be related to a molecular feature by some statistical procedure.

Firstly started by Cross in 1863 [3], observing that toxicity of alcohols to mammals increased as water solubility of such alcohols decreased, QSPR/QSAR techniques have been largely developed using a large variety of parameters as descriptors of molecular structural characteristics. For example, Hammett sigma values [4] are often used as electronic parameters, or the octanol/water partition coefficient [5] for lipophilicity description. Other frameworks have been devised to account for the shape, size, polarizability, and many other molecular structural features. In addition, the recent and fast development of computer architectures has increased the computational power to stages that allow the application of the quantum theory to regular organic molecules at fairly accurate computational levels (semiempirical or ab initio methods with appropriate basis sets) within reasonable time limits and affordable costs. A number of reviews have been published [1,2,6], concerning the historical development, generation of descriptors and different methodologies in the QSPR/QSAR fields.

Arising from the quantum point of view of a molecule, and according to the postulate that its density function contains all accessible information contained within this molecule, Molecular Quantum Similarity Measures (MQSM) [7-13] stand as a general efficient tool to solve actual chemical problems. MQSM methodology has been successfully applied within pharmacological [14-19] and toxicological [20-22] problems. MQSM are based on the concept of molecular similarity [23,24] and are related to a self-evident molecular similarity principle: “the more similar two molecules are, the more similar properties will possess”. This last statement requires a procedure to compare the molecules, and MQSM easily establish the degree of similarity based on

the electronic distribution, namely first-order density functions (DF), of the molecular structures.

In this work, it is intended to use the MQSM methodology to correlate binding affinities of three sets of steroids, each related to a different enzyme, possessing diverse biological responses. For this purpose, this paper is structured as follows: a brief description of the proposed protocol is given first, next the presentation of the molecular sets, as well as the results achieved, are described, and the concluding remarks are finally given.

MATERIAL AND METHODS

MQSM and Carbó Indices. Up to date, the practical application of MQSM lays on the use of first-order density functions. These functions are quantum-mechanical observable elements producing information on the molecular electron distribution. Within this framework, two molecular structures are considered to be similar if their electron distributions are. Thus, among other possibilities [9], quantitative measures of the similarity between two molecules can be defined as the direct volume integral between their density functions, expressed as:

$$Z_{AB} = \iint \rho_A(\mathbf{r}_1) \delta(\mathbf{r}_1 - \mathbf{r}_2) \rho_B(\mathbf{r}_2) d\mathbf{r}_1 d\mathbf{r}_2, \quad (1)$$

where $\rho_A(\mathbf{r})$ and $\rho_B(\mathbf{r})$ are the first-order electron density functions of molecules A and B respectively, $\delta(\mathbf{r}_1 - \mathbf{r}_2)$ is Dirac's Delta function, and Z_{AB} is the resulting quantum similarity measure, in this case known as *Overlap* similarity measure, which simplifies Eq. (1) to:

$$Z_{AB} = \int \rho_A(\mathbf{r}) \rho_B(\mathbf{r}) d\mathbf{r}, \quad (2)$$

As MQSM depend on the relative position of both compared objects in space, a molecular alignment procedure is required. When studied molecular sets share common structural features, like the 4 rings structure present in all steroids, Topo-Geometrical Superposition Algorithm (TGSA) [25] is used, as it performs pairwise superpositions

according to the molecular backbone coincidence. Once superposed, the overall set of MQSM for a series of N molecules are computed and collected in the way of a $(N \times N)$ similarity matrix form, which can be used to extract the appropriate correlation information and build QSPR/QSAR descriptors. Once the similarity matrix has been computed, a possible scaling of the MQSM can be done by means Carbó Index [7], as in Eq. (3):

$$C_{ij} = Z_{ij} (Z_{ii} Z_{jj})^{-1/2}, \quad (3)$$

providing a weighted similarity measure inside the open interval (0,1). Within this scope, the closest the value to unity, the higher is the degree of similarity between both analyzed molecules, and vice versa otherwise. A singular case consists of $C_{ij} = 1$, indicating that both molecular structures are the same.

Promolecular Atomic Shell Approximation (ASA). To avoid expensive theoretical calculations, the promolecular atomic shell approximation (ASA) [27-29] has been used here to construct molecular first-order electron density. Within ASA the molecular density is expressed as a sum of discrete atomic density contributions, which are taken as a superposition of 1S Gaussian functions and fitted to atomic ab initio ones. Once the overall atomic densities are built, each molecular density function can be constructed by adding as appropriate these elementary atomic building blocks. Since it has been proved that the MQSM from fitted densities built in this way differ by up to a 2% from the ab initio ones [28], their use is clearly justified.

Molecular Alignment. MQSM are dependent on the relative position of the molecules under comparison. In this way, the Topo-Geometrical Approach (TGSA) [25] has been used to perform the needed pairwise molecular alignments. This molecular superposition method overlays the involved molecules, considered as rigid bodies, according to the maximal common substructure shared by the analysed compounds. As an illustrative example of TGSA results, a pairwise superposition between two steroids (both compounds of the first studied molecular set) is presented in Figure 1.

Figure 1, about here.

Molecular Modeling. All molecules have been sketched and cleaned using Weblab Viewer 4.0 [30]. The geometry optimization of the resulting structures has been carried out in gas phase with Mopac 6.0 [31] software package at the AM1 [32] computational level. Molecular electronic DF has been constructed using the ASA approach, as described above, using parameters fitted to Huzinaga basis set [29].

Treatment of Quantum Similarity Matrices and Model Building. Common chemometric tools may be applied to deal with similarity matrices. Particularly, Partial Least Squares (PLS) [33,34] stands as an ideal technique to obtain a generalized regression to model the association between the matrices **X** (descriptors) and **Y** (responses). In computational chemistry, its main use is to model the relationship between computed variables, which together characterize the structural variation of a set of N compounds and any property of interest measured on those N substances [35-37]. This variation of the molecular skeleton is condensed into the matrix **X**, whereas the analyzed properties are recorded into **Y**. In PLS, the matrix **X** is commonly built up from non-independent data, as it uses to have more columns than rows; hence it is not called the *independent* matrix, but predictor or descriptor matrix. A good review, as well as its practical application in QSAR, is found in ref. [38], and a detailed tutorial in ref. [39].

Unlike regression, PLS is not based on the assumption of independent and precise **X** variables, but it is rather based on the more realistic assumption that **X** contains more or less collinear and noisy parameters. PLS summarizes these **X** variables by means of a few orthogonal score vectors ($\mathbf{t}_a \in \mathbf{T}$), and the matrix **Y** is also resumed in few score vectors ($\mathbf{u}_a \in \mathbf{U}$) which are not orthogonal. Plots of columns from **T** and **U** provide visual representation of the configuration of the observations in the **X** or **Y** spaces, respectively. The PLS procedure allows to derive a number of *factors* and *weights*, which are used to describe the desired properties. QSPR models are built up from these factors and weights.

In this work, all obtained models are evaluated by commonly used statistical parameters: goodness-on fit (r^2) [40], root mean square error between experimental and

predicted values (s) [40] by *leave-one-out* [41] and predictive capacity (q^2) [40]. In addition, and in order to avoid chance correlations and excess of parameters, models are submitted to random tests, where the properties are randomly permuted in their positions and the entire modeling procedure is repeated a number of times, a thousand in our case. If satisfactory correlations are found within the random test, the model obtained should not be trusted, as the methodology used may be potentially capable to correlate any kind of data.

RESULTS AND DISCUSSION

In this section, it is intended to prove the usefulness of the proposed procedure with three molecular sets constituted by steroids, where in each case the associated biological response is related to the binding activity of the involved molecules to an enzyme. The first example is composed by 50 steroids and the studied activity is the binding affinity to aromatase enzyme [42]. The second example, built up from 93 6-azasteroids, the studied properties are related to both binding affinity to 5α -reductase (type I) and 3β -hydroxy- Δ^5 -steroid dehydrogenase/3-keto- Δ^5 -steroid isomerase (3-BHSD), and selectivity to bind to 3-BHSD relative to 5α -reductase (type I) [43]. The last steroid set is the widely studied Crammer's 31 steroid set, where the corticosteroid binding affinity is studied for such molecular series [44,45].

Steroid Binding to Aromatase Enzyme. The importance of aromatase enzyme, which consists of a cytochrome P450 complex that plays a role in the conversion of androgens to estrogens, lies on its implication on breast cancer [46]. Since this enzyme catalyzes the testosterone to estradiol, a typical estrogen, conversion, its activity is directly involved in the evolution and development of estrogen-dependent tumors, such as breast cancer [46,47]. Designing inhibitors of this enzyme may lead to therapeutically useful drugs able to control or avoid the development of affected cells. For this purpose, a set of 50 steroids [42], inhibitors of aromatase enzymes, with experimentally determined binding affinity have been studied with MQSM. All data regarding molecular structures and binding activity of the compounds is summarized in Table 1.

Table 1, about here.

As stated in the methodological section, the PLS technique is used to create QSAR models for this molecular set from descriptors derived from quantum similarity measures. Model have been created up to 10 descriptors, and the optimal number of parameters to be used in the final model, which in this case it has been fixed to 5 PLS factors has been chosen from the evolution of both correlation (r^2) and predictivity (q^2), as presented in Figure 2.

Figure 2, about here.

As observed in Figure 2, a sharp increase in the predictive capacity occurs up to the third PLS factor, and a slow increase until the sixth one. A further increase in the number of parameters results, in despite of the expected increase of r^2 , in a decrease of the predictive power of the generated model. In this way, the chosen number of PLS factors used was 5, as the results provide the optimal balance between predictivity and descriptors applied, and a further increase does not lead to noticeable improved results and could lead to overfitted models. Thus, the final equation and statistical results for this molecular set are:

$$\begin{aligned} \text{Binding Activity} &= 0.132 \cdot \mathbf{f}_1 + 6.141 \cdot \mathbf{f}_2 + 6.440 \cdot \mathbf{f}_3 + 4.510 \cdot \mathbf{f}_4 + 6.721 \cdot \mathbf{f}_5 \\ r^2 &= 0.839 & q^2 &= 0.734 & s &= 0.710 \end{aligned}$$

The results are also graphically presented by plotting the experimental binding activities, versus the predicted ones from the PLS procedure, as shown in Figure 3. In addition, the random test was carried out permuting 1000 times the activity values and reconstructing the QSAR model each time. The results are graphically compiled in Figure 4, where each point represents the values of r^2 and q^2 achieved in each permutation.

Figures 3 and 4, about here.

As seen from the statistical and graphical results, satisfactory results are obtained for this molecular set. Most of the compounds are correctly predicted, within a

close margin, even few deviations are present, as displayed in Figure 3. The random test shows a clear separation between the original point (+) and the permuted ones (●). The results of the previously published work [42] are comparable, yielding to a range of results $r^2 \in [0.73, 1.0]$ and $q^2 \in [0.66, 0.75]$, using from 5 to 87 parameters based on ^{13}C spectrometric data.

Steroid binding and selectivity to 5 α -reductase. 5 α -reductase is responsible for the biological conversion of testosterone into androgen dihydrotestosterone (DHT). Since DHT is related to increased body hair, hair line recession, prostatic enlargement, and, at high concentration levels, DHT is also associated to several disorders, like prostatic cancer, male baldness and acne, several investigations have point their attention to 5 α -reductase [48-62], due to it has been detected at higher levels in those affected areas. Therefore, inhibition of this enzyme becomes a faithful therapeutically efficient treatment for disorders caused by high levels of DHT. Also, it has been proved that, for an optimal treatment, inhibition rate should be complemented with selectivity to other enzymes [60,61]. In this way, ideal inhibitors should possess low binding affinity to 5 α -reductase and a high one to 3-BHSD.

For this molecular set, composed by 93 azasteroids, two QSAR models have been developed to correlate the two relevant properties: the necessary nanomolar concentration to inhibit a sample of 5 α -reductase (expressed as $\log K_i$) [43] and the ability to selectively inhibit 5 α -reductase relative to 3-BHSD, denoted as S_i , which is the ratio of K_i for 3-BHSD and K_i for 5 α -reductase (expressed as $\log S_i$) [43]. All data for this molecular set is summarized in Table 2.

Table 2, about here.

Following the previous exposed protocol, the optimal QSAR models are fixed to 5 PLS factors, whose results are presented below for $\log K_i$:

$$\log K_i = 0.221 \cdot f_1 + 2.351 \cdot f_2 + 3.829 \cdot f_3 + 2.367 \cdot f_4 + 2.747 \cdot f_5$$

$$r^2 = 0.688 \qquad q^2 = 0.594 \qquad s = 0.888$$

and $\log S_i$:

$$\log S_i = 0.034 \cdot f_1 + 2.982 \cdot f_2 + 3.719 \cdot f_3 + 2.353 \cdot f_4 + 2.179 \cdot f_5$$

$$r^2 = 0.699 \qquad q^2 = 0.600 \qquad s = 0.959$$

As seen from the statistical results, acceptable relationships with valuable predictive capacity are obtained. These predictive capacities can be visually observed in Figures 5 and 6, where experimental versus predicted values are represented for $\log K_i$ and $\log S_i$, respectively.

Figures 5 and 6, about here

As seen in Figures 5 and 6, accurate predictions are obtained for both activities for most of the compounds present in the set. Even if a few deviations are present, the proposed model can still be used for classification or discriminant analysis, allowing to order the compounds by discrete ranges of predicted activity. Also, satisfactory results arise from the random test, where after 1000 permutations were carried out. In the $\log K_i$ case, the highest values of r^2 and q^2 were 0.406 and 0.236, whereas the mean values were 0.205 and -0.211, respectively. When analyzing $\log S_i$, the maximum values were 0.373 and 0.245, and the mean values 0.203 and -0.232, respectively. Since both activities are quite correlated, $r^2=0.612$, it is not a coincidence that the statistical results, as well as the slopes of the proposed models, become similar. The previous study [43] also obtained similar results when using simple statistical tools, like linear regression, PLS or PCR, to correlate the activities, selecting a small set of optimal descriptors from a bunch of 224 parameters (topological, geometric, electronic and hybrid ones). However, the usage of neural networks considerably improved the results, yielding to high predictive capacity.

Steroid binding affinity to corticosteroid binding globulin. The importance of corticosteroids rests on their role played in human physiology. Depending on the side group of the binding enzyme, its action ranges from nutrient metabolism regulation to electrolyte balance. In addition, and properly activated, several therapeutical applications have been found, such as treatments for rheumatoid arthritis, allergies,

Crohn's disease, uveitis or minimization of the complications associated to premature births [63-67]. Consequently, many research efforts [16,18,44,45,68] have been devoted to obtain significant models for this kind of compounds.

The exposed protocol is applied to this molecular set composed by 31 steroids, whose binding affinity is expressed in terms of pK. All data regarding the structures is summarized in Table 3. The optimal QSAR model using MQSM is composed of 4 PLS factors, whose results are presented below:

$$\text{pK} = 1.237 \cdot \mathbf{f}_1 + 4.795 \cdot \mathbf{f}_2 + 5.255 \cdot \mathbf{f}_3 + 5.591 \cdot \mathbf{f}_4$$

$$r^2 = 0.751 \qquad q^2 = 0.654 \qquad s = 0.636$$

Table 3, about here.

As seen from the previous statistical results, a satisfactory QSAR model is found with a remarkable predictive capacity, as judged from the q^2 value achieved. A clearer picture of the obtained results can be seen by browsing Figure 7, where the experimental pK are plotted versus the predicted ones. As seen in Figure 7, most of the compounds are correctly predicted, except for four ones, which are overestimated, being one the 31th structure (fluorine-substituted), which is commonly overpredicted [16,18,44,45,68]. The random test is satisfactorily overcome, yielding to mean values of 0.281 and -0.322 for r^2 and q^2 , respectively, whereas the maximum values were 0.596 and 0.326 after 1000 permutations of the pK data.

Figure 7, about here

According to literature [18,45], the obtained results with the proposed model are within the range covered by previous works, where a variety of both molecular descriptor generation and numerical data treatment were used.

CONCLUSIONS

A general protocol for the QSAR model generation for steroids using Molecular Quantum Similarity Measures has been presented, where different molecular activities related to binding activity of these compounds to different enzymes have been tested for correlation search. In all three application examples, satisfactory correlations were obtained using a relative small number of molecular descriptors.

Even if other methodologies may provide better results, it must be emphasized that the procedure used along this work, as well as the generation of molecular descriptors, has been kept untouched. In other words, the exposed QSAR protocol consists of a methodological pathway made of unbiased and universal MQSM descriptors able to characterize different molecular activities for steroids without introducing further information than those provided by quantum similarity based on electronic density functions. Additional refinements or statistical tools may be applied to the procedure in order to improve the results according to each molecular set under study; however, the exposed protocol may constitute an excellent starting point for subsequent research.

ACKNOWLEDGEMENTS

This research has been partially supported by Project No. SAF2000-0223-C03-01 from the Spanish *Ministerio de Ciencia y Tecnología*. One of us (X. Gironés) benefits from a predoctoral fellowship from the University of Girona. Financial support from the *Fundació María Francisca de Roviralta* is also acknowledged.

REFERENCES

1. Jurs, P.C. Quantitative Structure-Property Relationships (QSPR). In *Encyclopedia of Computational Chemistry*. Schleyer, P.v.R.; Allinger, N.L.; Clark, T.; Gasteiger, J.; Kollman, P.A.; Schaefer III, H.F.; Schreiner, P.R. (eds.) John Wiley and Sons Ltd., Chichester, UK, 1998, Vol. 4, pp 2320-2330.
2. Kubinyi, H. Quantitative Structure-Activity Relationships in Drug Design. In *Encyclopedia of Computational Chemistry*. Schleyer, P.v.R.; Allinger, N.L.; Clark,

- T.; Gasteiger, J.; Kollman, P.A.; Schaefer III, H.F.; Schreiner, P.R. (eds.) John Wiley and Sons Ltd., Chichester, UK, 1998, Vol. 4, pp 2309-2319.
3. Quoted in: Borman, S. New QSAR Techniques Eyed for Environmental Assessments. *Chem. Eng. News* **1990**, 68, 20-23.
4. Hammett, L. P. The Effect of Structures upon the Reactions of Organic Compounds. Benzene Derivatives. *J. Am. Chem. Soc.* **1937**, 59, 96-103.
5. Hansch, C.; Fujita, T. σ - ρ - π analysis. A method for the correlation of biological activity and chemical structure. *J. Am. Chem. Soc.* **1964**, 86, 5175-5180.
6. Waterbeemd, H.v.d. (ed.) *Structure-Property Correlations in Drug Research*. Academic, R.G. Landes Company, Austin, 1996.
7. Carbó, R.; Arnau, J.; Leyda, L. How similar is a molecule to another? An electron density measure of similarity between two molecular structures. *Int. J. Quantum Chem.*, **1980**, 17, 1185-1189.
8. Carbó-Dorca, R.; Besalú, E. A general survey of molecular quantum similarity. *TEOCHEM*, **1998**, 451, 11-23.
9. Carbó-Dorca, R.; Amat, L.; Besalú, E.; Lobato, M. Quantum Similarity. In *Advances in Molecular Similarity*; Carbó-Dorca, R.; Mezey, P.G., Eds.; JAI Press, Greenwich, CT, 1998, Vol. 2, pp 1-42.
10. Carbó-Dorca, R. Fuzzy sets and Boolean tagged sets; vector semispaces and convex sets; quantum similarity measures and ASA density functions; diagonal vector spaces and quantum chemistry. In *Advances in molecular similarity*; Carbó-Dorca, R.; Mezey, P.G., Eds.; JAI Press, Greenwich, CT, 1998, Vol. 2, pp 43-72.
11. Besalú, E.; Carbó, R.; Mestres, J.; Solà, M. Foundations and recent developments on molecular quantum similarity, *Top. Curr. Chem.*, **1995**, 173, 31-62.
12. Carbó, R.; Besalú, E.; Theoretical foundations of quantum molecular similarity. In *Molecular similarity and reactivity: from quantum chemical to phenomenological approaches*; Carbo, R., Ed.; Kluwer, Amsterdam, 1995, pp 3-30.
13. Carbó-Dorca R.; Besalú, E.; Amat, L.; Fradera, X. Quantum molecular similarity measures: concepts, definitions and applications to quantitative structure-properties relationships. In *Advances in molecular similarity*; Carbó-Dorca, R.; Mezey, P.G., Eds.; JAI Press, Greenwich, CT, 1996, Vol. 1, pp 1-42.
14. Carbó, R.; Besalú, E.; Amat, L.; Fradera, X. Quantum molecular similarity measures (QMSM) as a natural way leading towards a theoretical foundation of quantitative structure-properties relationships. *J. Math. Chem.*, **1995**, 18, 237-246.

15. Fradera, X.; Amat, L.; Besalú, E.; Carbó-Dorca, R. Application of molecular quantum similarity to QSAR. *Quant. Struct.-Act. Relat.*, **1997**, 16, 25-32.
16. Lobato, M.; Amat, L.; Carbó-Dorca, R. Structure-activity relationships of a steroid family using quantum similarity measures and topological quantum similarity indices. *Quant. Struct.-Act. Relat.*, **1997**, 16, 465-472.
17. Amat, L.; Robert, D.; Besalú, E.; Carbó-Dorca, R. Molecular quantum similarity measures tuned 3D QSAR: an antitumoral family validation study. *J. Chem. Inf. Comput. Sci.*, **1998**, 38, 624-631.
18. Robert, D.; Amat, L.; Carbó-Dorca, R. Three-dimensional quantitative structure-activity relationships from tuned molecular quantum similarity measures. Prediction of the corticosteroid-binding globulin binding affinity for a steroid family. *J. Chem. Inf. Comput. Sci.*, **1999**, 39, 333-344.
19. Robert, D.; Gironés, X.; Carbó-Dorca, R. Facet diagrams for quantum similarity data. *J. Comput.-Aided Mol. Des.*, **1999**, 13, 597-610.
20. Gironés, X.; Amat, L.; Robert, D.; Carbó-Dorca, R. Use of electron-electron repulsion energy as a molecular descriptor in QSAR and QSPR studies. *J. Comput.-Aided Mol. Des.*, **2000**, 14, 477-485.
21. Robert, D.; Carbó-Dorca, R., Aromatic compounds aquatic toxicity QSAR using quantum similarity measures. *SAR QSAR Environ. Res.*, **1999**, 10, 401-422.
22. Gironés, X.; Amat, L.; Carbó-Dorca, R. Using molecular quantum similarity measures as descriptors in quantitative structure-toxicity relationships. *SAR QSAR Environ. Res.* **1999**, 10, 545-556.
23. Dean, P.M., Molecular Similarity. In *3D QSAR in Drug Design: Theory, Methods and Applications*; Kubinyi, H., Ed.; ESCOM Science Publishers B.V., Leiden, The Netherlands, 1993, pp 150-172.
24. Richards, W.G., Molecular Similarity and Dissimilarity. In *Modeling of Biomolecular Structures and Mechanisms*, Pullman, A., Jortner, J., Pullman, B., Eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1995, pp 365-369.
25. Gironés, X.; Robert, D.; Carbó-Dorca, R. TGSA: a molecular superposition program based on Topo-Geometrical Considerations., *J. Comp. Chem.*, **2001**, 22, 255-263.
26. Carbó-Dorca, R. Stochastic Transformation of Quantum Similarity Matrices and Their Use in Quantum QSAR (QQSAR) Models. *Int. J. Quantum. Chem.*, **2000**, 79, 163-177.

27. Constants, P.; Carbó, R. Atomic shell approximation: electron density fitting algorithm restricting coefficients to positive values. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1046-1053.
28. Amat, L.; Carbó-Dorca, R. Quantum similarity measures under atomic shell approximation: first-order density fitting using elementary Jacobi rotations. *J. Comput. Chem.* **1997**, *18*, 2023-2039.
29. Amat, L.; Carbó-Dorca, R. Fitted electronic density functions from H to Rn for use in quantum similarity measures: cis-diamminedichloroplatinum(II) complex as an application example. *J. Comput. Chem.* **1999**, *20*, 911-920.
30. WebLab Viewer Pro 4.0, Molecular Simulations Inc., 2000. A free trial version is available from its website: <http://www.msi.com>.
31. Lovanov, V. Mopac 6.0, University of Florida, 1996.
32. Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. AM1: A New General Purpose Quantum Chemical Molecular Model. *J. Am. Chem. Soc.* **1985**, *107*, 3902-3909.
33. Höskuldsson A., Prediction Methods in Science and Technology. Thor, Copenhagen, 1996.
34. Tenenhaus, M., Regression de PLS. Editions Technip, Paris, 1997
35. Wold, S.; Johansson, E.; Cocchi, M., PLS – Partial Least-Squares Projections to Latent Structures. In *3D QSAR in Drug Design, Theory, Methods and Applications*; Kubinyi, H. (ed), ESCOM Science, Leiden, 1993, pp 523-550
36. Wold, S., PLS for multivariate linear modelling. In *Methods and Principles in Medicinal Chemistry*. Vol. 2: *Chemometric Methods in Molecular Design*; van de Waterbeemd, H. (ed), VCH, Weinheim, 1995, pp 195-218.
37. Cramer III, R.D.; Patterson, D.E.; Bunce, J.D. Comparative Molecular Field Analysis (CoMFA). 1. Effect on Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959-5967.
38. Wold, S.; Sjöström, M.; Eriksson, L. Partial Least Squares Projections to Latent Structures (PLS) in Chemistry. In *Encyclopedia of Computational Chemistry*; von Ragué Schleyer, P.; Allinger, N.L.; Clark, T.; Gasteiger, J.; Kollman, P.A.; Schaefer III, H.F.; Screiner, P.R. (ed); John Wiley and Sons Ltd., Chichester, UK, 1994, Vol. 4, pp 2006-2021.
39. Geladi, P.; Kowalski, B.R. Partial Least-Squares Regression: A Tutorial. *Anal. Chim. Acta.* **1986**, *185*, 1-17.

40. Montgomery, D.C.; Peck, E.A. Introduction to linear regression analysis. Wiley, New York, 1992.
41. Wold, S. Cross-validatory estimation of a number of components in factor and principal component models. *Technometrics* **1978**, *20*, 397-405.
42. Beger, R.D.; Buzatu, D.A.; Wilkes, J.G.; Lay Jr, J.O. ¹³C NMR Quantitative Spectrometric Data-Activity Relationship (QSDAR) Models of Steroids Binding the Aromatase Enzyme. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1360-1366.
43. Bakken, G.A.; Jurs, P.C. QSARs for 6-Azasteroids as Inhibitors of Human Type 1 5 α -Reductase: Prediction of Binding Affinity and Selectivity Relative to 3-BHSD. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1255-1265.
44. Cramer, R.D.; Patterson, D.E.; Bunce, J.D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959-5967.
45. Maw, H.H.; Hall, L.H. E-State modeling of corticosteroid binding affinity validation of model for small data set. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1248-1254.
46. Brodie, A.M.; Santen, R.J. Aromatase and its inhibitors in breast cancer treatment – overview and perspective. *Breast Cancer Res. Treat.* **1994**, *30*, 1-6.
47. Brodie, A.M. Aromatase inhibitors in the treatment of breast cancer. *J. Steroid Biochem. Mol. Biol.* **1994**, *49*, 281-287.
48. Singh, S.M.; Gauthier, S.; Labrie, F. Androgen receptor antagonists (antiandrogens): structure-activity relationships. *Curr. Med. Chem.* **2000**, *7*, 211-247.
49. Clifford, G.M.; Farmer, R.D.T.; Medical therapy for benign prostatic hyperplasia: a review of the literature. *Eur. Urol.* **2000**, *38*, 2-19.
50. Bartsch, G.; Rittmaster, R.S.; Klocker, H. Dihydrotestosterone and the concept of 5 α -reductase inhibition in human benign prostatic hyperplasia. *Eur. Urol.* **2000**, *37*, 367-380.
51. Weisser, H.; Tunn, S.; Debus, M.; Krieg, M. 5 α -reductase inhibition by finasterine (Proscar) in epithelium and stroma of human benign prostatic hyperplasia. *Steroids* **1994**, *59*, 616-620.
52. Tolman, R.L.; Sahoo, S.P.; Bakshi, R.K.; Gratale, D.; Patel, G.; Patel, S.; Toney, J.; Chang, B.; Harris, G.S. 4-methyl-3-oxo-4-aza-5 α -androst-1-ene-17 β -aryl-carboxamides: an approach to combined androgen blockade. *J. Steroid Biochem. Mol. Biol.* **1997**, *60*, 303-309.

53. Njar, V.C.O.; Kato, K.; Nnane, I.P.; Grigoryev, D.N.; Long, B.J.; Brodie, A.M.H. Novel 17-azolyl steroids, potent inhibitors of human cytochrome 17 α -hydroxylase-C_{17,20}-lyase (P450_{17 α}): potential agents for the treatment of prostate cancer. *J. Med. Chem.* **1998**, *41*, 902-912.
54. Ling, Y.; Li, J.; Liu, Y.; Kato, K.; Klus, G.T.; Brodie, A. 17-imidazolyl, pyrazolyl, and isoxazolyl androstene derivatives. Novel steroidal inhibitors of human cytochrome C_{17,20}-lyase (P450_{17 α}). *J. Med. Chem.* **1997**, *40*, 297-3304.
55. Thiboutot, D.; Bayne, E.; Thorne, J. Gilligard, K.; Flanagan, J.; Shao, Q.; Light, J.; Helm, K. Immunolocalization of 5 α -reductase isozyme in acne lesions and normal skin. *Arch. Dermatol.* **2000**, *136*, 1125-1129.
56. Guara, A.; Occhiato, E.G.; Banzo, G.; Conti, A.; Serio, M. 5 α -reductase inhibitors, chemical and clinical models. *Steroids* **1998**, *63*, 355-361.
57. Guarna, A.; Mechetti, F.; Occhiato, E.G.; Scarpi, D. Benzo[c]-quinolin-3-ones: a novel class of potent and selective nonsteroidal inhibitors of human 5 α -reductase 1. *J. Med. Chem.* **2000**, *43*, 3718-3735.
58. Hogan, D.J.; Chamberlain, M. Male pattern baldness. *South. Med. J.* **2000**, *93*, 657-662.
59. Russell, D.W.; Wilson, J.D. Steroid 5 α -reductase: two genes/two enzymes. *Annu. Rev. Biochem.* **1994**, *63*, 25-61.
60. Frye, S.V.; Haffner, C.D.; Maloney, P.R.; Hiner, R.N.; Dorsey, G.F.; Noe, R.A.; Unwalla, R.J.; Batchelor, K.W.; Bramson, H.N.; Stuart, J.D.; Schweider, S.L.; van Arnold, J.; Bickett, D.M.; Moss, M.L.; Tian, G.; Lee, F.W.; Tippin, T.K. James, M.K.; Grizzle, M.K.; Long, J.E.; Croom, D.K. Structure-activity relationships for inhibition of Type 1 and 2 human 5 α -reductase and human adrenal 3 β -hydroxy- Δ^5 -steroid dehydrogenase/3-keto- Δ^5 -steroid isomerase by 6-azaandrost-4-en-3-ones: optimization of C17 substituent. *J. Med. Chem.* **1995**, *38*, 2621-2627.
61. Frye, S.V.; Haffner, C.D.; Maloney, P.R.; Mook, R.A.; Dorsey, G.F.; Hiner, R.N.; Cribbs, C.M.; Wheeler, T.N.; Ray, J.A.; Andrews, R.C.; Batchelor, K.W.; Bramson, H.N.; Stuart, J.D.; Schweiker, S.L.; van Arnold, J.; Croom, S.; Bickett, D.M.; Moss, M.L.; Tian, G.; Unwalla, R.J.; Lee, F.W.; Tippin, T.K.; James, M.K.; Grizzle, M.K.; Long, J.E.; Schuster, S.V. 6-azasteroids: structure-activity relationships for inhibition of 5 α -reductase and 3 β -hydroxy- Δ^5 -steroid dehydrogenase/3-keto- Δ^5 -steroid isomerase. *J. Med. Chem.* **1994**, *37*, 2352-2360.

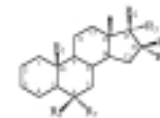
62. Kurup, A.; Garg, R.; Hansch, C. Comparative QSAR analysis of 5 α -reductase inhibitors. *Chem. Rev.* **2000**, *100*, 909-924.
63. Foye, W.O. *Principles of medicinal chemistry*, 3rd ed.; Lea & Febiger: Philadelphia, 1989.
64. McCluskey, P.J.; Towler, H.M.; Lightman, S. Management of chronic uveitis. *S. Br. Med. J.* **2000**, *320*, 555-558.
65. Slavkin, H.C. The body's skin frontier and the challenges of wound healing: keloids. *J. Am. Dental Assoc.* **2000**, *131*, 362-365.
66. Rampton, D.S. Management of Crohn's disease. *Br. Med. J.* **1999**, *319*, 1480-1485.
67. Baud, O.; Foix-L'Hélias, L.; Kamiski, M.; Audibert, F.; Jarreau, P.; Papiernick, E.; huon, C.; Lepercq, J.; Dehan, M.; Lacaze-Masmonteil, T. Antenatal glucocorticoid treatment and cystic periventricular leukomalacia in very premature infants. *New Engl. J. Med.* **1999**, *341*, 1190-1196.
68. Gregorio, C.; Kier, L.B.; Hall, L.H. QSAR modeling with the electrotopological state indices: corticosteroids. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 557-561.

Table Captions

Table 1.- Structures and binding activities to aromatase for a set of 50 steroids [42].

Table 2.- Structures, binding affinities ($\log K_i$) to 5α -reductase, and selectivities ($\log S_i$) to 3-BHSD [43].

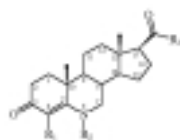
Table 3.- Structures and binding affinities to corticosteroid binding globulin for a set of 31 steroids [44,45].

Table 1.- Structures and binding activities to aromatase for a set of 50 steroids [42].

#	R ₁	R ₂	R ₃	R ₄	R ₅	R ₆	R ₇	=O in 3	=O in 7	Other	Activity
1	CH ₂ OH	=O		H	H		H	No	Yes	Δ^5	-2.92
2	CH ₂ OH	OH	H	H	H		H	No	Yes	Δ^5	-3.54
3	CHO	=O		H	H		H	No	Yes	Δ^5	-3.00
4	H	=O		H	H		H	No	Yes	Δ^5	-3.26
5	CH ₃	OH	H	H	H		H	No	Yes	Δ^5	-2.62
6	CH ₂ OH	=O		H	H		H	No	Yes	Δ^3, Δ^5	-3.06
7	CHO	=O		H	H		H	No	Yes	Δ^3, Δ^5	-2.14
8	H	=O		H	H		H	No	Yes	Δ^3, Δ^5	-2.36
9	CH ₂ OH	=O		H	H		H	No	No	Δ^5	-1.89
10	CH ₂ OH	OH	H	H	H		H	No	No	Δ^5	-2.88
11	CHO	=O		H	H		H	No	No	Δ^5	-2.03
12	CH ₃	=O		H	H		H	No	No	Δ^5	-0.97
13	CH ₃	=O		H	Br		H	No	No	Δ^5	-2.93
14	CH ₃	=O		H	H		H	No	Yes	Δ^5	-1.28
15	CH ₃	=O		H	H		H	No	Yes	Δ^3, Δ^5	-1.23
16	CH ₃	OH	H	H	H		H	No	Yes	Δ^3, Δ^5	-2.61
17	CH ₃	OH	H	H	H		H	No	No	Δ^5	-2.36

18	CH ₃	=O	H	H	H	No	No	Δ^3, Δ^5	-0.65
19	CH ₃	OH	H	H	H	No	No	Δ^3, Δ^5	-2.19
20	H	=O	H	H	H	Yes	No	Δ^4	-1.03
21	CH ₃	=O	H	H	H	No	No	Δ^4	0.00
22	CH ₂ OH	=O	H	H	H	No	No	Δ^4	0.46
23	CH ₂ OH	=O	H	H	H	Yes	No	Δ^4	-0.84
24	CH ₃	=O	H	H	=O	Yes	No	Δ^4	0.15
25	CH ₃	=O	H	H	see Other	Yes	No	$\Delta^4, 6,7-\alpha\text{-CF}_2$	-0.13
26	CH ₃	=O	H	H	see Other	No	No	$\Delta^4, 6,7-\alpha\text{-CH}_2$	0.87
27	CH ₃	OH	H	H	see Other	No	No	$\Delta^4, 6,7-\alpha\text{-CH}_2$	-0.51
28	CH ₃	OH	H	H	H	No	No	Δ^4	-1.35
29	CH ₂ OH	OH	H	H	H	No	No	Δ^4	-0.67
30	CH ₂ OC(O)CH ₃	=O	H	H	H	No	No	Δ^4	-0.89
31	CH ₃	=O	H	Br	H	No	No	Δ^4	-0.79
32	CH ₃	=O	H	H	H	No	No	Δ^4	-1.09
33	CF ₃	=O	H	H	H	No	No	Δ^4	-1.08
34	CH ₃	=O	H	H	CH ₃	Yes	No	Δ^4	0.56
35	CH ₃	=O	H	H	H	Yes	No	Δ^4	0.87
35	CH ₃	=O	H	H	CH ₂ CH ₃	Yes	No	Δ^4	1.56
37	CH ₃	=O	H	H	H	Yes	No	Δ^4	0.94
38	CH ₃	=O	H	H	(CH ₂) ₂ CH ₃	Yes	No	Δ^4	0.94
39	CH ₃	=O	H	H	H	Yes	No	Δ^4	0.78
40	CH ₃	=O	H	H	(CH ₂) ₃ CH ₃	Yes	No	Δ^4	0.65
41	CH ₃	=O	H	H	H	Yes	No	Δ^4	0.53
42	CH ₃	=O	H	H	CH(CH ₃) ₂	Yes	No	Δ^4	0.21
43	CH ₃	=O	H	H	H	Yes	No	Δ^4	0.04
44	CH ₃	=O	H	H	C ₆ H ₅	Yes	No	Δ^4	-0.04
45	CH ₃	=O	H	H	H	Yes	No	Δ^4	0.24

46	CH ₃	=O	H	H	CH ₂ C ₆ H ₅	H	Yes	No	Δ^4	-0.24
47	CH ₃	=O	H	H	H	CH ₂ C ₆ H ₅	Yes	No	Δ^4	0.61
48	CH ₃	=O	H	H	CH=CH ₂	H	Yes	No	Δ^4	0.91
49	CH ₃	=O	H	H	H	CH=CH ₂	Yes	No	Δ^4	-0.32
50	CH ₃	=O	H	H	=CHCH ₃		Yes	No	Δ^4	0.96

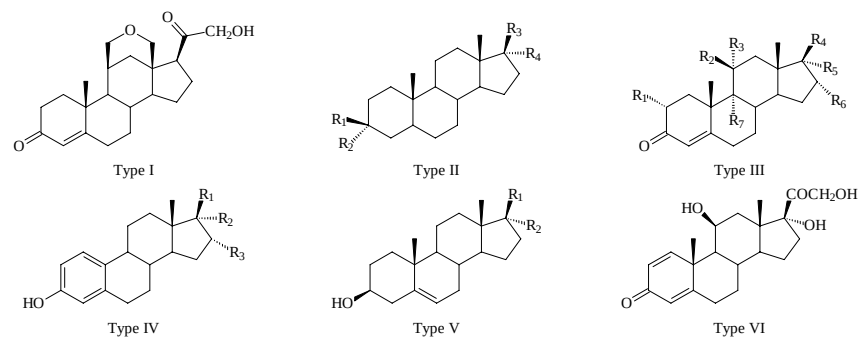
Table 2.- Structures, binding affinities (log K_i) to 5 α -reductase, and selectivities (log S_i) to 3-BHSD [43].

#	R ₁	R ₂	R ₃	Other	log K_i	log S_i
1	H	H	NH ₂		2.38	-1.38
2	H	H	2-C(CH ₃) ₃ -Ph-NH		1.43	0.50
3	H	H	5-Cl,2-C(CH ₃) ₃ -Ph-NH		0.88	2.01
4	H	H	5-Br,2-C(CH ₃) ₃ -Ph-NH		0.62	2.29
5	H	H	4-Br,2-C(CH ₃) ₃ -Ph-NH		0.72	1.87
6	H	H	2,5-bis(C(CH ₃) ₃)-Ph-NH		0.70	2.00
7	H	H	2-C(CH ₃) ₃ ,5-Ph-Ph-NH		0.66	2.26
8	H	H	2-C(CH ₃) ₃ ,5-CF ₃ -Ph-NH		0.94	2.26
9	H	H	2-C(CH ₃) ₃ ,5-(4-Cl-Ph)-Ph-NH		0.53	2.55
10	H	H	2-C(CH ₃) ₃ ,5-(4-C(CH ₃) ₃ -Ph)-Ph-NH		0.11	3.84
11	H	H	2,5-bis(CF ₃)-Ph-NH		0.60	1.88
12	H	H	2-(4-C(CH ₃) ₃ -Ph),5-CF ₃ -Ph-NH		3.04	-1.76
13	H	H	3,5-bis(CF ₃)-Ph-NH		1.41	0.35
14	H	H	3,5-bis(C(CH ₃) ₃)-Ph-NH		0.90	-0.01
15	H	H	4-Cl-Ph-C(CH ₂) ₄ -NH		0.83	2.37
16	H	H	4-Cl-Ph-C(CH ₂) ₅ -NH		0.49	2.76

17	H	H	2,4-di-Cl-Ph-C(CH ₂) ₂ -NH		0.83	1.29
18	H	H	4-C(CH ₃) ₃ -Ph-C(CH ₂) ₄ -NH		0.48	2.70
19	H	H	4-C(CH ₃) ₃ -Ph-C(CH ₂) ₅ -NH		0.18	3.17
20	H	H	4-C(CH ₃) ₃ -Ph-C(CH ₂) ₆ -NH		0.56	2.62
21	CH ₃	H	2-C(CH ₃) ₃ ,5-CF ₃ -Ph-NH		-0.30	2.60
22	Cl	H	2-C(CH ₃) ₃ ,5-CF ₃ -Ph-NH		-0.22	2.79
23	CH ₃	H	2,5-bis(CF ₃)-Ph-NH		-0.70	1.98
24	Cl	H	2,5-bis(CF ₃)-Ph-NH		-0.70	2.98
25	CH ₃	H	(1-(4-Cl-Ph)c-pentyl)NH		-0.52	2.73
26	Cl	H	(1-(4-Cl-Ph)c-pentyl)NH		-0.22	2.91
27	H	H	N(CH ₂ CH ₃) ₂		2.88	-1.10
28	H	H	N(CH ₂ CH ₃) ₂	Δ^1	3.76	-1.86
29	H	H	N(CH ₂ CH ₃) ₂	1,2- α -CH ₂	4.15	-2.17
30	H	H	N(CH ₂ CH ₃) ₂	2- α , β -CH ₃	3.54	-0.84
31	H	COCH ₃	N(CH ₂ CH ₃) ₂		4.60	-0.07
32	H	CN	N(CH ₂ CH ₃) ₂		3.92	-0.74
33	H	CH ₂ CO ₂ H	N(CH ₂ CH ₃) ₂		4.00	-1.74
34	H	CH ₃	N(CH ₂ CH ₃) ₂		2.26	-1.70
35	H	CH ₂ CH ₃	N(CH ₂ CH ₃) ₂		3.11	-1.42
36	H	CH ₂ CH ₂ CH ₃	N(CH ₂ CH ₃) ₂		4.38	-1.66
37	H	i-Pr	N(CH ₂ CH ₃) ₂		3.53	-0.49
38	H	(CH ₂) ₃ CH ₃	N(CH ₂ CH ₃) ₂		3.89	-0.34
39	H	(CH ₂) ₅ CH ₃	N(CH ₂ CH ₃) ₂		3.77	0.27
40	H	CH ₂ Ph	N(CH ₂ CH ₃) ₂		4.00	0.20
41	H	CH ₃	N(CH ₂ CH ₃) ₂	Δ^1	3.11	-1.77
42	H	CH ₃	N(CH ₂ CH ₃) ₂		4.00	-3.00
43	Cl	H	N(CH ₂ CH ₃) ₂		1.71	-0.28
44	Br	H	N(CH ₂ CH ₃) ₂		1.99	-0.87
45	I	H	N(CH ₂ CH ₃) ₂		2.84	-1.66
46	CH ₂ N(CH ₃) ₂	H	N(CH ₂ CH ₃) ₂		4.00	-0.92

47	CH ₃	H	N(CH ₂ CH ₃) ₂		1.60	-0.35
48	CH ₃	H	N(CH ₂ CH ₃) ₂	Δ ¹	2.45	-0.49
49	CH ₂ CH ₃	H	N(CH ₂ CH ₃) ₂		3.61	-1.94
50	CH ₃	CH ₃	N(CH ₂ CH ₃) ₂		1.91	-0.87
51	H	H	NHC(CH ₃) ₃		2.91	-0.74
52	H	H	NHC(CH ₃) ₃	Δ ¹	3.38	-0.04
53	H	H	NHC(CH ₃) ₃	1,2-α-CH ₂	4.69	-2.21
54	H	CH ₃	NHC(CH ₃) ₃		1.94	-1.12
55	H	CH ₃	NHC(CH ₃) ₃		3.15	-1.35
56	CH ₃	H	NHC(CH ₃) ₃		1.08	-0.12
57	CH ₃	CH ₃	NHC(CH ₃) ₃		1.78	-0.42
58	H	H	i-But		0.95	0.05
59	H	CH ₃	i-But		0.49	-0.45
60	Br	H	i-But		0.51	-0.14
61	CH ₃	H	i-But		-0.40	0.48
62	H	H	NH-1-Ad		1.04	0.85
63	H	H	NH-1-Ad	Δ ¹	1.87	1.13
64	H	CH ₃	NH-1-Ad		0.93	-0.50
65	H	CH ₃	NH-1-Ad	Δ ¹	2.45	-0.65
66	Br	H	NH-1-Ad		0.65	0.46
67	CH ₃	H	NH-1-Ad		0.04	0.91
68	Br	CH ₃	NH-1-Ad		0.93	-0.02
69	CH ₃	CH ₃	NH-1-Ad		0.79	0.46
70	H	H	NHCHPh ₂		1.48	0.70
71	H	H	NHCHPh ₂	Δ ¹	0.68	1.31
72	H	CH ₃	NHCHPh ₂		0.81	0.24
73	H	(CH ₂) ₂ CH ₃	NHCHPh ₂	Δ ¹	2.53	-0.28
74	CH ₃	H	NHCHPh ₂		0.56	0.70
75	H	H	OCH ₃		2.18	-1.10
76	H	H	O-2-Ad		0.84	1.42

77	H	H	N(CH ₃)OCH ₃		3.36	-1.66
78	H	H	Piperazine		3.93	-1.59
79	H	H	Morpholine		3.34	-1.06
80	H	H	Thiomorpholine		2.76	-1.26
81	H	H	Piperidine		1.93	0.11
82	H	H	NHCH(4-F-Ph) ₂		1.30	1.20
83	H	H	NHCH(4-Cl-Ph) ₂		1.30	1.41
84	H	H	NHNPh ₂		1.15	1.06
85	H	H	N(OH)C(CH ₃) ₃		1.58	0.21
86	H	H	NHCH(c-Hex) ₂		1.30	2.30
87	H	H	NHCPh ₃		0.91	1.09
88	H	H	(CH ₂) ₂ CH ₃		1.08	-0.04
89	H	H	(CH ₂) ₇ CH ₃		0.00	0.88
90	H	H	CH ₂ -c-Hex		0.60	0.60
91	H	H	2,6-difluorophenyl		1.59	-0.55
92	H	H	1-naphthyl		1.18	-0.28
93	H	H	2,4,6-triisopropylphenyl		2.45	0.83

Table 3.- Structures and binding affinities to corticosteroid binding globulin for a set of 31 steroids [44,45].

#	Type	R ₁	R ₂	R ₃	R ₄	R ₅	R ₆	R ₇	pK
1	I								6.279
2	II	OH	H	OH	H				5.000
3	V	OH	H						5.000
4	III	H	H	H	=O		H	H	5.763

5	II	H	OH		=O					5.613
6	III	H	OH	H	COCH ₂ OH	H	H	H		7.881
7	III	H	OH	H	COCH ₂ OH	OH	H	H		7.881
8	III	H	OH	=O	COCH ₂ OH	OH	H	H		6.892
9	V	=O								5.000
10	III	H	H	H	COCH ₂ OH	H	H	H		7.653
11	III	H	H	H	COCH ₂ OH	OH	H	H		7.881
12	II	=O		OH	H					5.919
13	IV	OH	H	H						5.000
14	IV	OH	H	OH						5.000
15	IV	=O		H						5.000
16	II	OH	H		=O					5.225
17	V	COCH ₃	H							5.225
18	V	COCH ₃	OH							5.000
19	III	H	H	H	COCH ₃	H	H	H		7.380
20	III	H	H	H	COCH ₃	OH	H	H		7.740
21	III	H	H	H	OH	H	H	H		6.724
22	VI									7.512
23	III	H	OH	H	COCH ₂ OCOCH ₃	OH	H	H		7.553
24	III	H	=O		COCH ₃	H	H	H		6.779
25	III	H	H	OH	COCH ₂ OH	H	H	H		7.200
26	III	H	H	H	OH	H	H	H		6.144
27	III	H	H	H	COCH ₃	OH	OH	H		6.247
28	III	H	H	H	COCH ₃	H	CH ₃	H		7.120
29	III	H	H	H	COCH ₃	H	H	H		6.817
30	III	CH ₃	OH	H	COCH ₂ OH	OH	H	H		7.688
31	III	CH ₃	OH	H	COCH ₂ OH	OH	H	F		5.797

Figure 1.- Pairwise superposition of two steroids according to the TGSA superposition procedure.

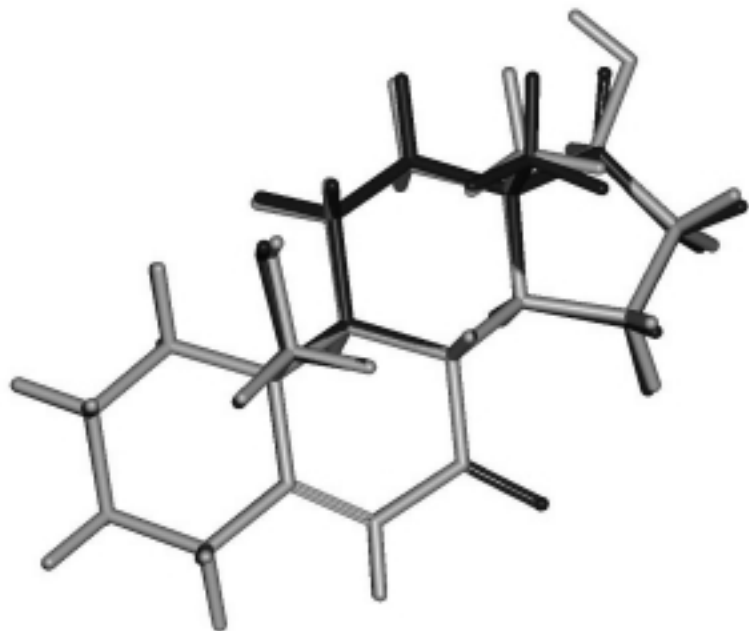


Figure 2.- Evolution of r^2 (continuous line) and q^2 (dashed line) versus the number of PLS factors involved in the construction of the QSAR model for a set of 50 steroids [42].

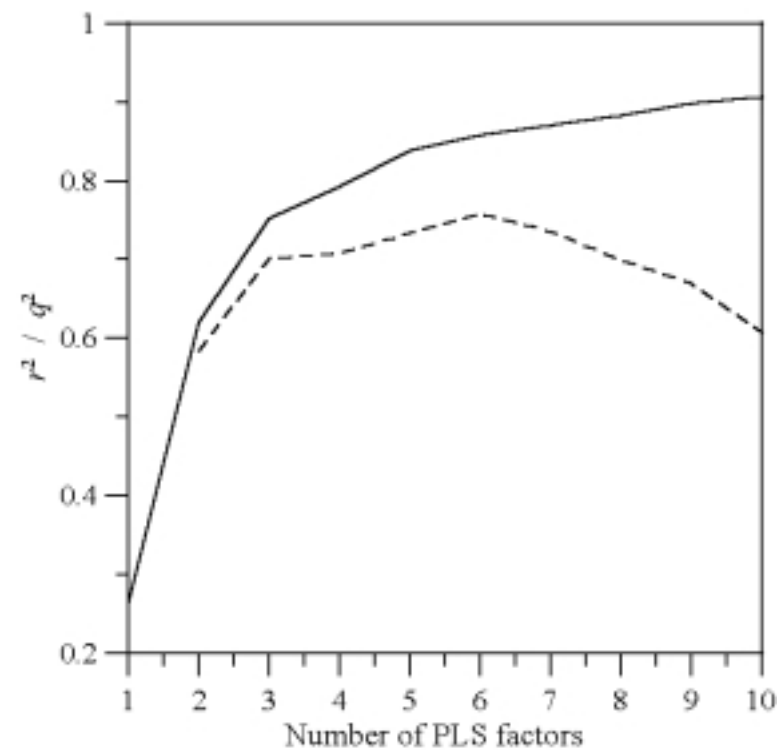


Figure 3.- Experimental versus predicted binding activity for a set of 50 steroids.

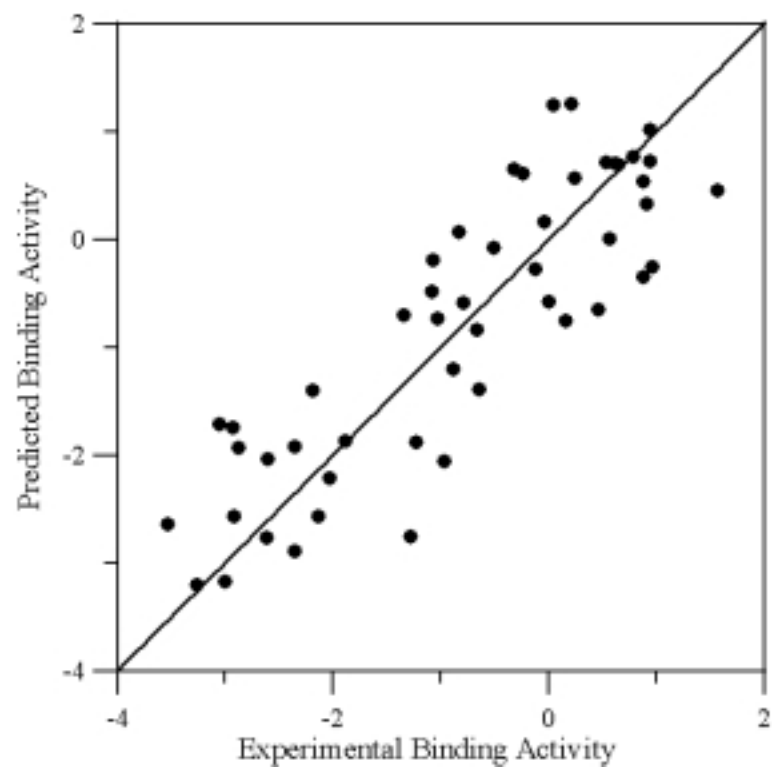


Figure 4.- Random test results after permuting the binding activity values for a set of 50 steroids 1000 times. The real value is labeled with a cross.

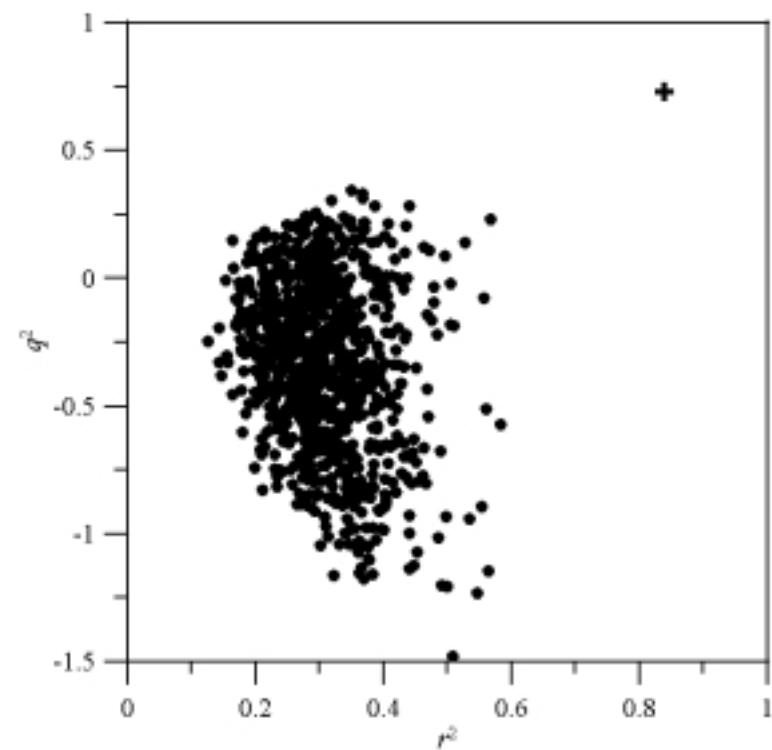


Figure 5.- Experimental versus predicted $\log K_i$ for a set of 93 azasteroids.

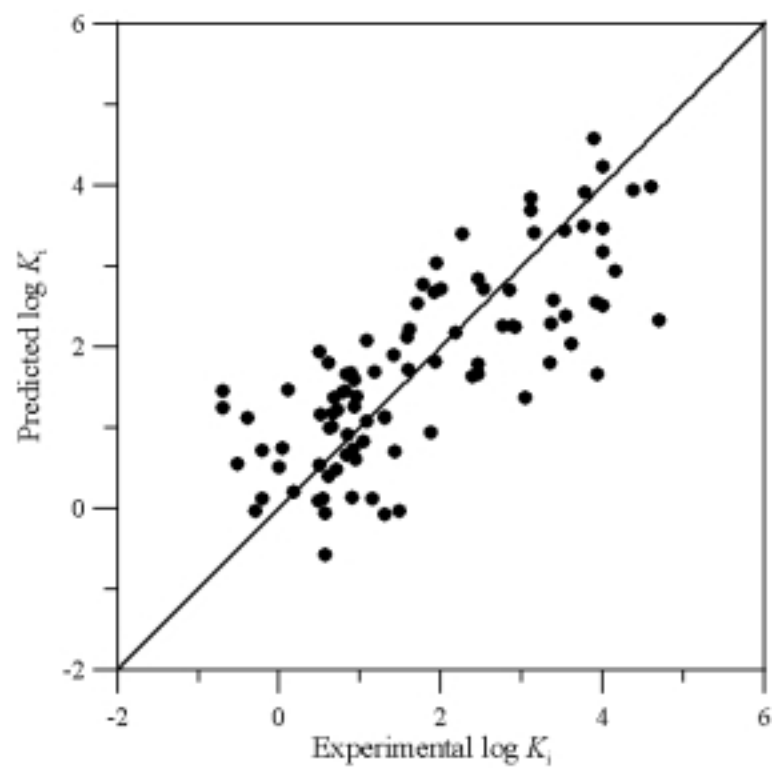


Figure 6.- Experimental versus predicted $\log S_i$ for a set of 93 azasteroids.

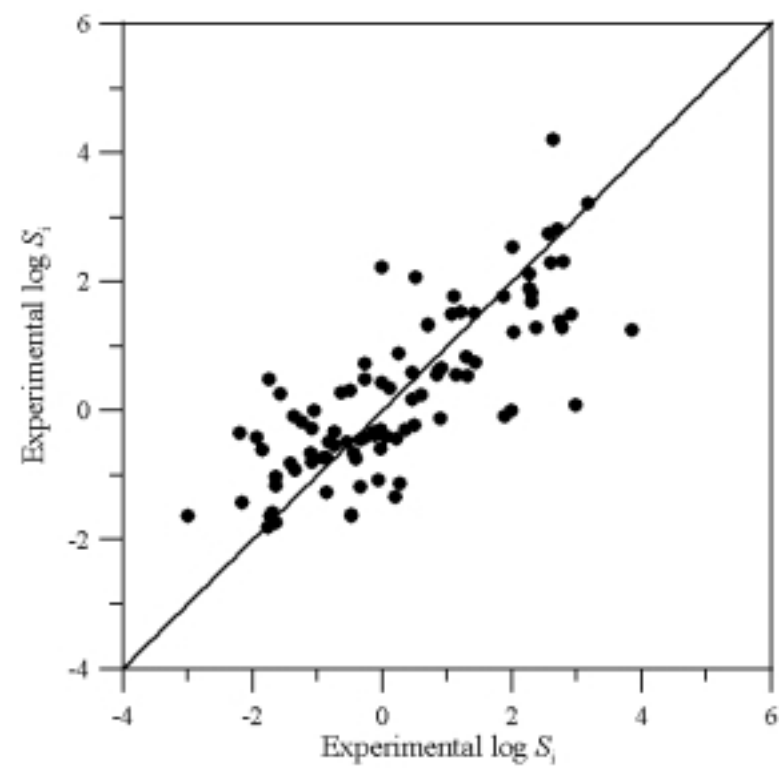
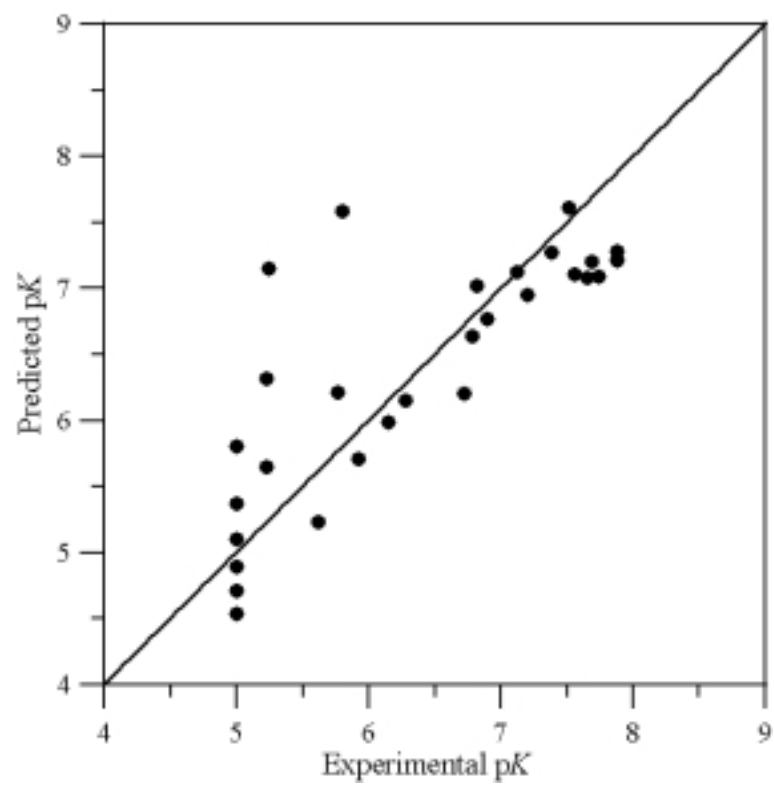


Figure 7.- Experimental versus predicted pK for a set of 31 steroids.



7 Consideracions Finals

Aquest últim capítol vol donar una pinzellada final al treball, intentant resumir de manera global els resultats obtinguts en els àmbits de recerca estudiats, així com comentar altres estudis que s’han realitzat a banda d’aquests, on s’ha col·laborat de manera menys intensiva. Com a colofó d’aquest capítol es llisten una sèrie d’objectius de futur, en els quals podria ser interessant fer estudis, així com el llistat total de publicacions generades al llarg de la realització d’aquesta tesi doctoral.

7.1 *Conclusions Generals*

Apart de les conclusions específiques de cada capítol, així com les exposades en els articles adjunts, es poden extreure les següents inferències d’aquest treball:

- I. S’ha contribuït en el desenvolupament, tant teòric com computacional, de la semblança molecular quàntica, així aquest ha estat aplicat satisfactòriament en diferents àmbits.
- II. S’han codificat aquests nous avenços en forma de programes informàtics capaços de dur a terme els càlculs de forma senzilla i transparent al possible usuari, alhora que s’ha treballat en la seva optimització en diferents plataformes computacionals.
- III. Les diferents contribucions en aquest desenvolupament, com són l’energia d’interacció electró-electró, o bé l’aplicació de densitats de Forats de Fermi, amplien, alhora que enriqueixen, la teoria inicial de la semblança quàntica, permetent la seva expansió en nous àmbits.
- IV. La possibilitat de realitzar el càlcul de semblança entre diverses molècules emprant l’algorisme de sobreposició topo-geomètrica, enlloc de la regla de la màxima semblança emprada anteriorment, permetent ampliar el rang de mida de molècules a estudiar. Aquest fet, unit a l’ús de funcions ajustades i de diferents llindars en el càlcul,

i en conjunció amb l'abaratiment dels recursos computacionals permet abordar l'estudi de sistemes compostats per un nombre elevat de molècules o de macromolècules.

- V. S'ha demostrat l'aplicabilitat de la semblança quàntica en el desenvolupament de descriptors per a estudis de correlació estructura-activitat (QSAR), obtenint-se resultats valuosos per a sistemes moleculars d'interès divers emprant metodologies estadístiques relativament senzilles. Tot i que en alguns casos els resultats presents a la literatura arriben a ser millors, s'ha de considerar que el protocol QSAR exposat dona poques opcions de càlcul, tals com la tria de l'operador, l'escalat, i de manera aliena a la semblança molecular, la metodologia estadística a utilitzar. D'aquesta manera, la construcció de models QSAR basats en semblança molecular es troba en sí mateixa limitada i auto-continguda. A banda, en aquells sistemes on no s'aconsegueixin resultats acceptables, o requereixin una major precisió, els descriptors creats a partir de semblança poden constituir un bon punt de partida o esdevenir un excel·lent complement per a d'altres descriptors o procediments.
- VI. De manera paral·lela, s'ha treballat exhaustivament en la presentació dels mateixos resultats, així com en la facilitat per a la seva obtenció i visualització. Aquest punt, de major importància en els apartats de representacions gràfiques de densitats i sobreposició molecular, proporciona una major i més ràpida comprensió dels resultats, així com obre diverses aplicacions didàctiques o docents.
- VII. Referint-se a l'últim punt, s'ha intentat aplicar els últims avenços en interfícies gràfiques, així com unir els resultats obtinguts per a ser l'entrada a programari divers, a fi de millorar-ne l'aspecte, quedant aquest punt en continu desenvolupament.

7.2 Altres Aspectes Estudiats

Conjuntament amb altres membres del grup a l'Institut i amb d'altres investigadors durant l'estada de recerca a Praga, i simultàniament a les meves línies, s'ha col·laborat de manera menys intensiva en d'altres camps, els quals cito manera resumida a continuació:

- o Classificació de compostos segons la teoria de les facetes (de l'anglès, *Facet Theory*), amb la qual, l'activitat (p.e. biològica) de sèries moleculars podia ser estimada de manera discreta segons la regió de l'espai, definida en base a una tècnica estadística. Veure contribució #04 de l'apartat 7.4.

- o S'ha realitzat una incursió dins de la Topologia (#31), incloent paràmetres d'origen quàntic alhora de calcular les matrius topològiques, enlloc de simplement distàncies i connectivitats.
- o Estudis sobre la natura dels enllaços en diferents compostos d'interès divers. Per exemple, es va realitzar un estudi sobre el N_5^+ (#13) emprant anàlisi de poblacions i Forats de Fermi. També es van realitzar estudis de compostos organolitiats (#26), així com un anàlisi comparatiu de compostos hipervalents (#29), tals com el CLi_6 i el SF_6 emprant orbitals atòmics derivats de la teoria dels Àtoms en Molècules (AIM, de l'anglès *Atoms in Molecules*) de Bader.
- o S'han fet alguns estudis amb xarxes neurals en QSAR, prenent com a base, entre d'altres, descriptors derivats de càlculs de semblança (#17,#24 i #32).
- o S'han estudiat els efectes inductius de diferents substituents en la reactivitat química (#20), el qual va inspirar més tard els càlculs d'autosemblança basats en densitats de Forats de Fermi (#21,#29,#33).

7.3 Objectius de Futur

Al llarg d'aquest treball s'han desenvolupat diferents conceptes, les aplicacions dels quals no han estat estudiades amb suficient profunditat, així com han sorgit noves idees i nou programari amb diferents i prometedores possibilitats. Així doncs, la intenció d'aquest apartat és fer un petit llistat de treballs als que podria valer la pena dedicar-hi esforços en el futur:

- o Unificació de programari per a aplicacions relacionades, així com dotar-lo d'una interfície gràfica amigable a l'usuari. D'aquesta manera, la utilització i capacitat de distribució del programari desenvolupat esdevé més eficient per a dues raons: 1) no cal executar un programa específic per a cada *sub-tasca* d'un estudi, amb l'estalvi de temps que comporta en la verificació dels resultats parcials de cada programa i execució del següent; 2) comoditat i seguretat en l'execució del programa. La comoditat ve donada per la possibilitat de seleccionar les dades a partir d'un sistema de menús enlloc de *barallar-se* amb fitxers de comandes o dades d'entrada de consola, mentre que la seguretat prové de la incapacitat (sempre que la captura d'errors estigui ben implementada) de seleccionar opcions conflictives entre elles o mostrar missatges d'error, permetent modificar les opcions entrades. Aquesta possibilitat avarca tant la creació de contorns de densitat, sobreposició molecular i aplicació de la semblança en QSAR.

- o Paral·lelització de codi, dins d'aquelles arquitectures computacionals que ho permeten, per a execució simultània o càlcul en paral·lel. Aquest aspecte, molt estudiat a d'altres grups de recerca, permet ampliar el rang d'estudi al disposar de més potència de càlcul. Tot i enunciat molt ràpid, el camp, així com l'ús eficient de la paral·lelització requereix un estudi profund del problema concret a tractar, així com de les seves dependències de càlcul.
- o Realització de nous ajustos a diferents conjunts de funcions de base dins d'altres metodologies, com per exemple, ajustos a paràmetres del mètode semiempíric AM1, d'aquesta manera el mètode de optimització de geometria i el modelatge molecular esdevenen mútuament coherents.
- o Inclusió de flexibilitat en la sobreposició molecular, permetent la rotació d'enllaços i lleugeres variacions en els angles i distàncies d'enllaç. D'aquesta manera, les sobreposicions obtingudes, així com les mesures de semblança associades podrien esdevenir més verosímils.
- o Profunditzar intensivament en la interpretació dels models QSAR basats en semblança, així com trobar explicacions amb sentit químic als resultats.

7.4 Llistat Total de Publicacions

En aquest apartat final es recull un llistat del total de les publicacions que s'han generat al llarg d'aquest treball, bé pertanyin directament als camps de recerca que s'han tractat, bé s'hi hagi contribuït d'una forma més indirecte. Les publicacions es llisten en ordre (més o menys) cronològic, incloent els autors, títols, referència o estat de la publicació segons el cas, i abstract de la mateixa.

#01

Autors:	Gironés, X. ; Amat, L.; and Carbó-Dorca, R.
Títol:	A comparative study of isodensity surfaces using ab initio and ASA density functions.
Referència / estat:	J. Mol. Graph. & Mod. 16 (1998) 190-196
Abstract:	In this article, we report a visual comparison between several of the available methods for constructing electronic density functions. The density forms studied include ab initio atomic shell approximation and promolecular densities. A graphical comparison is made for six different molecules at different levels of density function values. The differences between the various density functions are analyzed by considering a molecular quantum self-similarity measure and the required computational time for all molecules at all computation levels considered.

#02

Autors: **Gironés, X.**; Amat, L.; and Carbó-Dorca, R.

Títol: Descripció de propietats moleculars i activitats biològiques emprant l'energia de repulsió electró-electró.

Referència / estat: Scientia Gerundensis 24 (1999) 197-208

En aquest treball es descriu l'ús de les mesures de semblança molecular quàntica (MSMQ) per a caracteritzar propietats i activitats biològiques moleculars, i definir descriptors emprables per a construir models QSAR i QSPR. L'estudi que es presenta consisteix en la continuació d'un treball recent, on es descriuen relacions entre el paràmetre log P i MSMQ, donant així una alternativa a aquest paràmetre hidrofòbic empíric. L'actual contribució presenta una nova mesura, capaç d'extendre l'ús de les MSMQ, que consisteix en

Abstract: l'energia de repulsió electró-electró ($\langle V_{ee} \rangle$). Aquest valor, disponible normalment a partir de programari de química quàntica, considera la molècula com a una sola entitat, i no cal recórrer a l'ús de contribucions de fragments. La metodologia s'ha aplicat a cinc tipus diferents de compostos on diferents propietats moleculars i activitats biològiques s'han correlacionat amb $\langle V_{ee} \rangle$ com a únic descriptor molecular. En tots els casos estudiats, s'han obtingut correlacions satisfactòries.

#03

Autors: **Gironés, X.**; Amat, L.; and Carbó-Dorca, R.

Títol: Using molecular quantum similarity measures as descriptors in quantitative structure-toxicity relationships.

Referència / estat: SAR & QSAR Envir. Research 10 (1999) 545-556

In this paper molecular quantum similarity measures (MQSM) are used to describe molecular toxicity and to construct Quantitative Structure-Toxicity Relationships (QSTR) models. This study continues the recently described relationships between MQSM and log P values, which permits to use the theoretical MQSM as an alternative to the empirical hydrophobic parameter in QSPR studies. In addition a new type of MQSM is presented in this work: it is based on the expectation value of electron-electron repulsion energy. The molecular properties studied here, as application examples are aquatic toxicity, toxicology on Bacteria and inhibition of a macromolecule employing four different molecular sets.

Abstract:

#04

Autors: Robert, D.; **Gironés, X.**, and Carbó-Dorca, R.

Títol: Facet diagrams for quantum similarity data.

Referència / estat: J. Comp.-Aided Molec. Design 13 (1999), 597-610

The objective of this work is to demonstrate that an appropriate treatment of quantum similarity matrices can reveal hidden data grouping related to relevant structural features and even to biological properties of interest. Classical scaling is used here to extract the information contained in the similarity relationships between the overall Cramer steroid set. Facet Theory is invoked to relate, in a qualitative way, the spatial regions to structural characteristics as well as to the CBG binding affinity.

Abstract:

#05

Autors: **Gironés, X.**; Amat, L.; Carbó-Dorca, R.; and Robert, D.

Títol: Use of electron-electron repulsion energy as a molecular descriptor in QSAR and QSPR studies.

Referència / estat: J. Comp.-Aided Molec. Design 14 (2000) 477-485

Electron-Electron repulsion energy ($\langle V_{ee} \rangle$) is presented as a new molecular descriptor to be employed in QSAR and QSPR studies. This energy parameter, fundamentally connected to molecular quantum similarity measures (MQSM), can be considered as a complement to steric and electronic parameters in description of molecular properties and biological responses of organic compounds. The present strategy considers the molecule as a whole, thus there is no need to employ contributions of isolated fragments as in many approximated calculations. The procedure has been tested in a widespread set of molecules: alcohols, alkanamides, indole derivatives and 1-alkylimidazoles. Molecular properties, as well as toxicity, are correlated using as a parameter, and extensions to the method are given for handling difficult systems. In almost all studied cases, satisfactory linear relationships were finally obtained.

#06

Autors: Carbó-Dorca, R.; Amat, L.; Besalu, E.; **Gironés, X.**; and Robert, D.

Títol: Quantum Mechanical Origin of QSAR: Theory and Applications.

Referència / estat: J. Mol. Struct. Theochem 504 (2000) 181-228

In this paper, search is carried out on how to develop the formalism where Quantum Similarity Measures (QSM) become a natural product of the theoretical framework. This fact is latter used to establish a fundamental connection between Quantum Theory and QSAR, which is analysed in turn within the realm of discrete quantum chemistry. In order to perform such task several theoretical tools are revised in a previous step. The first section is devoted to construct the concept of Tagged Set. Next, the definition of Quantum Object (QO) is clarified by means of Quantum Theory background ideas and the previous Tagged Set formalism. In the definition of QO, Density Functions (DF) play a fundamental role and a possible simplified mathematical picture is presented. On the road to prepare the problem solving tools, Convex Sets result to be prominent, and the notion of Vector Semispace appears as a consequence. The Transformation Rule, a device to connect Wavefunctions with DF, is defined in a new step. Various products of this preliminary discussion are described, among others the concept of Kinetic Energy and other observable distributions. QSM as a source of discrete representation of molecular structures is made evident in this context. Further theoretical development intends to study discretization, the transformation of infinite dimensional Functional Spaces into n-dimensional ones. This results adds new perspectives to the discrete representation of QO, because: a) Provides a source of new QO descriptors, like a generalisation of scalar product and new similarity indices. b) Describes the QSPR theoretical background enabling the construction of the adequate mathematical tools in order to discuss connected problems: limitations of linear models, tuned-QSAR, p-valued classification of QO, ... c) Allows the construction of spund and general alternatives of Hammett's sigma or log P parameters.

#07

Autors: Carbo-Dorca,R.; Besalú,E.; and **Gironés,X.**

Títol: Extended density functions.

Referència / estat: Adv. Quantum Chem. 38 (2000) 3-63

Density Functions play a fundamental role in the definition of Quantum Theory, due to this they are the basic materials used in order to define Quantum Objects and from this intermediate step, they constitute the support of Quantum Similarity Measures. Here, the connection of Wavefunctions with Extended Density Functions is analysed. Various products of this preliminary discussion are described, among others the concept of Kinetic Energy Distributions.

Abstract: Another discussed set of concepts, directly related with the present paper, is constituted by the Extended Hilbert Space definition, where their vectors are defined as column structures or diagonal matrices, containing both wavefunctions and their gradients. The shapes of new density distributions are described and analysed. All the steps above summarised are completed and illustrated, when possible, with practical application examples and visualisation pictures.

#08

Autors: Robert,D.; **Gironés, X.**; and Carbó-Dorca,R.

Títol: Quantification of the influence of single-point mutations on the Haloalkane dehalogenase activity: a molecular quantum similarity study.

Referència / estat: J. Chem. Inf. Comput. Sci. 40 (2000) 839-846

Controlled modifications in certain protein aminoacid residuals can yield to changes in its function or stability. The examination of the aminoacid structural features and their relation to these changes are made here by using quantum molecular similarity techniques. As a case example, the effect of deliberate mutations in position 172 of the haloalkane dehalogenase enzyme, yielding to variations on the dehalogenation of 1,2-dibromoethane, is discussed qualitatively and quantitatively. Molecular Quantum Similarity is presented

Abstract: here as an appropriate chemical tool to extract the relevant information from a set of homogeneous compounds, namely the aminoacids. A valuable binary classification is obtained only by representing the two-dimensional classical scaling solution using the most predictive principal coordinates as axes. In addition, satisfactory quantitative relationships are found out, comparable to those attained by previous studies on this same data set using other techniques. This kind of analysis provides a consistent, unbiased and homogeneous set of molecular descriptors, and entails a possible alternative to the use of physico-chemical properties as variables.

#09

-
- Autors: Robert,D.; **Gironés,X.**; and Carbó-Dorca,R.
Títol: Molecular Quantum Similarity Measures as descriptors for Quantum QSAR.
Referència / estat: Polycycl. Aromat. Comp. 19 (2000) 51-71
Abstract: Molecular quantum similarity establishes quantitative similarity measures between molecular structures by means of the comparison of their electronic distributions. From this theoretical framework, different molecular descriptors can be derived and used as QSAR parameters. In particular, molecular quantum similarity matrices and the electron-electron interaction energy are examined. A brief exposition of the methodology and the QSAR protocol is given, and a summary of the results obtained using these descriptors is presented.

#10

-
- Autors: Carbó-Dorca, R.; Robert, D.; Amat, L.; **Gironés, X.**; and Besalú, E.
Títol: Molecular Quantum Similarity in QSAR and Drug Design.
Referència / estat: Springer Lecture Notes in Chemistry 73 (2000) Springer (Llibre Monogràfic)
Abstract: The authors introduce the concept of Molecular Quantum Similarity, developed in their laboratory, in a didactic form. The basis of the concept combines quantum theoretical calculations with molecular structure and properties even for large molecules. They give definitions and procedures to compute similarities molecules and provide graphical tools for visualization of sets of molecules as n-dimensional point charts.

#11

-
- Autors: **Gironés, X.**; Gallegos, A.; and Carbó-Dorca,R.
Títol: Modeling Antimalarial Activity: Application of Kinetic Energy Density Quantum Similarity Measures as Descriptors in QSAR.
Referència / estat: J. Chem. Inf. Comp.Sci. 40 (2000) 1400-1407
Abstract: In this work, the application of the recently described Kinetic Energy Density Function in the evaluation of the antimalarial activity is demonstrated. First, this new type of Density Function is briefly presented from its theoretical foundations, and its inclusion in the Molecular Quantum Similarity framework is discussed. The application of Kinetic Energy-based Quantum Similarity Measures to QSAR is presented with 2 molecular sets composed by artemisinin derivatives, in which the 50% inhibition of synthesis and reduction of hidrofolate (IC_{50}) in different Plasmodium Falciparum clones are analyzed. Satisfactory correlations are obtained for all antimalarial activities in all molecular sets. Molecular Quantum Similarity analysis provides a consistent, unbiased and homogeneous set of molecular descriptors and is a feasible alternative to the use of physicochemical descriptors.

#12

-
- Autors:** Carbó-Dorca, R.; Amat, L.; Besalú, E.; **Gironés, X.**; and Robert, D.
- Títol:** Quantum Molecular Similarity: Theory and applications to the evaluation of molecular properties, biological activities and toxicity. (Capítol de llibre)
- Referència / estat:** Math. And Comp. Chem.: Fundamentals of Molecular Similarity (2001) 187-320 Proceedings of 4th Girona Seminar on Molecular Similarity
- Abstract:** Here is developed the formalism, where Quantum Similarity Measures (QSM) arise as a natural product of a self contained theoretical framework based on Quantum Theory. This fact is used latter to establish a fundamental connection between Quantum Theory and QSAR. This basic relationship is found within a discrete quantum mechanical formalism: a fashion to represent molecular structures by means of n-dimensional vectors. The first paper section is devoted to describe Tagged Sets devices, a general procedure with Fuzzy Set theory appearing as a particular case. As a consequence of the theoretical development, Convex Sets result to be prominent and the notion of Vector Semispace becomes necessary for the coherence of the mathematical framework. The definition of Quantum Object (QO) is ready made using a mix of old quantum theoretical and new concepts. In the definition of QO, Density Function (DF) plays a fundamental role. A simplified mathematical picture: the Atomic Shell Approximation (ASA), is presented at the same time as the Transformation Rule, an ideal device to connect quantum wavefunctions with DF. This is followed by the way the DF construction can be generalized, by using inward matrix products. Beyond the electronic DF structure, various alternative theoretical elements are described too, among others, Kinetic Energy and other observable Density Distributions. The way to discuss the appearance of new DF is associated to the idea of extended Hilbert spaces (EHS). EHS formal structure produces an alternative quantum mechanical operator form, which can be employed to represent expectation values in a unique formulation. Finally, QSM is made evident in this context as a natural source of molecular structure discrete description. These results add new insight to n-dimensional QO representations, because: A) Provide several general unbiased sources of new QO descriptors. B) Describe in a general way the theoretical background of the relationship between molecular structure and properties (QSPR), biological activity (QSAR) and toxicity (QSTR). C) Enable the construction of the adequate mathematical tools to solve QSAR problems, as the tuned-QSAR technique. D) Allows the construction of sound and general descriptors as an alternative to Hammett's σ or log P molecular parameters. E) Produce a set of background matrices able to be used for the evaluation of Quantum Topology Indices, All the steps above summarised are completed and illustrated, when possible, providing several application examples and pictures.
- Especial:** Carbó-Dorca, R.; **Gironés, X.**; and Mezey, P.G., eds. New York: Kluwer Publishers / Plenum Press, 2001

#13

-
- Autors: Ponec, R.; Roithová, J.; **Gironés, X.**; and Jug, K.
Títol: On the nature bonding in N_5^+ ion.
Referència / estat: J. Mol. Struct. Theochem Vol.545 (2001) 255-264
Abstract: The structure of the recently synthesized N_5^+ ion was analyzed using two recently proposed methodologies, namely the generalized population analysis and the Fermi hole analysis. It has been shown that as a remnant of its synthetic precursor, the azide ion namely, the molecule contains the system of two partially perturbed three-center four-electron (3c-4e) bonds localized in the corresponding terminal NNN fragments. This structural assignment agrees with recently proposed analysis based on other methodologies.

#14

-
- Autors: **Gironés, X.**; Carbó-Dorca, R.; and Mezey, P.G.
Títol: Application of Promolecular ASA Densities to Graphical Representation of Density Functions of Macromolecular Systems.
Referència / estat: J. Mol. Graphics & Mod.19, (2001) 343-348
Abstract: In this paper we report the application of the Promolecular Atomic Shell Approximation (Promolecular ASA) to the graphical representation of the density function (DF) of large macromolecular systems. Promolecular ASA DF, constructed from previously computed and fitted atomic densities, provides a fast and practical representation of Molecular IsoDensity Contours (MIDCOs). These representations can be extended to macromolecular systems composed by >1000 atoms easily and with low computational costs, allowing the visualization of protein DF. The method is at first presented with a small molecule (2,4,6-trinitrophenol), comparing the resulting ASA MIDCOs with direct ab initio contours. For macromolecular tests the Promolecular ASA densities are also applied to the generation of macromolecular density surfaces of two proteins: myoglobin (2541 atoms) and gene V protein (1362 atoms).

#15

-
- Autors: Gallegos, A; Robert, D; **Gironés, X.**; Carbó-Dorca, R.
Títol: Structure-Toxicity Relationships of Polycyclic Aromatic Hydrocarbons using Molecular Quantum Similarity.
Referència / estat: J. Comput-Aided Mol. Desing 15 (2001) 67-80
Abstract: The establishment of quantitative structure-activity relationship (QSAR) models for the toxicity of polycyclic aromatic hydrocarbons (PAHs) is described. Two properties: in vitro percutaneous absorption in rat skin and discrete levels of carcinogenic activity are examined using molecular quantum similarity measures (MQSM). The results show that MQSM produces comparable, or even better, results than other approaches using physicochemical, topological and quantum-chemical molecular descriptors. Furthermore, a careful analysis puts into evidence that most of the information characterized by the original descriptors is in fact contained in the molecular density functions, the basis of MQSM. The present paper, together with several other ones reported by our laboratory, prove that MQSM might be appropriate theoretical tools for QSAR and computer-aided drug design, comparable to other highly predictive QSAR methodologies.

#16

-
- Autors: **Gironés, X.**; Robert, D.; and Carbó-Dorca, R.
Títol: TGSA: a molecular superposition program based on Topo-Geometrical Considerations.
Referència / estat: J.Comput. Chem. 22 (2001) 255-263
Abstract: In this paper, a new molecular alignment procedure to provide general-purpose, fast, automatic, and user-intuitive three-dimensional molecular alignments is presented. This procedure, called Topo-Geometrical Superposition Approach (TGSA), is only based on atom types and interatomic distances comparisons, hence the procedure can handle large sets of compounds within affordable computational costs. The algorithm has been implemented into the program named TGSA99, and it has been tested over eight different sets of compounds: flavilium salts, amino acids, indole derivatives, AZTs, steroids, anilide derivatives, poly-aromatic-hydrocarbons and inhibitors of thrombine. The performance is evaluated by means of computational time, number of atoms superposed and index of fit between the structures compared

#17

-
- Autors: Renners, I.; Carbó-Dorca, R.; Grauel, A.; Ludwig, L.A.; Robert, D.; and **Gironés, X.**
Títol: Toxicity Prediction by Using Genetically Optimized B-Spline Networks Based on Molecular Quantum Similarity.
Referència / estat: 2th International ICSC Symposium on Neural Computation-NC'2000
Abstract: Predicting the toxicology of chemicals is of special interest for environmental and health problems. Furthermore a system which can give a clue to toxicity is highly economical because it avoids useless and expensive pilot batches. To model the Quantitative Structure-Activity Relationships (QSARs) to toxicity we computed Molecular Quantum Similarity Measures (MQSM) for each molecule. These MQSM values then were used as input to a genetically optimized B-spline networks (BSNs), a network type which can be interpreted as a Takagi-Sugeno neuro-fuzzy system, to preserve the possibility of applying rule extraction methods.

#18

-
- Autors: Besalú, E.; **Gironés, X.**; Amat, L.; and Carbó-Dorca, R.
Títol: Molecular Quantum Similarity and the Fundaments of QSAR.
Referència / estat: Accouts of Chemical Research, *acceptat*
Abstract: A general overview on quantum similarity and applications to QSAR is presented. Concepts regarding quantum similarity from its theoretical foundation and consecutive development involving mathematical formulation, taxonomy, similarity measures and indexing will be presented. The conceptual part is complemented with several application examples concerning the well-known Crammer 31 steroids set. The practical part covers approximate quantum similarity calculations, molecular superposition and statistics. In this way, the reader will find both, basic general information and applicability of the developed concepts.

#19

Autors: **Gironés, X.**; Amat, L.; and Carbó-Dorca, R.

Títol: Modelling Large Macromolecular Structures using Promolecular Densities.

Referència / estat: J. Chem. Inf. Comput. Sci., *acceptat*.

Abstract: A procedure to easily construct fitted density functions is presented. This methodology, based on the promolecule approach, is able to handle large macromolecular systems, like proteins. The usual procedure dealing with fitted densities has been improved by adding some restrictions, which allow faster calculations. As a main application example, molecular isodensity contours (MIDCOs) are constructed for two proteins, one of them composed by more than 50000 atoms. MIDCOs, as a visual representation of the molecular density function, and thus, an important descriptor of the molecular charge distribution, constitute a powerful tool in the understanding of molecular systems. MIDCOs are presented for two proteins, allowing to explore their surfaces, as well as to analyse their shape. Also, as quantum mechanical calculation example, Molecular Quantum Self-Similarity Measures are calculated for several proteins.

#20

Autors: Ponec, R.; **Gironés, X.**; and Carbó-Dorca, R.

Títol: Molecular basis of LFER. On the nature of inductive effects in aliphatic series.

Referència / estat: J. Chem. Inf. Comput. Sci., *acceptat*.

Abstract: The nature of polar substituent effect in aliphatic series was analyzed in terms of quantum chemical model based on Longuet-Higgins theory of proton-affinity. It has been shown that the effect is generally composed of two additive components, the electrostatic one and the polarization one. The relative weights of these components depends in any given case on the type of reaction series. Thus, for example, the substituent induced variation of gas phase acidity is due to electrostatic component classically described by Kirwood-Westheimer theory. On the other hand, the gas phase basicity in a series of aliphatic amines is primarily determined by the polarization term that in traditional description correspond to the concept of +I effect.

#21

Autors: **Gironés, X.**; and Carbó-Dorca, R.

Títol: Sobre les relacions lineals d'energia lliure: Mesures de Semblança Molecular Quàntica sobre Funcions de Densitat Electrònica Modificades.

Referència / estat: Scientia Gerundensis, 25 (2001) 5-15.

Abstract: En aquest treball es presenta l'ús de funcions densitat electrònica modificades per a incrementar el paper que pren una regió molecular concreta, considerada com a responsable de la reactivitat molecular, tot i mantenint la mida de la funció densitat original. Aquestes densitats s'utilitzen per a fer mesures d'autosemblança molecular quàntica, presentant-se com una alternativa a l'ús de fragments moleculars aïllats en estudis de relació estructura propietat. El treball es complementa amb un exemple pràctic, on es correlaciona l'autosemblança molecular a partir de densitats modificades amb l'energia d'una reacció isodèsmica.

#22

-
- Autors: Besalú, E.; **Gironés, X.**; and Carbó-Dorca, R.
Títol: Mathematical Foundations of Quantum Similarity.
Referència / estat: Math. And Comp. Chem. Kluwer Academic/Plenum Publishers (Llibre Monogràfic), *en fase final de redacció*.

#23

-
- Autors: **Gironés, X.**; Gallegos, A.; and Carbó-Dorca, R.
Títol: Antimalarial Activity of Synthetic 1,2,4-Trioxanes and Cyclic Peroxy Ketals, a Quantum Similarity Study.
Referència / estat: J.Comp.-Aided Molec. Design, *acceptat*.
Abstract: In this work, the antimalarial activity of two series of synthetic 1,2,4-trioxanes and a set of cyclic peroxy ketals are tested for correlation search by means of Molecular Quantum Similarity Measures (MQSM). QSAR models, dealing with different biological responses (IC_{50} , IC_{90} and ED_{90}) of the parasite Plasmodium Falciparum, are constructed using MQSM as molecular descriptors and are satisfactorily correlated.

#24

-
- Autors: Renners, I.; Ludwig, L.A.; Grauel, A.; Benfenati, E.; Pelagatti, S.; Robert, D.; Carbó-Dorca, R.; **Gironés, X.**
Títol: Modeling toxicity with molecular descriptors and similarity measures via B-Spline networks.
Referència / estat: 8th Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU'2000)
Abstract: Predicting the toxicology of chemicals is of special interest for environmental and health problems. Furthermore a system which can give a clue to toxicity is highly economical because it avoids useless and expensive pilot batches. To model the Quantitative Structure-Activity Relationships (QSARs) regarding toxicity we computed molecular descriptors and Molecular Quantum Similarity Measures (MQSMs) for each molecule. These values then were independently used as input to small B-spline networks (BSNs), a network type which can be interpreted as a Takagi-Sugeno neuro-fuzzy system, to preserve the possibility of applying rule extraction methods.

#25

-
- Autors: **Gironés, X.**; and Carbó-Dorca, R.
Títol: Using Molecular Quantum Similarity Measures under Stochastic Transformation to describe Physical Properties of Molecular Systems.
Referència / estat: J. Chem Inf. Comp. Sci., *acceptat*.
Abstract: The application of Molecular Quantum Similarity Measures (MQSM) to correlate physicochemical properties is reported. Satisfactory Quantitative Structure-Property Relationship (QSPR) models are obtained for three molecular sets, where boiling points and chromatographic retention times and indices are studied. In this work, MQSM are scaled using a stochastic transformation and related to molecular properties using the Partial Least Squares technique.

#26

-
- Autors: Ponec, R.; Roithová, J.; **Gironés, X.**; Lain, L.; Torre, A.; Bochicchio, R.
Títol: On the nature of C-Li bonding in lithiated hydrocarbons and lithiocarbons.
Referència / estat: J. Phys. Chem. A, 160 (2002) 1019-1025.
Abstract: The nature of C-Li bonding in CH₃Li, (CH₃Li)₄ and CLi₆ was analysed using two recently proposed new methodologies namely the AIM generalized population analysis and Fermi hole analysis. The calculations were performed at HF level of the theory using 6-311G** basis sets. The main issue was the realistic estimation of the polarity of C-Li bonds. The analysis confirms that C-Li bonds in CH₃Li are predominantly ionic but some slight modifications of the contemporary picture of bonding in CLi₆ is proposed.

#27

-
- Autors: **Gironés, X.**; and Carbó-Dorca, R.
Títol: Molecular Quantum Similarity-based QSARs for Binding Affinities of Steroids.
Referència / estat: J. Chem. Inf. Comput. Sci., *enviat*.
Abstract: The application of Molecular Quantum Similarity Measures (MQSM) to correlate biological activities for three different sets of steroids is reported. A general protocol for the generation of descriptors is detailed, thus covering molecular superposition, electronic density fitting and quantum similarity calculation issues. Satisfactory Quantitative Structure-Activity Relationship (QSAR) models ($r^2 \in [0.69, 0.94]$ and $q^2 \in [0.59, 0.73]$), comparable to previous studies, are obtained in all cases, where steroid binding affinities to different enzymes are studied. In this work, MQSM, properly scaled using Carbó Index, are related to activity using a Partial Least Squares routine.

#28

-
- Autors: Ponec, R.; and **Gironés, X.**
Títol: Chemical structures from the analysis from of domain averaged Fermi Holes.
Títol: Hypervalence and the nature of bonding in isocoordinated molecules SF₆ and CLi₆.
Referència / estat: *En escriptura*.

#29

-
- Autors: Ponec, R.; and **Gironés, X.**
Títol: Molecular Quantum Similarity Measures with Fermi Hole Densities: Modelling Hammett Sigma.
Referència / estat: Proceedings of the 5th Girona Seminar on Molecular Similarity.

#30

-
- Autors: **Gironés, X.**; and Carbó-Dorca, R.
Títol: Modelling Toxicity using Molecular Quantum Similarity Measures.
Referència / estat: SAR QSAR Environ. Res, *en revisió*. Dins del monogràfic dedicat a la tercera sessió del 5th Girona Seminar on Molecular Similarity
Abstract: In this work, a general procedure for the application of Molecular Quantum Similarity Measures (MQSM) into the Quantitative Structure-Toxicity Relationships (QSTR) field is presented. The methodology is presented from its theoretical background, beginning with the formulation of MQSM and the role of density functions. Other topics, like molecular modelling, electronic density fitting, similarity scaling and molecular superposition are also analysed. The theoretical part is complemented with three different application examples dealing with acute aquatic toxicity related to different fish species. The molecular sets are constituted by different kinds of benzene derivatives. Acceptable correlations are obtained for the aquatic toxicity using quantum similarity derived descriptors, allowing a fairly confident classification of the studied compounds according to their toxicity level.

#31

-
- Autors: Gallegos, A.; **Gironés, X.**; and Carbó-Dorca, R.
Títol: Topological Quantum Similarity Measures: Applications in QSAR.
Referència / estat: SAR QSAR Environ. Res, *en revisió*. Dins del monogràfic dedicat a la tercera sessió del 5th Girona Seminar on Molecular Similarity
Abstract: A novel method for computing new descriptors to construct Quantitative Structure-Toxicity Relationships is presented. First, a brief review on the classical graph theory is presented and, then, the link with molecular similarity is drawn. In the applications section, molecular topological indices are calculated using the interatomic Molecular Quantum Similarity Measure with a Coulomb weight operator. The use of similarity matrices instead of classical topological ones has been adopted according to the connection between molecular topology and the general theory of Quantum Similarity. Afterwards, the molecular descriptors, which include the structural information necessary to properly describe the system, are employed to derive numerical correlation with toxicities. The QSAR model is built using a multilinear regression technique. Finally, some application examples are presented, including polycyclic aromatic hydrocarbons and aquatic toxicants, demonstrating the applicability of the exposed methodology.

#32

-
- Autors: Espinosa, G.; Arenas, A.; Giralte, F.; Amat, L.; Gironés, X.; and Carbó-Dorca, R.
- Títol: QSAR for TD₅₀ of Aromatic Compound by using an Integrated SOM-Fuzzy Artmap based Neural System with Topological and Quantum Molecular Similarity Descriptors.
- Referència / estat: SAR QSAR Environ. Res, *en revisió*. Dins del monogràfic dedicat a la tercera sessió del 5th Girona Seminar on Molecular Similarity
- Abstract: An integrated methodology using self-organized maps (SOMs) and fuzzy ARTMAP neural networks has been applied to predict the toxicity index TD₅₀ of 104 aromatic compounds with nitrogen-containing substituents. SOMs have been applied to select from a given pool of descriptors the most relevant subset needed to build reliable QSAR models based on fuzzy ARTMAP. The pool of descriptors used in the present work incorporated both topological and quantum information. The quantum descriptors included measures of quantum similarity, i.e. of the resemblance between two molecules and of the properties of the molecules themselves (self-similarity). The topological maps for each molecular descriptor and for the target activity variable were created and then classified into clusters on the basis of visual inspection, correlation and curvilinear component analyses. The best subset of descriptors was obtained by choosing a representative from each cluster, in particular the index that presented the highest correlation with the target variable, and additional indices afterwards in order of decreasing correlation. The selection process ended when a dissimilarity measure between the maps for the different sets of descriptors reached a minimum value, indicating that the inclusion of more descriptors did not add supplementary information. The optimal subset of descriptors was finally used as input to a fuzzy ARTMAP architecture modified to effect predictive capabilities. The proposed QSAR model predicted toxicity TD₅₀ values (carcinogenicity) significantly better than backpropagation neural network models previously reported in the literature for the same dataset.

#33

-
- Autors: **Gironés, X.**; Carbó-Dorca, R.; and Ponec, R.
- Títol: Modelling Electronic Effects: Quantum Similarity using Fermi Hole Densities
- Referència / estat: J. Chem. Inf. Comput. Sci., *enviat*.
- Abstract: A new procedure to calculate Molecular Quantum Self-Similarity Measures (MQS-SM) based on Fermi Hole Density Functions is presented. The modification introduced consists of defining a Fermi Hole around a molecular region, thus altering the regular electronic distribution. This simple operation allows concentrating most of the attention of electron density to those molecular bays assumed to be responsible for a certain molecular response. As Fermi Hole densities behave as any regular density function, they can be included into the quantum similarity framework, thus obtaining MQS-SM restricted to a molecular substructure, but still taking into account the rest of the molecular backbone. The methodology is complemented with an application example, where an estimation of the well-known Hammett sigma (σ) is calculated for a set of benzoic acid derivatives, using MQS-SM as a single descriptor.