



Universitat de Girona

# ESTRUCTURA COMPUTACIONAL I APLICACIONS DE LA SEMBLANÇA MOLECULAR QUÀNTICA

**Lluís AMAT BARNÉS**

**ISBN: 84-688-6952-X**

**Dipòsit legal: GI-469-2004**

<http://hdl.handle.net/10803/8027>

**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.

**UNIVERSITAT DE GIRONA**

INSTITUT DE QUÍMICA COMPUTACIONAL

**ESTRUCTURA COMPUTACIONAL I  
APLICACIONS DE LA SEMBLANÇA  
MOLECULAR QUÀNTICA**

Memòria presentada per:

**Lluís Amat Barnés**

per a la defensa i obtenció del grau de doctor

Gener 2003



Ramon Carbó-Dorca Carré, Catedràtic d'Universitat del  
Departament de Química de la Universitat de Girona,

CERTIFICO

Que en Lluís Amat Barnés, llicenciat en Ciències Químiques, ha realitzat sota la meva direcció, a l'Institut de Química Computacional i al Departament de Química d'aquesta Universitat, el treball titulat **Estructura computacional i aplicacions de la semblança molecular quàntica** 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 a efectes legals signo aquest certificat.

Prof. Ramon Carbó-Dorca

Institut de Química Computacional

Departament de Química

Universitat de Girona

Girona, 8 de gener de 2003



*Als meus pares i germans*



# Índex

<b>1</b>	<b>Introducció</b>	<b>1</b>
1.1	Continguts de la tesi	3
1.2	Agraïments	9
	<b>Referències</b>	<b>10</b>
<b>2</b>	<b>Semblança molecular quàntica</b>	<b>11</b>
2.1	Introducció	12
2.2	Descripció dels sistemes quàntics	14
2.3	Funció de densitat electrònica de primer ordre	15
2.4	Densitat electrònica d'un fragment molecular	16
2.5	Mesures de semblança quàntica	17
2.5.1	<i>MQSM de solapament o recobriment</i>	18
2.5.2	<i>MQSM de Coulomb</i>	19
2.6	Mesures d'autosemblança quàntica	19
2.7	Mesures de semblança quàntica de moment	20
2.8	Índexs de semblança quàntica	21
2.9	Superposició molecular	22
	<b>Referències</b>	<b>23</b>
<b>3</b>	<b>Funcions densitat ASA</b>	<b>31</b>
3.1	Mètode de Hartree-Fock	32
3.1.1	<i>Aproximació LCAO i equacions de Hartree-Fock-Roothaan</i>	34
3.2	Funció densitat <i>ab initio</i>	36
3.3	Funció densitat aproximada	37
3.4	Mètode d'ajust de mínims quadrats	38
3.5	Funcions densitat ASA	39
3.5.1	<i>Mètode de mínims quadrats adaptat a les funcions ASA</i>	40
3.6	Funcions densitat <i>PASA</i>	41
3.7	Funció error quadràtic integral	43
3.8	Optimització dels coeficients ASA mitjançant rotacions de Jacobi	44
3.8.1	<i>Esquema computacional de l'algorisme EJR</i>	46
3.9	Ajustos atòmics	50
3.9.1	<i>Optimització dels exponents ASA</i>	50
3.9.2	<i>Optimització conjunta de coeficients i exponents ASA</i>	53
3.9.3	<i>Programa GATOMIC</i>	54
3.9.4	<i>Ajust ASA atòmic d'una base de funcions 3-21G</i>	56
	<i>Ll. Amat, R. Carbó-Dorca, J. Comput. Chem. 1997, 18, 2023-2039</i>	59



3.10	Milliores en l'algorisme d'optimització dels coeficients ASA	77
3.10.1	<i>Esquema computacional del càlcul del sinus EJR</i>	78
3.10.2	<i>Ajust atòmic d'una base de funcions d'Huzinaga</i>	79
	<i>Ll. Amat, R. Carbó-Dorca, J. Comput. Chem. 1999, 20, 911-920</i>	81
3.10.3	<i>Altres ajustos atòmics</i>	91
3.11	Ajustos moleculars	91
3.11.1	<i>Esquema computacional</i>	91
3.11.2	<i>Exemples d'ajustos moleculars</i>	92
	<i>Ll. Amat, R. Carbó-Dorca, J. Chem. Inf. Comput. Sci. 2000, 40, 1188-1198</i>	95
3.12	Classificació de camins de reacció mitjançant mesures de semblança	107
3.13	Ús de les funcions PASA per reduir el nombre de cicles SCF	112
	<i>Ll. Amat, R. Carbó-Dorca, Int. J. Quantum Chem. 2002, 87, 59-67</i>	115
	<b>Discussió</b>	<b>125</b>
	<b>Referències</b>	<b>127</b>
<b>4</b>	<b>Superposició molecular</b>	<b>131</b>
4.1	Descripció de la posició relativa de les molècules	132
4.2	Superposició molecular referent al màxim de la MQSM	134
4.2.1	<i>Deducció de l'algorisme de superposició del màxim de semblança</i>	134
4.2.2	<i>Maximització de les MQSM</i>	138
4.3	Derivades analítiques de les MQSM definides sobre funcions ASA	142
4.3.1	<i>Primera derivada de les MQSM</i>	143
4.3.2	<i>Segona derivada de les MQSM</i>	145
4.4	Esquema general del programa MOLSIMIL	148
4.5	Algorisme de superposició topo-geomètric	149
4.5.1	<i>Exemple de sobreposició on hi intervenen àtoms pesants</i>	151
	<b>Discussió</b>	<b>155</b>
	<b>Referències</b>	<b>157</b>
<b>5</b>	<b>Anàlisi QSAR</b>	<b>159</b>
5.1	Models de regressió multilinear	161
5.1.1	<i>Regressió lineal</i>	162
5.1.2	<i>Variança, desviació estàndard i covariança</i>	163
5.1.3	<i>Coefficient de correlació i de determinació</i>	163
5.1.4	<i>Coefficient de múltiple determinació</i>	165
5.1.5	<i>Importància estadística dels models MLR</i>	166
5.2	Capacitat de predicció dels models MLR	167
5.2.1	<i>Coefficient de la validació creuada</i>	167
5.2.2	<i>Càlcul del vector de prediccions a partir de la matriu de predicció</i>	168
5.3	Mètodes d'anàlisi multivariant	169
5.3.1	<i>Anàlisi de components principals</i>	170

5.3.2	<i>Tècnica de mínims quadrats parcials</i>	171
5.3.3	<i>Anàlisi de coordenades principals</i>	173
5.4	Selecció de les variables per obtenir el millor model <i>MLR</i>	175
5.5	Validació estadística dels models <i>QSAR</i>	176
5.5.1	<i>Validació creuada</i>	177
5.5.2	<i>Test d'aleatorietat sobre el vector de les activitats</i>	177
5.5.3	<i>Dividir les dades en una sèrie d'exploració més un conjunt de test</i>	178
	<b>Referències</b>	<b>179</b>
<b>6</b>	<b>Matrius de semblança en anàlisi <i>QSAR</i></b>	<b>183</b>
6.1	Evolució de les anàlisis <i>QSAR</i> basades en <i>MQSM</i>	184
6.1.1	<i>Descripció de les molècules: funció de densitat electrònica</i>	185
6.1.2	<i>Superposició molecular</i>	186
6.1.3	<i>Tractament estadístic de les matrius de semblança</i>	187
6.2	Influència de determinats factors en les anàlisis <i>QSAR</i>	188
6.2.1	<i>Influència de la densitat electrònica</i>	190
6.2.2	<i>Influència de la superposició molecular</i>	194
6.2.3	<i>Influència de la geometria molecular</i>	198
	<b>Discussió</b>	<b>201</b>
	<b>Referències</b>	<b>203</b>
<b>7</b>	<b>Aproximació <i>QS-SM</i> de fragments</b>	<b>207</b>
7.1	<i>LFER</i> i models de <i>QSAR</i> clàssica	207
7.2	Efectes hidrofòbics	210
7.2.1	<i>QS-SM com a alternativa a log P</i>	212
7.3	Efectes electrònics dels substituents	214
7.3.1	<i>QS-SM com a substitut de la constant <math>\sigma</math> de Hammett</i>	215
	<i>R. Ponec, Ll. Amat, R. Carbó-Dorca, J. Comput.-Aided Mol. Des. 1999, 13, 259-270</i>	219
7.3.2	<i>Correlacions entre els valors esperats de l'espai de moment i la constant <math>\sigma</math></i>	231
7.4	Aplicacions <i>QSAR</i>	232
7.4.1	<i>Descripció de l'aproximació <i>QSAR</i> emprant <i>QS-SM</i> de fragments</i>	232
7.4.2	<i>Exemples <i>QSAR</i></i>	235
	<i>Ll. Amat, R. Carbó-Dorca, R. Ponec, J. Med.Chem. 1999, 42, 5169-5180</i>	239
7.5	Identificació dels fragments moleculars responsables de l'activitat biològica	251
7.5.1	<i>Generalització del mètode <i>QS-SM</i> de fragments</i>	251
7.5.2	<i>Correlació de la constant <math>\sigma</math> amb altres fragments moleculars</i>	253
7.5.3	<i>Estudi de l'activitat dels esteroides de Cramer</i>	254
	<i>Ll. Amat, E. Besalú, R. Carbó-Dorca, R. Ponec J.Chem.Inf.Comput.Sci. 2001, 41, 978-991</i>	257
7.5.4	<i>Influència de factors estructurals en l'aproximació <i>QS-SM</i> de fragments</i>	271
	<b>Discussió</b>	<b>275</b>
	<b>Referències</b>	<b>277</b>
	<b>Conclusions</b>	<b>281</b>



# 1. Introducció

---

La química computacional és una disciplina relativament moderna que ha experimentat un important progrés en els darrers anys. Un dels sectors on la química computacional destaca per la seva implicació i ressonància és la indústria farmacèutica, on l'aplicació de tècniques de disseny molecular assistides per ordinador (*Computer-Aided Molecular Design, CAMD*) ha estat clau perquè es redueixi el temps de síntesi de nous fàrmacs. Les sigles *CAMD* engloben un nombre considerable de procediments basats en l'ús d'ordinadors i encaminats a relacionar activitat amb estructura molecular. En l'actualitat, qualsevol empresa capdavantera en el disseny de fàrmacs ha introduït en la seva línia de producció laboratoris especialitzats en les noves tecnologies de modelatge molecular i simulació informàtica. Algunes de les tasques que se'ls encomana són, per exemple, estudis orientats a dilucidar els requeriments bàsics d'una determinada activitat (*farmacòfor*), fer simulacions de l'acoblament entre el fàrmac i l'enzim, proposar mecanismes per entendre els processos biològics, o predir l'activitat d'anàlegs no sintetitzats mitjançant les tècniques conegudes amb les sigles angleses *QSAR (Quantitative Structure-Activity Relationships)* o *QSPR (Quantitative Structure-Property Relationships)*.

El procés de síntesi de nous fàrmacs inclou des del descobriment d'un principi actiu fins a la comercialització del fàrmac corresponent, i suposa a les companyies farmacèutiques una inversió aproximada de 360 milions de dòlars i un període de desenvolupament que varia entre 10 i 14 anys. La part probablement més costosa de tot el procés és el descobriment d'un principi actiu o prototip amb un interès potencial per guarir o pal·liar els efectes d'una malaltia. Una vegada identificada la molècula potencialment activa, hi ha un procés de recerca de compostos anàlegs que millorin l'activitat biològica i les característiques farmacocinètiques, al mateix temps que facin disminuir els efectes secundaris i la toxicitat. Les noves tecnologies de química combinatoria i simulació informàtica afavoreixen aquesta fase on es construeixen enormes llibreries de molècules que són analitzades per complexos sistemes robotitzats

i informatitzats. Aquest mètode de descobriment de nous agents amb activitat biològica és interessant des del punt de vista que pot convertir noves classes estructurals de compostos en fàrmacs potencials, però està fonamentat en tècniques d'assaig i error que consumeixen molt de temps i diners. Els primers intents dirigits a incrementar la probabilitat de sintetitzar un anàleg més actiu o de descobrir un nou cap de sèrie es fonamentaren en trobar correlacions entre l'estructura química de la sèrie de compostos i la seva activitat. D'aquí va sorgir les famoses sigles *QSAR* que actualment és una paraula d'ús corrent tant en el procés de disseny de nous fàrmacs com en la racionalització de les propietats farmacològiques d'una sèrie química. Una vegada trobada la fórmula definitiva del candidat a fàrmac, es passa a les diferents etapes de proves biològiques: primer en cultiu, després en animals i finalment les fases clíniques en humans, que poden allargar-se fins als set anys. En moltes ocasions unes característiques farmacocinètiques inadequades, l'aparició d'efectes secundaris inacceptables, o la biotransformació a un metabòlit tòxic, poden fer que el compost, en principi prometedor, torni als laboratoris d'investigació com un intent fallit i s'hagi de començar el procés de nou.

Els mètodes *QSAR* agrupen totes les tècniques que intenten establir models empírics o teòrics del comportament de famílies de compostos biològicament actius, amb l'objectiu d'assolir de la manera més eficient possible òptims d'activitat mitjançant les dades de l'afinitat d'un nombre limitat de productes. Les tècniques *QSAR* s'apliquen una vegada s'ha determinat experimentalment l'activitat d'una sèrie de compostos anàlegs a un principi actiu. Normalment els membres de la sèrie d'exploració estan formats per un nucli comú i uns substituents o fragments variables, que són característics de cada producte de la sèrie. Llavors es defineixen uns descriptors moleculars sobre la sèrie analitzada que actuen de variables independents en un model matemàtic que els relaciona amb el vector d'activitats biològiques que és la variable dependent. L'anàlisi *QSAR* engloba tant la definició dels descriptors moleculars com les tècniques estadístiques requerides en la construcció dels models matemàtics. Freqüentment el model resultant s'utilitza en una fase de comprovació que consisteix en calcular l'activitat de membres de la mateixa família que no han format part de la sèrie d'exploració i contrastar-la amb els valors determinats experimentalment. El següent estadi és l'ús del model per predir l'activitat de productes no sintetitzats, amb la finalitat de distingir els anàlegs actius dels no actius, i així millorar l'efectivitat de tot el procés.

En els darrers anys han sorgit moltes aproximacions *QSAR* que consideren les propietats de les molècules en les tres dimensions. En la majoria de casos es tracta de metodologies fonamentades en mesures de semblança molecular,<sup>1-11</sup> les quals estan perfectament establertes dins la disciplina *CAMD* i són d'utilitat tant per generar paràmetres vàlids en *QSAR/QSPR* com per optimitzar la superposició entre estructures moleculars. La definició d'una mesura de semblança entre molècules és arbitrària i comporta l'aparició de diferents aproximacions de la semblança molecular. Una d'elles, que és l'objecte d'aquesta tesi, es fonamenta en la mecànica quàntica i utilitza les densitats de càrrega electrònica com a font de comparació. D'acord amb els postulats de la mecànica quàntica, un sistema quàntic pot ser completament caracteritzat a través de la funció d'ona que és solució de l'equació de Schrödinger, i per extensió, de la corresponent funció densitat (*Density Function, DF*). A partir d'aquesta premissa, s'han definit les mesures de semblança molecular quàntica (*Molecular Quantum Similarity Measures, MQSM*) com la integral de volum entre les respectives *DF* dels sistemes considerats, ponderada per un operador no diferencial i definit positiu. Una *MQSM* constitueix una forma molt simple d'obtenir relacions entre els objectes quàntics comparats mitjançant la identificació de les característiques de la densitat electrònica que varien d'un sistema a un altre. Les tècniques basades en la semblança quàntica tenen moltes aplicacions, majoritàriament, encara que no exclusivament, en la racionalització i predicció de l'activitat de fàrmacs.

## 1.1 Continguts de la tesi

Aquesta tesi doctoral, titulada *Estructura computacional i aplicacions de la semblança molecular quàntica*, constitueix un resum del treball que he realitzat a l'Institut de Química Computacional (IQC) de la Universitat de Girona, dins el grup d'enginyeria molecular quàntica dirigit pel professor Ramon Carbó-Dorca. L'objectiu del treball és presentar l'estructura computacional que se segueix en el nostre laboratori quan s'apliquen les *MQSM* en anàlisis *QSAR/QSPR*, i il·lustrar-ho amb algunes de les aplicacions més rellevants. El meu treball a l'IQC s'ha centrat fonamentalment en el desenvolupament de nous mètodes relacionats amb la semblança molecular quàntica i en la codificació dels programes informàtics que se'n deriven.

El principal inconvenient dels estudis pràctics de les *MQSM* orientats a determinar relacions estructura-activitat és el nombre i tipus de compostos que intervenen en els càlculs. Normalment s'estudien conjunts moleculars extensos, compostats per molècules grans que són difícils d'analitzar a nivell *ab initio*. A més les *MQSM* porten implícites en la seva definició la idea d'optimitzar la posició relativa de les molècules estudiades. La recerca de la superposició molecular òptima normalment esdevé la part més costosa de tot el procés perquè requereix computar repetidament la mesura de semblança. Tot plegat ha fet indispensable l'ús de *DF* aproximades en els estudis de *MQSM*.

El projecte inicial de recerca va ser el disseny d'un programa de càlcul de mesures de semblança quàntica. Es van plantejar dos objectius: d'una banda descriure el més acuradament possible les densitats electròniques de les molècules, i en segon terme desenvolupar un mètode d'optimització de la superposició molecular. Es va començar a treballar sobre el programa MOLSIMIL, existent a l'IQC, on es descriuen les molècules mitjançant una aproximació de tipus *CNDO*, i s'utilitzaven moments dipolars i quadropolars per orientar les molècules en l'espai. Aquesta etapa inicial, dedicada primordialment a l'aprenentatge dels aspectes bàsics de la teoria de la semblança quàntica i a la iniciació en les tècniques de programació, va concloure amb l'elaboració del meu treball de recerca, titulat *Estructura computacional de les mesures de semblança quàntica: programa MOLSIMIL96*. La recerca es va centrar bàsicament en la descripció de la densitat electrònica de les molècules mitjançant la que s'ha anomenat aproximació de capes atòmiques (*Atomic Shell Approximation, ASA*), la qual construeix la *DF* com una combinació lineal de funcions Gaussianes 1s centrades en els àtoms, amb la característica que els coeficients de l'expansió han de ser definits positius. En un primer estadi es va utilitzar una aproximació promolecular, basada en la definició de la densitat electrònica d'una molècula com a simple sumes de contribucions atòmiques, i de caire empíric, on els exponents de les capes atòmiques s'ajustaven de manera que la diferència entre el valor de l'autosemblança atòmica *ab initio* i la calculada amb la *DF* aproximada fos mínima. També, en aquest mateix període, vaig col·laborar en el disseny d'un algorisme de superposició molecular per cercar el màxim de la mesura de semblança, i que ha estat implementat en el programa MOSLIMIL.

Posteriorment s'ha aprofundit en algun dels projectes d'investigació iniciats en el treball de recerca, fins a concloure en els treballs exposats en la present tesi doctoral.

El capítol 2, *Semblança molecular quàntica*, és introductori. Es revisen els principals conceptes i definicions relacionats amb la teoria de les *MQSM*. Sense aprofundir massa en detalls computacionals, es dóna una idea de les metodologies emprades en el càlcul de les *MQSM*. Serà en els següents capítols on s'especificaran els algorismes matemàtics dissenyats per dur-les a la pràctica.

En el capítol 3, *Funcions densitat ASA*, es descriu el mètode d'ajust de les funcions ASA. Si bé en un inici es van utilitzar uns exponents deduïts empíricament, posteriorment s'han desenvolupat algorismes d'ajust de la densitat *ab initio* a una combinació lineal de funcions esfèriques centrades en els àtoms. Els coeficients de l'expansió lineal s'obtenen de la minimització de la funció error quadràtic integral, definida com la integral entre la diferència de les *DF ab initio* i aproximada al quadrat. Com ja s'ha comentat, la principal característica de les funcions ASA és la restricció addicional dels coeficients de l'expansió lineal ha ser definits positius. Així s'aconsegueix que la *DF* aproximada tingui les propietats d'una distribució de probabilitats. El principal avenç en aquest camp ha estat el disseny d'un algorisme d'ajust de les funcions ASA basat en la tècnica de rotacions de Jacobi. Inicialment es va idear com un ajust de *DF* atòmiques, però posteriorment també s'ha comprovat que funciona en sistemes moleculars. En el capítol 3 s'ha inclòs el codi esquematitzat d'algunes subrutines emprades en el programa d'ajust i es presenten alguns exemples de càlcul de bases atòmiques. Precisament, la disponibilitat de conjunts de bases atòmiques parametritzades ha permès generar la densitat electrònica de moltes sèries de compostos mitjançant una aproximació promolecular, que s'ha identificat amb les sigles *PASA*. En el capítol s'exposen diversos exemples d'aplicació de les funcions *PASA* en estudis de *MQSM*. A més, en el darrer apartat es mostra un possible ús de les funcions *PASA* en càlculs d'energia electrònica. En concret s'ha desenvolupat una tècnica que genera matrius densitat inicials útils en el càlcul iteratiu del camp autocoherent, que redueixen el nombre de cicles necessaris per assolir els criteris de convergència.



En el capítol 4, *Superposició molecular*, es descriu el programa MOLSIMIL de càlcul de mesures de semblança quàntica. S'inclou una descripció detallada d'un algorisme específic de les *MQSM* basades en el càlcul de la semblança entre funcions de densitat electrònica que permet determinar la superposició molecular òptima, adoptant el criteri de prendre el valor del màxim absolut. La representació de les superfícies de mesures de semblança ha permès comprovar que quan dos àtoms diferents d'hidrogen se sobreposen donen un màxim. Això és així perquè els màxims de semblança se situen en els punts on es concentra la major part de la densitat electrònica, com són les posicions dels nuclis atòmics. La metodologia proposada es fonamenta en la recerca de l'alineament molecular en el qual se sobreposen el màxim nombre d'àtoms de les dues molècules comparades. L'algorisme de maximització de les *MQSM* s'ha deduït a partir d'un cas extrem, com és considerar una densitat construïda mitjançant funcions delta de Dirac. Dins el mateix algorisme s'han desenvolupat diferents nivells de càlcul, en els quals les superposicions moleculars poc favorables són descartades mitjançant criteris de semblança atòmica. Després de la recerca global, es fa un refinament del millor alineament molecular trobat mitjançant un mètode de Newton per assolir el màxim absolut de semblança. En el capítol també s'ha inclòs la descripció d'un nou mètode de sobreposició molecular desenvolupat en el nostre laboratori que cerca la màxima semblança estructural entre les molècules comparades a partir únicament de criteris topològics i geomètrics. Un dels motius que ha portat a la programació d'aquest nou algorisme ha estat la influència que tenen els àtoms pesants en les superposicions resultants del criteri de màxima semblança.

Una altra línia d'investigació promoguda en el nostre laboratori i contigua al càlcul de les *MQSM*, és el tractament estadístic de les matrius de dades que contenen les mesures o índexs de semblança definits per a tots els possibles parells de molècules de la sèrie analitzada. En el capítol 5, *Anàlisis QSAR*, s'introdueixen les tècniques estadístiques més comunes que s'utilitzen en els estudis *QSAR*. No són metodologies exclusives de les *MQSM*, sinó que es poden aplicar a qualsevol conjunt de descriptors moleculars. En qualsevol circumstància, l'objectiu és construir models matemàtics que relacionin els descriptors moleculars amb les dades observables, com són les activitats biològiques. Alguns dels procediments estadístics descrits en el capítol fan referència a transformacions del conjunt de dades per reduir-ne la dimensió i a les tècniques més comunes emprades en la validació estadística dels models matemàtics resultants.

En els darrers capítols es mostren alguns exemples d'anàlisis *QSAR* emprant dues tecnologies diferents, fonamentades ambdues en el desenvolupament de la semblança molecular quàntica i que s'originen a partir de representacions diferents de les molècules. En la primera les molècules es descriuen mitjançant la funció densitat global i es defineixen les mesures de semblança entre tots els possibles parells de molècules que formen la sèrie analitzada. Són importants, entre d'altres, els processos de superposició molecular i l'anàlisi conformacional. La segona aproximació es fonamenta en mesures de semblança de fragments moleculars definides sobre una mateixa molècula. El principal aspecte a tenir en compte és la densitat electrònica utilitzada per definir els fragments moleculars.

En el capítol 6, *Matrius de semblança en anàlisis QSAR*, es descriu la primera aproximació basada en *MQSM* que s'ha utilitzat per relacionar els canvis en l'estructura d'una família de molècules amb les seves activitats. Els descriptors moleculars emprats en la generació dels models matemàtics són les matrius de mesures de semblança obtingudes després del procés d'alineament molecular. En el capítol es presenta una síntesi cronològica de l'evolució soferta en els diferents processos involucrats en el càlcul de les *MQSM*, així com un resum dels principals resultats obtinguts emprant les matrius de semblança com a descriptors moleculars en anàlisis *QSAR*. També s'ha volgut examinar la repercussió que tenen diferents factors en els resultats estadístics dels models *QSAR*. Així s'ha estudiat quin efecte té la utilització de *PASA DF* en lloc de densitats *ab initio*, la influència del mètode de superposició molecular escollit per alinear les molècules en l'espai, i les conseqüències de variacions en l'estructura molecular en les correlacions finals estructura-activitat.

En el capítol 7, *Aproximació QS-SM de fragments*, es mostra com determinades mesures de semblança quàntica poden substituir els principals paràmetres empírics utilitzats en les equacions de *QSAR* clàssiques. Es tracta d'una nova línia de recerca dins l'àmbit de la semblança molecular quàntica que s'ha iniciat a partir de l'estada del professor Robert Ponc del *Institute of Chemical Process Fundamentals* de l'acadèmia txeca de les ciències a l'IQC l'any 1997. La base dels nous descriptors moleculars ha estat mesures d'autosemblança quàntica (*Quantum Self-Similarity Measures, QS-SM*), calculades sobre la *DF* de tota la molècula o bé localitzades en el fragment responsable de l'activitat biològica. El principal avantatge de les *QS-SM* respecte a les *MQSM* és la

supressió del procés de superposició molecular. En els primers treballs s'analitzen les *QS-SM* que poden servir d'alternativa a paràmetres tan diversos com el coeficient de partició octanol-aigua ( $\log P$ ) o la constant  $\sigma$  de Hammett. El següent avenç ha estat la comprovació que els mateixos descriptors teòrics poden ser usats en qualitat de descriptors moleculars en la construcció de models *QSAR*. Així s'han reproduït alguns models *QSAR* que prèviament havien estat generats com una combinació lineal dels paràmetres  $\log P$  i la constant  $\sigma$  de Hammett, però ara emprant els seus equivalents teòrics basats en *QS-SM*. Una vegada corroborada l'eficàcia dels nous descriptors, s'han establert les bases d'una aproximació *QSAR* fonamentada en la semblança de fragments moleculars. En concret s'ha desenvolupat un mètode general capaç d'identificar les regions característiques d'una sèrie homogènia de compostos que millor descriuen una propietat molecular, sense cap restricció o especificació imposada a priori. El procediment permet la detecció de les regions moleculars comunes a tota la sèrie molecular que són responsables d'una alta resposta biològica.

També s'ha inclòs en la tesi un resum dels treballs iniciats a partir de l'estada a l'IQC dels professors David L. Cooper de la Universitat de Liverpool i Neil L. Allan de la Universitat de Bristol a l'estiu de l'any 1999. L'objectiu de la col·laboració va ser l'estudi de diferents aplicacions de les mesures de semblança basades en *DF* espacials de posició i de moment. En l'apartat *Classificació de camins de reacció mitjançant mesures de semblança* del capítol 3, s'analitza el comportament tipus Hammond i anti-Hammond d'unes reaccions de reordenació intramolecular. Els principals motius d'annexionar aquest treball en el capítol *Funcions densitat ASA* són, d'una banda, comprovar que és indistint utilitzar les *DF ab initio* i *PASA* per determinar el comportament de les reaccions estudiades, i en segon terme mostrar un exemple d'aplicació de les *MQSM* no relacionat amb les anàlisis *QSAR* dels capítols 6 i 7. És precisament en el capítol 7 on s'adjunta un segon exemple de la col·laboració amb N. L. Allan i D. L. Cooper, que fa referència a l'estudi de l'efecte dels substituents en sèries de compostos aromàtics per mitjà dels anomenats moments del moment. Ha estat una conseqüència dels treballs iniciats amb R. Ponec i corrobora els resultats obtinguts amb les *MQSM* emprant *DF* de posició.

## 1.2 Agraïments

Voldria agrair a totes les persones que en un moment o altre m'han ajudat en el desenvolupament d'algun dels treballs de recerca que presento en aquesta tesi. En primer lloc al meu director de tesi, el prof. Ramon Carbó-Dorca, que és la persona que ha proposat les línies de recerca a seguir, i gràcies als seus comentaris i indicacions a nivell científic ha estat possible realitzar el treball que aquí presento. També voldria recordar i agrair als companys de l'IQC amb els quals he treballat més estretament i que conjuntament hem publicat algun article de recerca. Cronològicament el primer ha estat en Xavier Fradera. Amb ell vaig coincidir durant la carrera de química, i també vam iniciar junts els nostres treballs de recerca a l'IQC. Un agraïment especial és per l'Emili Besalú, perquè sempre ha estat disposat a resoldre'm els problemes de caire científic i relacionats amb la programació que me s'han plantejat. També vull recordar en Pere Constans. Algun dels treballs que aquí presento són una continuació de la línia de recerca que ell va iniciar a l'IQC, sobretot els relacionats amb l'ajust de les *ASA DF* i l'algorisme de sobreposició molecular. Dels meus primers anys a l'IQC també vull mencionar l'amistat i el treball realitzat amb en Miquel Lobato. I més recentment en David Robert i en Xavier Gironés, que gràcies al seu interès i entusiasme ha estat possible avançar en diferents àmbits de la recerca, com ho reflecteixen els nombrosos articles que hem publicat junts. Per descomptat vull agrair l'ajuda que he rebut de la resta de persones amb les quals he coincidit tots aquests anys a l'IQC, i que han estat moltes. També em sento especialment agraït als professors visitants amb els quals he iniciat noves línies de recerca, en concret els investigadors Robert Ponec, Neil L. Allan i David L. Cooper. Els mesos que han estat a l'IQC han suposat, en certa manera, estades meves a altres centres de recerca però sense moure'm de casa. I finalment vull recordar als meus pares i germans, que sense el seu suport incondicional no hauria arribat fins aquí.

## Referències

1. M. A. Johnson, G. Maggiora (Eds.). Concepts and applications of molecular similarity. John Wiley & Sons, Inc., New York, 1990.
2. H. Kubinyi (ed.). 3D QSAR in drug design: theory methods and applications. ESCOM Science Publishers B.V., Leiden, The Netherlands, 1993.
3. F. Sanz, J. Giraldo, F. Manaut (eds.). QSAR and molecular modeling: concepts, computational tools and biological applications. Proceedings of the 10th European Symposium on SAR, QSAR and molecular modeling. Prous Science, Barcelona, 1995.
4. P. M. Dean (ed.). Molecular similarity in drug design. Blackie Academic & Professional, London, 1995.
5. K. Sen (Ed.). Molecular similarity. *Topics in Current Chemistry*. Springer Verlag, Berlin, volums 173 i 174, 1995.
6. R. Carbó (Ed.). Molecular similarity and reactivity: from quantum chemical to phenomenological approaches. Understanding chemical reactivity. Kluwer Academic, volum 14, Dordrecht, 1995.
7. M. Charton (Ed.). Advances in quantitative structure-property relationships. JAI Press, London, volum 1, 1996.
8. R. Carbó-Dorca, P. G. Mezzey (Eds.). Advances in molecular similarity. JAI Press, London, volum 1, 1996. Volum 2, 1998.
9. H. van de Waterbeemd, B. Testa, G. Folkers (eds.). Computer-assisted lead finding and optimization. Proceedings of the 11th European Symposium QSAR. Lausanne, 1996. Verlag: *Helvetica Chimica Acta*, Basel, i VCH, Weinheim, 1997.
10. R. Carbó-Dorca, D. Robert, Ll. Amat, X. Gironés, E. Besalú. Molecular quantum similarity in QSAR and drug design. *Lecture Notes in Chemistry*. Springer Verlag, Berlin, volum 73, 2000.
11. R. Carbó-Dorca, X. Gironés, P. G. Mezzey (Eds.). Fundamentals of molecular similarity. Kluwer Academic/Plenum Press, New York, 2001.

## 2. Semblança molecular quàntica

---

Un dels conceptes que ha contribuït de manera més notòria en el desenvolupament de la química és la idea de la semblança. No és d'estranyar que l'aplicabilitat d'aquest concepte sigui molt gran i inclogui pràcticament totes les àrees de la química.<sup>1</sup> Un exemple és la llei periòdica de Mendeleev, el descobriment de la qual està estretament connectat amb els esforços per classificar i sistematitzar les semblances en les propietats dels elements i els seus compostos més simples. A partir del significat intuïtiu de la semblança sorgeix un dels principis químics més poderosos, el principi d'analogia, que en els inicis de la química va servir de fonament per a la classificació de molècules i reaccions. El mateix principi serveix també per argumentar la idea que estructures semblants tenen propietats semblants, que al mateix temps, és el fonament de l'existència de diverses relacions empíriques entre l'estructura i l'activitat conegudes com *QSAR*.

Degut al fonamental paper que té la semblança en moltes situacions diferents, no sorprèn que la seva investigació sistemàtica hagi esdevingut el focus d'un intens interès científic. Però abans d'aplicar els criteris de semblança a qualsevol context és necessari establir-ne la seva quantificació. Una atenció preferent s'ha dedicat a la descripció de noves mesures quantitatives de la semblança molecular. En la bibliografia es pot trobar una gran varietat de descriptors moleculars, molts d'ells derivats de raonaments teòrics. En el present treball s'analitzaran unes mesures de semblança fonamentades en la teoria quàntica, les quals s'han deduït a partir de la idea que les propietats de les molècules, ja siguin químiques, físiques o biològiques, poden ser predeterminades per mitjà de la seva estructura electrònica. Com a conseqüència, algunes mesures que caracteritzen l'estructura electrònica són utilitzades en el disseny de nous descriptors moleculars teòrics.

## 2.1 Introducció

La primera mesura quantitativa de la semblança entre molècules fonamentada en els elements de la mecànica quàntica va ser formulada per R. Carbó i col·laboradors l'any 1980.<sup>2</sup> A partir de plantejar-se la simple qüestió de quant semblant són dues molècules, van proposar una mesura de semblança entre les densitats electròniques dels sistemes estudiats. Des d'aleshores ençà s'ha desenvolupat progressivament una àmplia teoria al voltant de les *MQSM*,<sup>2-76</sup> fins a arribar a l'actual esquema de treball que se segueix en el nostre laboratori. Alguns dels avenços s'han produït en la recerca de nous procediments i algorismes matemàtics associats amb les *MQSM*. Així, per exemple, mencionar la deducció de funcions de densitat electrònica aproximades per ser utilitzades en *MQSM*,<sup>13-15,19,20,26,34,41,42,55,60</sup> o el disseny de nous algorismes de superposició molecular.<sup>25,62</sup> Altres estudis s'han dirigit a la construcció d'una definició rigorosa de les mesures de semblança quàntica i a proporcionar els fonaments per al seu significat mecanicoquàntic.<sup>14,19,29-32,34,55,69</sup> Aquests articles aprofundeixen en la natura quàntica de la definició de la semblança, i permeten connectar-la amb diferents àmbits de la ciència.

Al llarg dels anys s'ha constatat que les *MQSM* tenen una àmplia aplicació en moltes àrees de la química. La més rellevant ha estat l'anàlisi *QSAR/QSPR*,<sup>19,27,28,39,45,47,51,54-59,63,67-73</sup> i no únicament enfocada a la racionalització i predicció de l'activitat de fàrmacs, sinó també a l'estudi de la toxicitat,<sup>49,50,66,69</sup> a la descripció de les constants de dissociació dels àcids carboxílics,<sup>46</sup> a l'estabilitat de proteïnes en mutacions d'un sol aminoàcid,<sup>53</sup> o fins i tot a la determinació de la quiralitat de molècules.<sup>48</sup> Malgrat que en els darrers anys s'ha dedicat una atenció preferent a les anàlisis *QSAR/QSPR*, les *MQSM* s'han aplicat amb èxit en altres àmbits. D'entre ells destaquen les aplicacions relacionades amb la reactivitat molecular,<sup>23</sup> el raonament del comportament d'algunes reaccions intramoleculares a partir del postulat de Hammond<sup>12,76</sup> i el principi de màxima duresa,<sup>44</sup> les anàlisis comparatives mitjançant *MQSM* de distribucions de densitat electrònica derivades de diferents metodologies de càlcul,<sup>18,22</sup> la determinació de la qualitat d'un conjunt de funcions de base,<sup>24</sup> o els estudis per millorar els paràmetres dels potencials emprats en càlculs de la teoria del funcional de la densitat.<sup>35,65</sup>

El principal camp d'aplicació de les *MQSM* són les molècules, però la definició de la semblança quàntica és suficientment general per incloure la comparació entre altres tipus d'objectes quàntics. Per exemple, semblança entre àtoms,<sup>21,22</sup> nuclis atòmics,<sup>36,43</sup> i fragments moleculars.<sup>45-48,63,71</sup> També s'ha establert una nova metodologia que connecta la semblança molecular quàntica i la teoria de grafs, a partir de la qual s'han definit els índexs topològics de semblança.<sup>28,34,68,72</sup>

Altres grups de recerca han desenvolupat noves tècniques basades en la semblança quàntica en les quals les densitats electròniques són comparades quantitativament mitjançant una gran varietat de mesures de semblança. A destacar els estudis fets per N. L. Allan i D. L. Cooper, que són els propulsors d'unes mesures de semblança definides sobre funcions densitat espacials de moment.<sup>76-84</sup> O els primers treballs del grup de W. G. Richards,<sup>85-88</sup> on es van estudiar mesures de semblança derivades de funcions densitat. Un altre exemple són les mesures proposades per J. Cioslowski.<sup>89-92</sup> També mencionar la línia d'investigació iniciada per R. Ponec, orientada a dilucidar els efectes dels electrons en determinades reaccions orgàniques mitjançant mesures de semblança.<sup>93-99</sup> Finalment citar els treballs de P.G. Mezey referents als càlculs de semblança relacionats amb la topologia molecular.<sup>100-103</sup> Però la densitat electrònica no ha estat l'únic descriptor mecanicoquàntic emprat en la quantificació de la semblança entre molècules. Una de les opcions més esteses és la definició de mesures de semblança entre potencials electrostàtics.<sup>104-117</sup>

La principal conseqüència de la gran diversitat de metodologies ha estat el desenvolupament de nous esquemes computacionals derivats bàsicament del tipus de mesura de semblança definit. Una àmplia revisió de les diferents tècniques fonamentades en la semblança molecular quàntica es pot trobar en els llibres [118,119], que són una síntesi de les conferències donades en els *Symposiums on Molecular Similarity* organitzats per l'IQC els anys 1995, 1997 i 1999.



## 2.2 Descripció dels sistemes quàntics

El progressiu desenvolupament de la semblança quàntica ha anat acompanyat en els darrers anys de l'interès a aprofundir en la naturalesa mecanicoquàntica de la definició de les *MQSM*.<sup>29-32,55,69</sup> L'objectiu és donar una definició i interpretació de les mesures de semblança quàntica d'acord amb els principis de la mecànica quàntica, i alhora construir un marc teòric per a les *MQSM* el més general possible. En primer lloc s'ha de tenir en compte la descripció quàntica d'un sistema microscòpic, que s'associa essencialment amb tres processos:

- a) Construir l'operador d'Hamilton,  $H$
- b) Calcular el parell energia–funció d'ona de l'estat,  $\{E, \Psi\}$ , a partir de l'equació de Schrödinger:  $H \Psi = E\Psi$
- c) Avaluar la *DF* de l'estat,  $\mathbf{r} = \Psi^* \Psi = |\Psi|^2$

Un dels postulats de la mecànica quàntica manifesta que coneguda la funció d'ona d'un estat, totes les propietats observables del sistema,  $\mathbf{w}$ , poden ser formalment deduïdes a partir d'ella, com a valors esperats,  $\langle \mathbf{w} \rangle$ , d'un operador hermític associat,  $\Omega$ , que actua sobre la corresponent funció:

$$\langle \mathbf{w} \rangle = \langle \Psi | \Omega | \Psi \rangle = \int \Psi^*(\mathbf{r}) \Omega(\mathbf{r}) \Psi(\mathbf{r}) d\mathbf{r} . \quad (2.1)$$

Si l'observable físic està associat a un operador hermític no diferencial, llavors es pot expressar el valor esperat de l'operador  $\Omega$  en funció de la densitat electrònica:

$$\langle \mathbf{w} \rangle = \int \Omega(\mathbf{r}) \mathbf{r}(\mathbf{r}) d\mathbf{r} = \langle \Omega | \mathbf{r} \rangle \quad (2.2)$$

La contribució de la mecànica quàntica al desenvolupament de la química i altres ciències ha significat un canvi substancial en la seva comprensió i interpretació. Des dels inicis de la mecànica quàntica,<sup>120-123</sup> s'ha plantejat la possibilitat d'explicar el comportament experimental i observable dels sistemes microscòpics, com són els àtoms i les molècules, mitjançant l'estudi de la funció de densitat de probabilitat que es construeix com el mòdul al quadrat de la funció d'ona del sistema considerat. En aquest

context, i com una extensió de la teoria quàntica, es defineixen les *MQSM* com una mesura del volum superposat entre les distribucions de densitat electrònica dels sistemes analitzats.

## 2.3 Funció de densitat electrònica de primer ordre

Si bé la funció d'ona  $\Psi$  d'un sistema microscòpic, i per extensió la *DF* de probabilitat definida com  $|\Psi|^2$ , contenen tota la informació que es pot conèixer del sistema que descriuen, en sistemes amb un gran nombre de partícules es fa difícil de tractar. En canvi, la distribució de densitat electrònica de primer ordre, expressada en termes de la funció d'ona com

$$\mathbf{r}^{(1)}(\mathbf{r}_1) = N \int \dots \int \Psi^*(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N) \Psi(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N) d\mathbf{x}_1 d\mathbf{x}_2 \dots d\mathbf{x}_N, \quad (2.3)$$

té l'avantatge de condensar tota la informació quan  $N$  és prou gran, i ser molt més manejable que  $\Psi$ . A més,  $\mathbf{r}^{(1)}(\mathbf{r}_1)$  és un observable físic sobre el qual altres propietats moleculars en depenen directa o indirectament. Degut a que en tot el treball només s'utilitzaran funcions densitat de primer ordre, a partir d'ara s'utilitzarà la lletra grega  $\mathbf{r}$  sense cap superíndex.

Emprant l'aproximació *LCAO-MO* (*Linear Combination of Atomic Orbitals – Molecular Orbital*)<sup>124</sup> cada orbital molecular s'expressa com una combinació lineal de funcions de base, que usualment són orbitals atòmics. Llavors la densitat electrònica de primer ordre es defineix com una doble suma sobre tots els parells de funcions de base:

$$\mathbf{r}(\mathbf{r}) = \sum_{mn} D_{mn} \mathbf{c}_m^*(\mathbf{r}) \mathbf{c}_n(\mathbf{r}), \quad (2.4)$$

on  $\{D_{mn}\}$  són els elements de la matriu densitat o també anomenada matriu de càrregues i ordres d'enllaç, i  $\{\mathbf{c}_m\}$  són els orbitals atòmics.

Els càlculs a nivell *ab initio* comporten un elevat cost computacional i únicament són possibles en sistemes moleculars petits i conjunts reduïts de molècules. En els primers treballs de *MQSM*,<sup>2-11</sup> es va utilitzar una aproximació de tipus *CNDO*, on la densitat electrònica es descrivia únicament per mitjà de les funcions esfèriques de les capes de valència. Posteriorment s'han desenvolupat algorismes d'ajust de la densitat *ab initio* a una combinació lineal de funcions esfèriques centrades en els àtoms. Un dels exemples és l'aproximació *ASA*,<sup>15,19,20,26,34,41,42,55,60</sup> on la densitat electrònica s'expressa com:

$$\mathbf{r}^{ASA}(\mathbf{r}) = \sum_i w_i |s_i(\mathbf{r} - \mathbf{r}_a)|^2, \quad (2.5)$$

i els coeficients  $\{w_i\}$  tenen la peculiaritat de ser definits positius per tal d'assegurar el significat físic de la densitat ajustada. El més usual és emprar funcions Gaussians  $1s$  per la seva simplicitat. En el capítol 3 es dóna una descripció més detallada de les funcions *ASA*, així com la metodologia emprada per determinar-ne els coeficients i exponents òptims.

## 2.4 Densitat electrònica d'un fragment molecular

Recentment s'ha desenvolupat una nova metodologia basada en mesures de semblança quàntica de fragments moleculars.<sup>45-48,63,71</sup> L'ús de descriptors moleculars definits sobre fragments es fonamenta, en part, en el teorema hologràfic de la densitat electrònica,<sup>125</sup> segons el qual tota la informació continguda en la densitat electrònica global d'una molècula també es troba inclosa en la densitat local de qualsevol fragment amb volum no zero de la molècula.

Emprant l'aproximació *LCAO-MO*, la *DF* d'un fragment *X* pertanyent a una molècula *A* es pot definir com:

$$\mathbf{r}_A^X(\mathbf{r}) = \sum_{\mathbf{n} \in X} \sum_{\mathbf{m} \in A} D_{\mathbf{m}} \mathbf{c}_{\mathbf{m}}^*(\mathbf{r}) \mathbf{c}_{\mathbf{n}}(\mathbf{r}). \quad (2.6)$$

En l'equació (2.6) el sumatori de  $\mathbf{n}$  es calcula sobre totes les funcions de base de la molècula *A*, mentre que el sumatori  $\mathbf{m}$  s'executa únicament per les funcions de base

centrades en els àtoms pertanyents al fragment estudiat  $X$ . Aquesta definició de la densitat d'un fragment proporciona una partició additiva de la densitat electrònica global de la molècula.<sup>126</sup> Per exemple, en el cas més simple de la divisió de la densitat, corresponent a definir tots els fragments formats per un sol àtom, es pot generar la densitat de tota la molècula sumant totes les contribucions atòmiques:

$$\mathbf{r}_A(\mathbf{r}) = \sum_a \mathbf{r}_a(\mathbf{r}) \quad \wedge \quad \mathbf{r}_a(\mathbf{r}) = \sum_{m \in a} \sum_{n \in A} D_m \mathbf{c}_m^*(\mathbf{r}) \mathbf{c}_n(\mathbf{r}). \quad (2.7)$$

Quant a les funcions *ASA*, la descripció d'un fragment molecular és molt més simple i admet una única definició. Així, la densitat electrònica associada a un fragment molecular  $X$  s'expressa segons l'equació:

$$\mathbf{r}_{X,A}^{ASA}(\mathbf{r}) = \sum_{i \in X} w_i |s_i(\mathbf{r} - \mathbf{r}_a)|^2. \quad (2.8)$$

## 2.5 Mesures de semblança quàntica

Les *MQSM* són un mitjà per quantificar les semblances entre densitats electròniques de sistemes quàntics diferents. A la pràctica, les *MQSM* es defineixen com la integral entre les densitats electròniques de dos objectes quàntics  $\{A, B\}$  ponderada per un operador bielectrònic definit positiu:

$$Z_{AB}(\Omega) = \int \int \mathbf{r}_A(\mathbf{r}_1) \Omega(\mathbf{r}_1, \mathbf{r}_2) \mathbf{r}_B(\mathbf{r}_2) d\mathbf{r}_1 d\mathbf{r}_2, \quad (2.9)$$

on  $\{\mathbf{r}_A, \mathbf{r}_B\}$  són les *DF* de primer ordre dels objectes  $A$  i  $B$ .

Donat un conjunt de  $n$  objectes  $M$  i les seves corresponents funcions densitat  $\mathbf{r}_I$ , es defineix la matriu de semblança  $\mathbf{Z} = \{Z_{IJ}(\Omega) | \forall I, J \in M\}$ , de dimensió  $(n \times n)$ , els elements de la qual són les mesures de semblança entre tots els possibles parells de funcions densitat del conjunt d'objectes considerat.

Dins l'aproximació *LCAO-MO*, una *MQSM* respon a la fórmula general:

$$Z_{AB}(\Omega) = \sum_{m \in A} D_{mm} \sum_{l, s \in B} D_{ls} \iint \mathbf{c}_m^*(\mathbf{r}_1) \mathbf{c}_n(\mathbf{r}_1) \Omega(\mathbf{r}_1, \mathbf{r}_2) \mathbf{c}_l^*(\mathbf{r}_2) \mathbf{c}_s(\mathbf{r}_2) d\mathbf{r}_1 d\mathbf{r}_2. \quad (2.10)$$

En aquest nivell de càlcul s'han de resoldre integrals de quatre centres.

Una fórmula molt més simplificada pren la integral de semblança quan s'utilitzen les densitats *ASA*,

$$Z_{AB}(\Omega) = \sum_{i \in A} \sum_{j \in B} w_i w_j Z_{ij}(\Omega), \quad (2.11)$$

essent  $Z_{ij}$  la integral entre la capa  $i$  pertanyent a l'àtom  $a$  de l'objecte  $A$  i la capa  $j$  pertanyent a l'àtom  $b$  de l'objecte  $B$ :

$$Z_{ij}(\Omega) = \iint |s_i(\mathbf{r}_1 - \mathbf{r}_a)|^2 \Omega(\mathbf{r}_1, \mathbf{r}_2) |s_j(\mathbf{r}_2 - \mathbf{r}_b)|^2 d\mathbf{r}_1 d\mathbf{r}_2. \quad (2.12)$$

Emprant *ASA DF* com a màxim s'hauran de calcular integrals de dos centres entre funcions  $1s$ .

Segons l'operador  $\Omega(\mathbf{r}_1, \mathbf{r}_2)$  escollit en l'equació (2.9) es defineixen diferents tipus de *MQSM*. Tot seguit es descriuen les dues mesures més comunes i que s'utilitzen més assíduament en la deducció de relacions estructura-activitat.

### 2.5.1 *MQSM de solapament o recobriment*

Es defineix com el solapament de les densitats electròniques dels objectes comparats, i ve donada per la integral:

$$Z_{AB} = \int \mathbf{r}_A(\mathbf{r}) \mathbf{r}_B(\mathbf{r}) d\mathbf{r}. \quad (2.13)$$

És la mesura més simple i intuïtiva de la similitud entre les densitats  $\mathbf{r}_A$  i  $\mathbf{r}_B$ , que ja va ser proposada en l'article original de l'any 1980,<sup>2</sup> i que ha servit de fonament per al

posterior desenvolupament de les *MQSM*. Respecte a l'equació general (2.9), equival a substituir l'operador bielectrònic  $\Omega(\mathbf{r}_1, \mathbf{r}_2)$  per la funció delta de Dirac  $\mathbf{d}(\mathbf{r}_1 - \mathbf{r}_2)$ .

### 2.5.2 *MQSM de Coulomb*

Una altra *MQSM* molt estesa en els estudis de semblança és la mesura de Coulomb,<sup>3</sup>

$$Z_{AB}(\mathbf{r}_{12}^{-1}) = \int \int \mathbf{r}_A(\mathbf{r}_1) |\mathbf{r}_1 - \mathbf{r}_2|^{-1} \mathbf{r}_B(\mathbf{r}_2) d\mathbf{r}_1 d\mathbf{r}_2, \quad (2.14)$$

on  $\Omega(\mathbf{r}_1, \mathbf{r}_2) = |\mathbf{r}_1 - \mathbf{r}_2|^{-1}$  és l'operador usual en els càlculs de l'energia de repulsió bielectrònica.

## 2.6 Mesures d'autosemblança quàntica

Un cas particular de *MQSM* és quan els objectes quàntics comparats són un mateix. Aleshores es defineix la mesura d'autosemblança (*Quantum Self-Similarity Measure, QS-SM*) d'acord amb:

$$Z_{AA}(\Omega) = \int \int \mathbf{r}_A(\mathbf{r}_1) \Omega(\mathbf{r}_1, \mathbf{r}_2) \mathbf{r}_A(\mathbf{r}_2) d\mathbf{r}_1 d\mathbf{r}_2. \quad (2.15)$$

Les quantitats  $Z_{AA}$  s'han utilitzat en qualitat de simples descriptors moleculars en anàlisis *QSAR*.<sup>33,45-48,63,71</sup> S'ha comprovat que és possible descriure certes propietats moleculars, com la hidrofobicitat i els efectes electrònics produïts pels substituents, emprant *QS-SM* calculades sobre la *DF* de tota la molècula o bé localitzades en el fragment responsable de l'activitat biològica. En el capítol 7 es descriu la metodologia desenvolupada al voltant de les *QS-SM* a més de presentar una gran varietat d'exemples.

## 2.7 Mesures de semblança quàntica de moment

N. L. Allan i D. L. Cooper han desenvolupat tota la teoria referent a les mesures de semblança quàntica basades en funcions de moment enlloc de posició.<sup>78-84</sup> Per sistemes moleculars la millor manera de calcular quantitats relacionades amb el moment és determinar primer la funció d'ona espacial de posició  $\Psi(\mathbf{r})$ , i després transformar-la analíticament per obtenir la funció espacial de moment  $\Psi(\mathbf{p})$ .<sup>81</sup> La transformació de Fourier que relaciona ambdues funcions és:

$$\Psi(\mathbf{p}) = \frac{1}{(2\pi)^{3/2}} \int \Psi(\mathbf{r}) \exp(-i\mathbf{p}\cdot\mathbf{r}) d\mathbf{r}, \quad (2.16)$$

la qual conserva la configuració d'una funció d'ona, i així la densitat electrònica, els orbitals individuals, les funcions de base i la funció d'ona total en l'espai de  $p$  estan relacionades entre elles de la mateixa manera que ho estan en l'espai de  $r$ . Les mesures de semblança quàntica entre funcions densitat espacials de moment s'expressa segons la fórmula:

$$I_{AB}(n) = \int p^n \mathbf{r}_A(\mathbf{p}) \mathbf{r}_B(\mathbf{p}) d\mathbf{p}. \quad (2.17)$$

Altres quantitats emprades en estudis de semblança són els valors esperats en l'espai del moment  $\langle p^n \rangle$ , els anomenats moments del moment, que es defineixen com

$$\langle p^n \rangle = \int p^n \mathbf{r}(\mathbf{p}) d\mathbf{p}, \quad (2.18)$$

i es poden calcular convenientment per integració numèrica. Com a casos particulars,  $\langle p^0 \rangle$  és igual al nombre d'electrons i  $\langle p^2 \rangle$  és dues vegades l'energia cinètica.

## 2.8 Índexs de semblança quàntica

Transformacions de les *MQSM* donen els índexs de semblança, que també es poden emprar en qualitat de descriptors moleculars en estudis *QSAR*. El primer en definir-se va ser l'índex de Carbó,<sup>2</sup> expressat com:

$$C_{AB} = Z_{AB} (Z_{AA} Z_{BB})^{-1/2}. \quad (2.19)$$

L'índex de Carbó equival a una normalització de la mesura  $Z_{AB}$  respecte als valors de les *QS-SM* de *A* i *B*. L'índex  $C_{AB}$  varia en l'interval  $[0,1]$ , amb la característica que quan més proper a u sigui el valor més semblants seran els objectes comparats, mentre que un valor pròxim a zero indica que els objectes són molt dissemblants.

Una altra transformació de les *MQSM* utilitzada en algun dels treballs que es presentaran en els propers capítols és la distància euclidiana,<sup>3</sup> que es defineix com:

$$D_{AB} = (Z_{AA} + Z_{BB} - 2Z_{AB})^{1/2} \quad (2.20)$$

La variació de  $D_{AB}$  es produeix en l'interval  $[0,\infty)$ , però al contrari de l'índex de Carbó, valors pròxims a zero impliquen una gran similitud entre els objectes comparats.

En la bibliografia es poden trobar altres índexs de semblança definits per altres autors. Els més usuals són l'índex de Hodgkin-Richards,<sup>104,109</sup> l'índex de Tanimoto,<sup>127</sup> i l'índex de Petke.<sup>114</sup> En un treball recent s'han agrupat els diferents índex en dues classes: els índexs de correlació (classe C) i els índex de distància (classe D).<sup>17</sup>



## 2.9 Superposició molecular

En l'estudi de sistemes moleculars s'utilitza normalment l'aproximació de Born-Oppenheimer,<sup>128,129</sup> que desacobla els moviments electrònic i nuclear. Aleshores les molècules es poden descriure mitjançant les coordenades espacials dels nuclis, normalment fixes en alguna conformació, i les densitats de probabilitat electrònica. Això fa que les *MQSM* siguin dependents de la posició relativa de les molècules comparades en l'espai. Aquesta dependència de la mesura de semblança es pot incloure d'una manera explícita en la definició general donada en l'equació (2.9) d'acord amb:

$$Z_{AB}(\Omega; \mathbf{Q}) = \int \int \mathbf{r}_A(\mathbf{r}_1) \Omega(\mathbf{r}_1, \mathbf{r}_2) \mathbf{r}_B(\mathbf{r}_2; \mathbf{Q}) d\mathbf{r}_1 d\mathbf{r}_2, \quad (2.21)$$

on el vector  $\mathbf{Q}$  representa les tres translacions més les tres rotacions de la molècula mòbil, *B*, respecte a la molècula fixa, *A*.

Una línia d'investigació molt important en l'àmbit de la semblança molecular és el desenvolupament d'algorismes per cercar l'alineament molecular òptim. En el nostre laboratori s'han dissenyat dos algorismes de superposició molecular, un fonamentat en la recerca de la disposició de les molècules en l'espai que fa màxim el valor de la integral de semblança,<sup>25</sup> i un segon mètode orientat a trobar la màxima similitud estructural de les molècules comparades.<sup>62</sup> En el capítol 4 s'explica el funcionament de les dues tècniques d'alineament molecular.

## Referències

1. D.H. Rouvray. Similarity in chemistry: past, present and future. *Topics in Current Chemistry* **1995**, 173, 1–30.
2. R. Carbó, L. Leyda, M. Arnau. How Similar is a molecule to another? An electron density measure of similarity between two electronic structures. *Int. J. Quantum Chem.* **1980**, 17, 1185–1189.
3. R. Carbó, Ll. Domingo. LCAO-MO similarity measures and taxonomy. *Int. J. Quantum Chem.* **1987**, 32, 517–545.
4. R. Carbó, B. Calabuig. Molsimil-88: molecular similarity calculations using a CNDO-like approximation. *Comp. Phys. Commun.* **1989**, 55, 117–126.
5. R. Carbó, B. Calabuig. Molecular similarity and quantum chemistry. Publicat en el llibre: Concepts and applications of molecular similarity, A. Johnson i G. M. Maggiora (Eds.). John Wiley & Sons, Inc. New York, 1990.
6. R. Carbó, B. Calabuig. Quantum molecular similarity measures and the  $n$ -dimensional representation of molecular set: phenyldimethylthiazines. *J. Mol. Struct. (Theochem)* **1992**, 254, 517–531.
7. R. Carbó, B. Calabuig. Molecular quantum similarity measures and  $n$ -dimensional representation of quantum objects. I. Theoretical foundations. *Int. J. Quantum Chem.* **1992**, 42, 1681–1693.
8. R. Carbó, B. Calabuig. Molecular quantum similarity measures and  $n$ -dimensional representation of quantum objects. II. Practical applications. *Int. J. Quantum Chem.* **1992**, 42, 1695–1709.
9. R. Carbó, B. Calabuig, E. Besalú, A. Martínez. Triple density molecular quantum similarity measures: a general connection between theoretical calculations and experimental results. *Molecular Engineering* **1992**, 2, 43–64.
10. R. Carbó, B. Calabuig. Quantum similarity measures, molecular cloud description, and structure-properties relationships. *J. Chem. Inf. Comput. Sci.* **1992**, 32, 600–606.
11. R. Carbó, B. Calabuig, L.Vera, E. Besalú. Molecular quantum similarity: theoretical framework, ordering principles, and visualization techniques. *Adv. in Quantum Chem.* **1994**, 25, 253–313.
12. M. Solà, J. Mestres, R. Carbó, M. Duran. Use of *ab initio* quantum similarity measures as an interpretative tool for the study of chemical reactions. *J. Am. Chem. Soc.* **1994**, 116, 5909–5915.
13. J. Mestres, M. Solà, M. Duran, R. Carbó. On the calculation of *ab initio* quantum molecular similarities for large systems: fitting the electron density. *J. Comput. Chem.* **1994**, 15, 1113–1120.
14. E. Besalú, R. Carbó, J. Mestres, M. Solà. Foundations and recent developments on molecular quantum similarity. *Topics in Current Chemistry* **1995**, 173, 31–62.
15. P. Constans, R. Carbó. Atomic shell approximation: electron density fitting algorithm restricting coefficients to positive values. *J. Chem. Inf. Comput. Sci.* **1995**, 35, 1046–1053.
16. R. Carbó, E. Besalú, Ll. Amat, X. Fradera. Quantum molecular similarity measures (QMSM) as a natural way leading towards a theoretical foundation of quantitative structure-properties relationships (QSPR). *J. Math. Chem.* **1995**, 18, 237–246.
17. R. Carbó, E. Besalú, Ll. Amat, X. Fradera. On quantum molecular similarity measures (QMSM) and indices (QMSI). *J. Math. Chem.* **1996**, 19, 47–56.

18. M. Solà, J. Mestres, R. Carbó, M. Duran. A comparative analysis by means of quantum molecular similarity measures of density distributions derived from conventional *ab initio* and density functional methods. *J. Chem. Phys.* **1996**, *104*, 636–647.
19. R. Carbó-Dorca, E. Besalú, Ll. Amat, X. Fradera. Quantum molecular similarity measures: concepts, definitions, and applications to quantitative structure-property relationships. Publicat en el llibre: Advances in molecular similarity. R. Carbó-Dorca, P. G. Mezey (Eds.). JAI Press, London, volum 1, pàgines 1–41, 1996.
20. P. Constans, Ll. Amat, X. Fradera, R. Carbó-Dorca. Quantum molecular similarity measures (QMSM) and the atomic shell approximation (ASA). Publicat en el llibre: Advances in molecular similarity. R. Carbó-Dorca, P. G. Mezey (Eds.). JAI Press, London, volum 1, pàgines 187–211, 1996.
21. J. Cioslowski, B. B. Stefanov, P. Constans. Efficient algorithm for quantitative assessment of similarities among atoms in molecules. *J. Comput. Chem.* **1996**, *17*, 1352–1358.
22. M. Solà, J. Mestres, J. M. Oliva, M. Duran, R. Carbó. The use of *ab initio* quantum molecular self-similarity measures to analyze electronic charge density distributions. *Int. J. Quantum Chem.* **1996**, *58*, 361–372.
23. X. Fradera, Ll. Amat, M. Torrent, J. Mestres, P. Constans, E. Besalú, J. Martí, S. Simon, M. Lobato, J. M. Oliva, J. M. Luis, J. L. Andrés, M. Solà, R. Carbó, M. Duran. Analysis of the changes on the potential energy surface of Menshutkin reactions induced by external perturbations. *J. Mol. Struct. (Theochem)* **1996**, *371*, 171–183.
24. M. Forés, M. Duran, M. Solà. A procedure for assessing the quality of a given basis set based on quantum molecular similarity measures. *Theor. Mol. Mod. Electr. Conf.* **1997**, *1*, 50–56.
25. P. Constans, Ll. Amat, R. Carbó-Dorca. Toward a global maximization of the molecular similarity function: superposition of two molecules. *J. Comput. Chem.* **1997**, *18*, 826–846.
26. Ll. Amat, R. Carbó-Dorca. Quantum similarity measures under atomic shell approximation: first order density fitting using elementary Jacobi rotations. *J. Comput. Chem.* **1997**, *18*, 2023–2039.
27. X. Fradera, Ll. Amat, E. Besalú, R. Carbó-Dorca. Application of molecular quantum similarity to QSAR. *Quant. Struct.-Act. Relat.* **1997**, *16*, 25–32.
28. M. Lobato, Ll. Amat, E. Besalú, R. Carbó-Dorca. Structure-activity relationships of a steroid family using quantum similarity measures and topological quantum similarity indices. *Quant. Struct.-Act. Relat.* **1997**, *16*, 465–472.
29. R. Carbó-Dorca. Fuzzy sets and Boolean tagged sets. *J. Math. Chem.* **1997**, *22*, 143–147.
30. R. Carbó-Dorca. Tagged sets, convex sets and quantum similarity measures. *J. Math. Chem.* **1998**, *23*, 353–364.
31. R. Carbó-Dorca. On the statistical interpretation of density functions: atomic shell approximation, convex sets, discrete quantum chemical molecular representations, diagonal vector spaces and related problems. *J. Math. Chem.* **1998**, *23*, 365–375.
32. R. Carbó-Dorca, E. Besalú. A general survey of molecular quantum similarity. *J. Mol. Struct. (Theochem)* **1998**, *451*, 11–23.
33. Ll. Amat, R. Carbó-Dorca, R. Ponc. Molecular quantum similarity measures as an alternative to log P values in QSAR studies. *J. Comput. Chem.* **1998**, *19*, 1575–1583.
34. R. Carbó-Dorca, Ll. Amat, E. Besalú, M. Lobato. Quantum similarity. Publicat en el llibre: Advances in molecular similarity. R. Carbó-Dorca, P. G. Mezey (Eds.). JAI Press, London, volum 2, pàgines 1–41, 1998.

35. M. Solà, M. Forés, M. Duran. Optimizing hybrid density functionals by means of quantum molecular similarity techniques. Publicat en el llibre: *Advances in molecular similarity*. R. Carbó-Dorca, P. G. Mezey (Eds.). JAI Press, London, volum 2, pàgines 187–203, 1998.
36. D. Robert, R. Carbó-Dorca. On the extension of quantum similarity to atomic nuclei: nuclear quantum similarity. *J. Math. Chem.* **1998**, *23*, 327–351.
37. D. Robert, R. Carbó-Dorca. A formal comparison between molecular quantum similarity measures and indices. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 469–475.
38. D. Robert, R. Carbó-Dorca. Analyzing the triple density molecular quantum similarity measures with the INDSCAL model. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 620–623.
39. Ll. Amat, D. Robert, E. Besalú, R. Carbó-Dorca. Molecular quantum similarity measures tuned 3D QSAR: an antitumoral family validation study. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 624–631.
40. X. Fradera, M. Duran, J. Mestres. Second-order quantum similarity measures from intracule and extracule densities. *Theor. Chem. Acc.* **1998**, *99*, 44–52.
41. X. Gironés, Ll. Amat, R. Carbó-Dorca. A comparative study of isodensity surfaces using *ab initio* and ASA density functions. *J. Mol. Graph. Model.* **1998**, *16*, 190–196.
42. Ll. Amat, R. Carbó-Dorca. Fitted electronic density functions from H to Rn for use in quantum similarity measures: cis-diammine-dichloroplatinum(II) complex as an application example. *J. Comput. Chem.* **1999**, *20*, 911–920.
43. D. Robert, R. Carbó-Dorca. Structure-property relationships in nuclei. Prediction of the binding energy per nucleon using a quantum similarity approach. *Nuovo Cimento* **1999**, *A111*, 1311–1321.
44. M. Solà, A. Toro-Labbé. The Hammond Postulate and the principle of maximum hardness in some intramolecular rearrangement reactions. *J. Phys. Chem. A* **1999**, *103*, 8847–8852.
45. R. Ponec, Ll. Amat, R. Carbó-Dorca. Molecular basis of quantitative structure-properties relationships (QSPR): a quantum similarity approach. *J. Comput.-Aided Mol. Design* **1999**, *13*, 259–270.
46. R. Ponec, Ll. Amat, R. Carbó-Dorca. Quantum similarity approach to LFER: substituent and solvent effects on the acidities of carboxylic acids. *J. Phys. Org. Chem.* **1999**, *12*, 447–454.
47. Ll. Amat, R. Carbó-Dorca, R. Ponec. Simple linear QSAR models based on quantum similarity measures. *J. Med. Chem.* **1999**, *42*, 5169–5180.
48. P. G. Mezey, R. Ponec, Ll. Amat, R. Carbó-Dorca. Quantum similarity approach to the characterization of molecular chirality. *Enantiomer* **1999**, *4*, 371–378.
49. D. Robert, R. Carbó-Dorca. Aromatic compounds aquatic toxicity QSAR using molecular quantum similarity measures. *SAR QSAR Environ. Res.* **1999**, *10*, 401–422.
50. X. Gironés, Ll. Amat, R. Carbó-Dorca. Using molecular quantum similarity measures as descriptors in quantitative structure-toxicity relationships. *SAR QSAR Environ. Res.* **1999**, *10*, 545–556.
51. D. Robert, Ll. Amat, R. Carbó-Dorca. 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.
52. R. Carbó-Dorca. Stochastic transformation of quantum similarity matrices and their use in quantum QSAR (QQSAR) models. *Int. J. Quantum Chem.* **2000**, *79*, 163–177.

53. D. Robert, X. Gironés, R. Carbó-Dorca. Quantification of the influence of single-point mutations on haloalkane dehalogenase activity: a molecular quantum similarity study. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 839–846.
54. X. Gironés, Ll. Amat, D. Robert, R. Carbó-Dorca. Use of electron-electron repulsion energy as a molecular descriptor in *QSAR* and *QSPR* studies. *J. Comput.-Aided Mol. Design* **2000**, *14*, 477–485.
55. R. Carbó-Dorca, Ll. Amat, E. Besalú, X. Gironés, D. Robert. Quantum mechanical origin of *QSAR*: theory and applications. *J. Mol. Struct. (Theochem)* **2000**, *504*, 181–228.
56. R. Carbó-Dorca, D. Robert, Ll. Amat, X. Gironés, E. Besalú, Molecular quantum similarity in *QSAR* and drug design. *Lecture Notes in Chemistry*, 73, Springer Verlag, Berlin, 2000.
57. D. Robert, X. Gironés, R. Carbó-Dorca. Molecular quantum similarity measures as descriptors for quantum *QSAR*. *Polycyclic Aromatic Compounds* **2000**, *19*, 51–71.
58. X. Gironés, A. Gallegos, R. Carbó-Dorca. Modeling antimalarial activity: application of kinetic energy density quantum similarity measures as descriptors in *QSAR*. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1400–1407.
59. D. Robert, Ll. Amat, R. Carbó-Dorca. Quantum similarity *QSAR*: study of inhibitors binding to Trombin, Trypsin, and Factor Xa, including a comparison with CoMFA and CoMSIA methods. *Int. J. Quantum Chem.* **2000**, *80*, 265–282.
60. Ll. Amat, R. Carbó-Dorca. Molecular electronic density fitting using elementary Jacobi rotations under atomic shell approximation. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1188–1198.
61. R. Carbó-Dorca. Inward matrix products: estensions and applications to quantum mechanical foundations of *QSAR*. *J. Mol. Struct. (Theochem)* **2001**, *537*, 41–54.
62. X. Gironés, D. Robert, R. Carbó-Dorca. TGSA: a molecular superposition program based on topogeometrical considerations. *J. Comput. Chem.* **2001**, *22*, 255–263.
63. Ll. Amat, E. Besalú, R. Carbó-Dorca, R. Ponec. Identification of active molecular sites using quantum-self-similarity measures. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 978–991.
64. X. Gironés, R. Carbó-Dorca, P. G. Mezey. Applications of promolecular ASA densities to graphical representation of density functions of macromolecular systems. *J. Mol. Graph. Model.* **2001**, *19*, 343–348.
65. J. Poater, M. Duran, M. Solà. Parameterization of the Becke3-LYP hybrid functional for a series of small molecules using quantum molecular similarity techniques. *J. Comput. Chem.* **2001**, *22*, 1666–1678.
66. A. Gallegos, D. Robert, X. Gironés, R. Structure-toxicity relationships of polycyclic aromatic hydrocarbons using molecular quantum similarity. *J. Comput.-Aided Mol. Design* **2001**, *15*, 67–80.
67. X. Gironés, A. Gallegos, R. Carbó-Dorca. Antimalarial activity of synthetic 1,2,4-trioxane and cyclic peroxy ketals, a quantum similarity study. *J. Comput.-Aided Mol. Design* **2001**, *15*, 1053–1063.
68. E. Besalú, A. Gallegos, R. Carbó-Dorca. Topological quantum similarity indices and their use in *QSAR*: application of several families of antimalarial compounds. *MATHC-Communications in mathematical and computational chemistry.* **2001**, *44*, 41–64.
69. R. Carbó-Dorca, Ll. Amat, E. Besalú, X. Gironés, D. Robert. Quantum molecular similarity measures: theory and applications to the evaluation of molecular properties, biological activities and toxicity. Publicat en el llibre: Fundamentals of molecular similarity. R. Carbó-Dorca, X. Gironés, P. G. Mezey (Eds.). Kluwer Academic/Plenum Press, New York, 2001.

70. X. Gironés, R. Carbó-Dorca. Using molecular quantum similarity measures under stochastic transformation to describe physical properties of molecular systems. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 317–325.
71. R. Ponec, X. Gironés, R. Carbó-Dorca. Molecular basis of LFER. The nature of inductive effects in aliphatic series. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 564–570.
72. E. Besalú, X. Gironés, Ll. Amat, R. Carbó-Dorca. Molecular quantum similarity and the fundamentals of QSAR. *Acc. Chem. Res.* **2002**, *35*, 289–295.
73. X. Gironés, R. Carbó-Dorca. Molecular quantum similarity-based QSAR's for binding affinities of several steroids sets. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1185–1193.
74. Ll. Amat, R. Carbó-Dorca. Use of promolecular ASA density functions as a general algorithm to obtain starting MO in SCF calculations. *Int. J. Quantum Chem.* **2002**, *87*, 59–67.
75. X. Gironés, Ll. Amat, R. Carbó-Dorca. Modeling large macromolecular structures using promolecular densities. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 847–852.
76. Ll. Amat, R. Carbó-Dorca, D. L. Cooper, N. L. Allan. Classification of reaction pathways via momentum-space and quantum molecular similarity measures. *Chem. Phys. Lett.* **2003**, *367*, 207–213.
77. Ll. Amat, R. Carbó-Dorca, D. L. Cooper, N. L. Allan, R. Ponec. Structure-property relationships and momentum-space quantities: Hammett  $\sigma$  constants. *Mol. Phys.* **2003**, *en premsa*.
78. D. L. Cooper, N. L. Allan. A novel approach to molecular similarity. *J. Comput.-Aided Mol. Design* **1989**, *3*, 253–259.
79. D. L. Cooper, N. L. Allan. Molecular dissimilarity: a momentum-space criterion. *J. Am. Chem. Soc.* **1992**, *114*, 4773–4776.
80. N. L. Allan, D. L. Cooper. A momentum space approach to molecular similarity. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 587–590.
81. N. L. Allan, D. L. Cooper. Momentum-space electron densities and quantum molecular similarity. *Topics in Current Chemistry* **1995**, *173*, 85–111.
82. P. T. Measures, K. A. Mort, N. L. Allan, D. L. Cooper. Applications of momentum-space similarity. *J. Comput.-Aided Mol. Design* **1995**, *9*, 331–340.
83. P. T. Measures, N. L. Allan, D. L. Cooper. Momentum-space similarity: some recent applications. Publicat en el llibre: Advances in molecular similarity. R. Carbó-Dorca, P. G. Mezey (Eds.). JAI Press, London, volum 1, pàgines 61–87, 1996.
84. P. T. Measures, K. A. Mort, N. L. Allan, D. L. Cooper. A quantum molecular similarity approach to anti-HIV activity. *J. Mol. Struct. (Theochem)* **1998**, *423*, 113–123.
85. P. E. Bowen-Jenkins, D. L. Cooper, W. G. Richards. *Ab initio* computation of molecular similarity. *J. Phys. Chem.* **1985**, *89*, 2195–2197.
86. P. E. Bowen-Jenkins, W. G. Richards. Molecular similarity in terms of valence electron density. *J. Chem. Soc. Chem. Comm.* **1986**, 133–135.
87. P. E. Bowen-Jenkins, W. G. Richards. Quantitative measures of similarity between pharmacologically active compounds. *Int. J. Quant. Chem.* **1986**, *30*, 763–768.

88. E. E. Hodgkin, W. G. Richards. A semi-empirical method for calculating molecular similarity. *J. Chem. Soc. Chem. Comm.* **1986**, 1342–1344.
89. J. Cioslowski, E. D. Fleishmann. Assessing molecular similarity from results of *ab initio* electronic structure calculations. *J. Am. Chem. Soc.* **1991**, *113*, 64–67.
90. J. B. Ortiz, J. Cioslowski. Molecular similarity indices in electron propagator theory. *Chem. Phys. Lett.* **1991**, *185*, 270–275.
91. J. Cioslowski, M. Challacombe. Maximum similarity orbitals for analysis of the electronic excited states. *Int. J. Quant. Chem.* **1992**, *S25*, 81–93.
92. J. Cioslowski. Differential density matrix overlap: an index for assessment of electron correlation in atoms and molecules. *Theor. Chim. Acta.* **1992**, *81*, 319–327.
93. R. Ponec. Topological aspects of chemical reactivity. On the similarity of molecular structures. *Collect. Czech. Chem. Commun.* **1987**, *52*, 555–561.
94. R. Ponec, M. Strnad. Similarity approach to chemical reactivity. Specificity of multibond reactions. *Collect. Czech. Chem. Commun.* **1990**, *55*, 2583–2589.
95. R. Ponec, M. Strnad. Topological aspects of chemical reactivity. Evans/Dewar principle in terms of molecular similarity approach. *J. Phys. Org. Chem.* **1991**, *4*, 701–705.
96. R. Ponec. Similarity measures, the last motion principle and selection rules in chemical reactivity. *Z. Phys. Chemie* **1987**, *268*, 1180–1188.
97. R. Ponec, M. Strnad. Electron correlation in pericyclic reactivity: a similarity approach. *Int. J. Quant. Chem.* **1992**, *42*, 501–508.
98. R. Ponec. Similarity approach to chemical reactivity. A simple criterion for discriminating between one-step and stepwise reactions mechanism in pericyclic reactivity. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 805–811.
99. R. Ponec, M. Strnad. Position invariant index for assessment of molecular similarity. *Croatica Chem. Acta* **1993**, *66*, 123–127.
100. P. G. Mezey. Shape-similarity measures for molecular bodies: a three-dimensional topological approach to quantitative shape-activity relations. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 650–656.
101. P. G. Mezey. Shape in chemistry: an introduction to molecular shape and topology. VCH Publishers, New York, 1993.
102. P. D. Walker, P. G. Mezey. Molecular electron density: a lego approach to molecule building. *J. Am. Chem. Soc.* **1993**, *115*, 12423–12430.
103. P. G. Mezey. Density domain bonding topology and molecular similarity measures. *Topics in Current Chemistry* **1995**, *173*, 63–83.
104. E. E. Hodgkin, W. G. Richards. Molecular similarity based on electrostatic potential and electric field. *Int. J. Quant. Chem., Quant. Biol. Symp.* **1987**, *14*, 105–110.
105. F. J. Luque, F. Sanz, F. Illas, R. Pouplana, Y. G. Smeyers. Relationships between the activity of some H<sub>2</sub>-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.
106. C. Burt, W. G. Richards, P. Huxley. The application of molecular similarity calculations. *J. Comput. Chem.* **1990**, *11*, 1139–1146.

107. A. M. Richard. Quantitative comparison of molecular electrostatic potentials for structure activity studies. *J. Comput. Chem.* **1991**, *12*, 959–969.
108. A. M. Meyer, W. G. Richards. Similarity of molecular shape. *J. Comput.-Aided Mol. Design* **1991**, *5*, 426–439.
109. A. C. Good. The calculation of molecular similarity: alternative formulas, data manipulation and graphical display. *J. Mol. Graph.* **1992**, *10*, 144–151.
110. A. C. Good, E. E. Hodgkin, W. G. Richards. Utilization of Gaussian functions for the rapid evaluation of molecular similarity. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 188–191.
111. A. C. Good, S. S. So, W. G. Richards. Structure-activity relationships from molecular quantum similarity matrices. *J. Med. Chem.* **1993**, *36*, 433–438.
112. A. Seri-leby, R. Salter, S. West, W. G. Richards. Shape similarity as a single independent variable in QSAR. *Eur. J. Med. Chem.* **1994**, *29*, 687–695.
113. W. G. Richards. Molecular similarity and dissimilarity. Publicat en el llibre: Modeling of biomolecular structures and mechanisms. A. Pullman (Ed.). Kluwer Academic Publishers, Netherlands, 1995.
114. J. D. Petke. Cumulative and discrete similarity analysis of electrostatic potentials and fields. *J. Comput. Chem.* **1993**, *14*, 928–933.
115. F. Manaut, F. Sanz, J. Jose, M. Milesi. Automatic search for maximum similarity between molecular electrostatic potential distributions. *J. Comput.-Aided Mol. Design* **1991**, *5*, 371–380.
116. E. Lozoya, M. Berges, J. Rodríguez, F. Sanz, M. I. Loza, V. M. Moldes, C. F. Massaguer. Comparison of electrostatic similarity approaches applied to a series of ketanserin analogues with 5-HT<sub>2A</sub> antagonistic activity. *Quant. Struct.-Act. Relat.* **1998**, *17*, 199–204.
117. M. De Cáceres, J. Villà, J. J. Lozano, F. Sanz. MIPSIM: similarity analysis of molecular interaction potentials. *Comp. Appl. Biosci.* **2000**, *16*, 568–569.
118. R. Carbó-Dorca, P. G. Mezey (Eds.). Advances in molecular similarity. JAI Press, London, volum 1, 1996. Volum 2, 1998.
119. R. Carbó-Dorca, X. Gironés, P. G. Mezey (Eds.). Fundamentals of molecular similarity. Kluwer Academic/Plenum Press, New York, 2001.
120. P. A. M. Dirac. The principles of quantum mechanics. Clarendon Press, Oxford, 1930.
121. M. Born. Atomic physics. Blackie and Son, London, 1945.
122. J. Von Neumann. Mathematical foundations of quantum mechanics. Princeton University Press, Princeton, NJ, 1955.
123. R. McWeeny. Methods of molecular quantum mechanics. Academic Press, London, 1978.
124. C. C. J. Roothaan. New developments in molecular orbital theory. *Rev. Mod. Phys.* **1951**, *23*, 69–89.
125. P. G. Mezey. The holographic electron density theorem and quantum similarity measures. *Mol. Phys.* **1999**, *96*, 169–178.
126. P. G. Mezey. Functional groups in quantum chemistry. *Advances in Quantum Chemistry* **1996**, *27*, 163–222.
127. J. T. Tou, R. C. González. Pattern recognition principles. Addison-Wesley, Reading, M. A., 1974.



128. M. Born, J. R. Oppenheimer. On the quantum theory of molecules. *Annln. Phys.* **1927**, 85, 457–484.

129. M. Born, K. Huang. Dynamical theory of crystal lattices. Clarendon Press, Oxford, 1954.

### 3. Funcions densitat ASA

---

Estudis teòrics precisos de sistemes moleculars a nivell *ab initio* estan normalment limitats pel nombre i tipus d'àtoms involucrats. Sovint la química computacional es veu obligada a utilitzar esquemes senzills que reproduueixin el més acuradament possible els models *ab initio* per així reduir-ne el temps de computació. Un exemple són els estudis de *MQSM*, on el pas limitant quan es treballa amb l'aproximació *LCAO* és el càlcul d'integrals de quatre centres com les definides en l'equació (2.10). Malgrat que la capacitat de càlcul de les computadores està en constant progressió, l'estudi sistemàtic de sistemes moleculars grans, capaços de simular els sistemes biològics reals, fa necessari el desenvolupament d'aproximacions suficientment precises que puguin reemplaçar les mesures *ab initio*. A més, un dels principals àmbits d'aplicació de les *MQSM*, i del qual se'n fa especial èmfasi en el present treball, és l'anàlisi *QSAR/QSPR*. Aquí s'hi barreja no només la dimensió dels sistemes estudiats, sinó també la necessitat d'optimar la posició relativa de les molècules comparades. Quan s'utilitza el criteri de màxima semblança, el procés d'optimització requereix repetides avaluacions de les integrals de semblança, fent del tot inviable actualment realitzar els càlculs a nivell *ab initio*. Tots aquests requeriments computacionals han motivat el desenvolupament del càlcul de funcions densitat simplificades, d'integració ràpida, i que donen mesures de semblança de suficient qualitat quan es comparen amb els valors *ab initio*.

Abans d'analitzar els diferents mètodes d'ajust de les *DF*, i especialment l'aproximació *ASA*, s'ha cregut oportú descriure els trets fonamentals del mètode de Hartree-Fock (HF) i l'aproximació *LCAO*.<sup>1</sup> Sempre que en algun exemple d'aplicació es parli de *DF ab initio* o exacta, es refereix a densitats electròniques obtingudes mitjançant el mètode HF. A més, en l'últim apartat d'aquest capítol es mostra un possible ús de les funcions *ASA* en la determinació de matrius inicials per al càlcul iteratiu del camp auto coherent.

### 3.1 Mètode de Hartree-Fock

Un dels mètodes més emprats en la resolució de problemes polieletrònics és l'aproximació de HF, basada en el principi variacional de l'energia. Les equacions de HF determinen els orbitals més adequats per construir una funció d'ona polieletrònica d'un sol determinant i amb electrons aparellats:

$$|\Psi\rangle = |\mathbf{f}_1\bar{\mathbf{f}}_1\mathbf{f}_2\bar{\mathbf{f}}_2 \dots \mathbf{f}_n\bar{\mathbf{f}}_n\rangle, \quad (3.1)$$

on  $\mathbf{f}_i = \mathbf{j}_i(\mathbf{r})\mathbf{s}_i(s)$  són orbitals de spin. Segons el principi variacional, quan més propera sigui la funció  $\Psi$  de la veritable funció d'ona del sistema, més petit serà el valor de l'energia. Consegüentment els millors spin orbitals són els que minimitzen l'energia electrònica.

En sistemes amb nombre d'electrons  $N$  parell, i amb els orbitals de spin restringits ha tenir la mateixa funció espacial per les funcions de spin  $\mathbf{a}$  i  $\mathbf{b}$ , existeixen  $N/2$  orbitals espacials doblement ocupats. A més, si  $\mathbf{f}_i$  són ortonormals l'energia aproximada es pot escriure en funció d'integrals sobre coordenades únicament espacials. Així, per un sistema amb electrons aparellats es defineix:

$$E = 2 \sum_i H_{ii} + \sum_i \sum_j (2J_{ij} - K_{ij}), \quad (3.2)$$

on

$$H_{ii} = \int \mathbf{j}_i(1) \left[ -\frac{\nabla_1^2}{2} - \sum_c \frac{Z_c}{\mathbf{r}_{1c}} \right] \mathbf{j}_i(1) d\mathbf{V}_1 \quad (3.3)$$

$$J_{ij} = \int \int \mathbf{j}_i^*(1) \mathbf{j}_i(1) \frac{1}{\mathbf{r}_{12}} \mathbf{j}_j^*(2) \mathbf{j}_j(2) d\mathbf{V}_1 d\mathbf{V}_2 \quad (3.4)$$

$$K_{ij} = \int \int \mathbf{j}_i^*(1) \mathbf{j}_j(1) \frac{1}{\mathbf{r}_{12}} \mathbf{j}_j^*(2) \mathbf{j}_i(2) d\mathbf{V}_1 d\mathbf{V}_2 \quad (3.5)$$

El valor de l'energia E depèn dels orbitals  $\mathbf{j}_i$  amb els quals es construeix la funció polieletrònica (3.1). És necessari tenir en compte que, perquè l'expressió emprada per l'energia sigui vàlida, els orbitals  $\mathbf{j}_i$  han de ser ortonormals:

$$S_{ij} = \int \mathbf{j}_i^* \mathbf{j}_j d\mathbf{V} = \mathbf{d}_{ij} \quad (3.6)$$

Això obliga a aplicar un mètode anàleg al dels multiplicadors de Lagrange per trobar els punts extrems del funcional:

$$G[\mathbf{j}_i] = E[\mathbf{j}_i] + \sum_i \sum_j \mathbf{l}_{ij} (S_{ij} - \mathbf{d}_{ij}) \quad (3.7)$$

Les equacions que resulten són molt complexes i de difícil solució. Però degut a que els determinants són invariants respecte a les transformacions unitàries, s'escull el conjunt de funcions  $\mathbf{j}_i$  que fa mínim  $G[\mathbf{j}_i]$  i a més que els multiplicadors  $\mathbf{l}_{ij}$  valguin zero si  $i \neq j$ . Llavors, la igualtat que ha de complir qualsevol de les funcions  $\mathbf{j}_i$  perquè l'equació (3.7) sigui un extrem és:

$$\left\{ \hat{h}(1) + \sum_{j \neq i} [2\hat{J}_j(1) - \hat{K}_j(1)] \right\} \mathbf{j}_i(1) = \varepsilon_i \mathbf{j}_i(1), \quad (3.8)$$

que és iguala per a tots els electrons. En l'expressió (3.8)  $\varepsilon_i$  és l'energia de l'orbital  $i$ ,

$$\hat{h}(1) = -\frac{\nabla_1^2}{2} - \sum_c \frac{Z_c}{\mathbf{r}_{1c}} \quad (3.9)$$

$$\hat{J}_j(1) = \int \mathbf{j}_j^*(2) \frac{1}{\mathbf{r}_{12}} \mathbf{j}_j(2) d\mathbf{V}_2 \quad (3.10)$$

i es defineix l'operador  $\hat{K}$  per la propietat:

$$\hat{K}_j(1) \mathbf{j}_i(1) = \mathbf{j}_j(1) \int \mathbf{j}_i^*(2) \frac{1}{\mathbf{r}_{12}} \mathbf{j}_j(2) d\mathbf{V}_2 \quad (3.11)$$

L'equació de Hartree-Fock integrodiferencial (3.8) es pot abreviar:

$$\hat{F}(1)\mathbf{j}_i(1) = \varepsilon_i \mathbf{j}_i(1) \quad (3.12)$$

on  $\hat{F}$  és l'operador de Fock.

Les equacions de HF no són lineals perquè l'operador de Fock depèn de les funcions  $\mathbf{j}_i$ . En atenció a això, la solució de l'equació de valors propis plantejada en la igualtat (3.12) requereix d'un mètode iteratiu. El més utilitzat és el mètode anomenat de camp autocoherent (*Self-Consistent-Field, SCF*), donant com a resultat un conjunt d'orbitals ortonormalitzats amb energies orbitalàries  $\{\varepsilon_i\}$ . Per determinar els orbitals ocupats se seleccionen les  $N$  solucions que tenen menor  $\varepsilon_i$ . El determinant de Slater format amb aquests orbitals és la funció d'ona de l'estat fonamental corresponent a l'aproximació de HF. És la millor aproximació variacional de l'estat fonamental de la molècula utilitzant un únic determinant.

### 3.1.1 Aproximació LCAO i equacions de Hartree-Fock-Roothaan

A la pràctica, amb el mètode HF només es poden resoldre problemes atòmics degut a la complicació de l'equació de valors propis (3.12). Els problemes moleculars no es van poder resoldre fins que Roothaan va combinar les idees de Hartree i Fock amb la hipòtesi que els orbitals òptims  $\mathbf{j}_i$  es poden expressar com a combinacions lineals de les funcions de base adequades.<sup>1</sup> Així, qualsevol orbital  $\mathbf{j}_i$  s'expressa mitjançant una combinació lineal de les funcions de base  $\mathbf{c}_m$  segons:

$$\mathbf{j}_i(\mathbf{r}) = \sum_{m=1}^k C_{mi} \mathbf{c}_m(\mathbf{r}) \quad (3.13)$$

on  $C_{mi}$  són els coeficients dels orbitals moleculars obtinguts en el procés d'optimització de l'energia. Normalment les funcions de base emprades en càlculs de funcions d'ona moleculars són els orbitals atòmics dels àtoms de la molècula estudiada, més o menys modificats per tenir en compte que no es tracta d'àtoms aïllats. Aleshores cada orbital

molecular s'expressa com una combinació lineal d'orbitals atòmics, d'aquí les sigles *LCAO* (*Linear Combination of Atomic Orbitals*).

L'energia pot escriure's en funció dels coeficients  $C_{mi}$ :

$$E = 2 \sum_i \sum_m \sum_n C_m^* C_n H_{mm}^c + \sum_i \sum_j \sum_m \sum_n \sum_l \sum_s C_m^* C_n C_{lj}^* C_{sj} [2(\mathbf{m}|\mathbf{l}\mathbf{s}) - (\mathbf{ns}|\mathbf{l}\mathbf{n})] \quad (3.14)$$

on es defineix:

$$H_{mm}^c = \int \mathbf{c}_m^*(1) \hat{h}(1) \mathbf{c}_n(1) d\mathbf{V}_1 \quad (3.15)$$

$$(\mathbf{m}|\mathbf{l}\mathbf{s}) = \iint \mathbf{c}_m^*(1) \mathbf{c}_n(1) \frac{1}{r_{12}} \mathbf{c}_l^*(2) \mathbf{c}_s(2) d\mathbf{V}_1 d\mathbf{V}_2 \quad (3.16)$$

Les condicions d'ortonormalitat equivalen a l'equació:

$$\sum_m \sum_n (C_m^* C_n S_{mn} - \mathbf{d}_{ij}) = 0 \quad \forall i, j \quad (3.17)$$

on:

$$S_{mn} = \int \mathbf{c}_m^* \mathbf{c}_n d\mathbf{V} \quad (3.18)$$

El requeriment (3.17) s'inclou en la definició de l'energia (3.14) mitjançant el mètode de multiplicadors de Lagrange:

$$G(C_{mi}) = E(C_{mi}) + \sum_i \sum_j \mathbf{I}_{ij} \left( \sum_m \sum_n C_m^* C_n S_{mn} - \mathbf{d}_{ij} \right) \quad (3.19)$$

Segons se seleccionin les  $\mathbf{I}_{ij}$  s'obtindran diferents sèries d'orbitals moleculars.

Escollint les solucions que simplifiquen el problema, és a dir  $\mathbf{I}_{ij} = -2\varepsilon_i \mathbf{d}_{ij}$ , s'obté:

$$G(C_{mi}) = 2 \sum_i \left[ \sum_m \sum_n C_m^* C_n (H_{mm}^c - \varepsilon_i S_{mn}) - 1 \right] + \sum_i \sum_j \sum_m \sum_n \sum_l \sum_s C_m^* C_n C_{lj}^* C_{sj} [2(\mathbf{m}|\mathbf{l}\mathbf{s}) - (\mathbf{ns}|\mathbf{l}\mathbf{n})] \quad (3.20)$$

La condició de mínima energia s'aconsegueix derivant (3.20) respecte als coeficients  $C_{ni}$  i igualant a zero:

$$\sum_{m=1}^k (F_{mm} - \varepsilon_i S_{mm}) C_{mi} = 0 \quad n=1,2,\dots,k \quad (3.21)$$

on

$$F_{mm} = H_{mm}^c + \sum_j \sum_I \sum_s C_{Ij}^* C_{sj} [2(\mathbf{m}\mathbf{m}|\mathbf{l}\mathbf{s}) - (\mathbf{m}\mathbf{s}|\mathbf{l}\mathbf{n})] \quad (3.22)$$

Les igualtats (3.21) també es poden expressar com:

$$\sum_m F_{mm} C_{mi} = \sum_m \varepsilon_i S_{mm} C_{mi} , \quad (3.23)$$

que s'anomenen equacions de Hartree-Fock-Roothaan. S'han de resoldre de manera iterativa doncs els propis elements  $F_{mm}$  contenen les incògnites  $C_{mi}$ .

### 3.2 Funció densitat *ab initio*

La densitat electrònica d'una molècula de capa tancada descrita mitjançant una funció d'ona d'un sol determinant i amb tots els orbitals moleculars  $\mathbf{j}_i$  ocupats amb dos electrons es pot expressar com:

$$\mathbf{r}(\mathbf{r}) = 2 \sum_i^{N/2} |\mathbf{j}_i(\mathbf{r})|^2 = 2 \sum_i^{N/2} \mathbf{j}_i^*(\mathbf{r}) \mathbf{j}_i(\mathbf{r}) . \quad (3.24)$$

També es pot escriure en funció dels orbitals atòmics de base de la següent manera

$$\mathbf{r}(\mathbf{r}) = \sum_{mn} D_{mn} \mathbf{c}_m^*(\mathbf{r}) \mathbf{c}_n(\mathbf{r}) , \quad (3.25)$$

on els elements de la matriu densitat,  $D_{mn}$ , es determinen a partir dels coeficients moleculars segons

$$D_{mm} = 2 \sum_i^{N/2} C_{mi}^* C_{ni} \quad (3.26)$$

A més, essent  $\mathbf{r}(\mathbf{r}) d\mathbf{r}$  la probabilitat de trobar un electró en l'element de volum  $d\mathbf{r}$  en el punt  $\mathbf{r}$ , s'ha de complir que la integral d'aquesta densitat de càrrega doni el nombre total d'electrons,

$$\int \mathbf{r}(\mathbf{r}) d\mathbf{r} = 2 \sum_i^{N/2} \int |\mathbf{j}_i(\mathbf{r})|^2 d\mathbf{r} = 2 \sum_i^{N/2} 1 = N. \quad (3.27)$$

### 3.3 Funció densitat aproximada

Les funcions densitat simplificades se solen expressar com una combinació lineal de funcions 1s centrades en els àtoms:

$$\mathbf{r}(\mathbf{r}) = \sum_i c_i s_i(\mathbf{r} - \mathbf{r}_a). \quad (3.28)$$

És força corrent utilitzar la titlla sobre la lletra grega  $\mathbf{r}$  per diferenciar la  $DF$  aproximada de l'exacta. Els coeficients de l'expansió,  $\{c_i\}$ , es calculen minimitzant la funció error quadràtic integral entre la  $DF$  *ab initio* i la  $DF$  aproximada, que es pot definir com:<sup>2-4</sup>

$$\mathbf{e}^{(2)} = \int |\mathbf{r}(\mathbf{r}) - \tilde{\mathbf{r}}(\mathbf{r})|^2 d\mathbf{r}. \quad (3.29)$$

Una altra possibilitat és determinar els coeficients  $\{c_i\}$  de manera que minimitzin l'error quadràtic produït en l'energia de repulsió de Coulomb:<sup>5-8</sup>

$$\mathbf{e}^{(2)} = \iint \frac{[\mathbf{r}(\mathbf{r}_1) - \tilde{\mathbf{r}}(\mathbf{r}_1)][\mathbf{r}(\mathbf{r}_2) - \tilde{\mathbf{r}}(\mathbf{r}_2)]}{|\mathbf{r}_1 - \mathbf{r}_2|} d\mathbf{r}_1 d\mathbf{r}_2. \quad (3.30)$$

En ambdós casos el procés d'optimització està subjecte a la restricció que la integral de  $\mathbf{r}(\mathbf{r})$  en tot l'espai doni el nombre total d'electrons,

$$\int \tilde{\mathbf{r}}(\mathbf{r}) d\mathbf{r} = N. \quad (3.31)$$



### 3.4 Mètode d'ajust de mínims quadrats

En la literatura es poden trobar diferents algorismes d'ajust de les funcions densitat, encara que la tècnica més usada és la de mínims quadrats i la manera més comuna d'incloure la condició de normalització (3.31) en l'optimització dels coeficients és mitjançant un multiplicador de Lagrange. Una de les primeres aplicacions de les funcions densitat ajustades amb el mètode mínims quadrats va ser el càlcul d'energies electròniques.<sup>2-8</sup> També en l'àmbit de les *MQSM* s'ha utilitzat una tècnica de mínims quadrats per generar *DF* simplificades.<sup>9,10</sup> L'objectiu en ambdós casos és produir *DF* aproximades que reproduïxin els resultats *ab initio* amb la màxima precisió i el mínim cost computacional.

La funció error quadràtic integral definida en l'equació (3.29) es pot expressar en notació matricial d'acord amb:

$$\mathbf{e}^{(2)} = \mathbf{Z} + \mathbf{c}^T \mathbf{S} \mathbf{c} - 2\mathbf{c}^T \mathbf{t} + \mathbf{I}(N - \mathbf{c}^T \mathbf{n}) \quad (3.32)$$

on s'ha inclòs un multiplicador de Lagrange per assegurar el compliment de la restricció (3.31). En l'equació (3.32)  $\mathbf{Z}$  és la *QS-SM* de solapament de la *DF ab initio*:

$$\mathbf{Z} = \int |\mathbf{r}(\mathbf{r})|^2 d\mathbf{r} = \sum_{\mathbf{m}} D_{\mathbf{m}} \sum_{1s} D_{1s} \int \mathbf{c}_{\mathbf{m}}^*(\mathbf{r}) \mathbf{c}_{\mathbf{n}}(\mathbf{r}) \mathbf{c}_{1}^*(\mathbf{r}) \mathbf{c}_{s}(\mathbf{r}) d\mathbf{r}, \quad (3.33)$$

$\mathbf{c}$  és un vector columna que conté els coeficients de l'expansió,  $\mathbf{S}$  és la matriu de solapament, els elements de la qual es defineixen com:

$$S_{ij} = \int s_i(\mathbf{r} - \mathbf{r}_a) s_j(\mathbf{r} - \mathbf{r}_b) d\mathbf{r}, \quad (3.34)$$

els elements del vector  $\mathbf{t}$  són la integral de recobriment entre la *DF ab initio* i les funcions que es volen ajustar,

$$t_i = \int s_i(\mathbf{r} - \mathbf{r}_a) \mathbf{r}(\mathbf{r}) d\mathbf{r}, \quad (3.35)$$

i finalment, els elements de  $\mathbf{n}$  es calculen segons:

$$n_i = \int s_i(\mathbf{r} - \mathbf{r}_a) d\mathbf{r} \quad (3.36)$$

Derivant l'expressió (3.32) respecte als coeficients de l'expansió i aplicant la condició de mínim s'obté l'equació lineal:

$$\mathbf{S}\mathbf{c} = \mathbf{t}' \tag{3.37}$$

on el vector  $\mathbf{t}'$  és la suma

$$\mathbf{t}' = \mathbf{t} + \mathbf{I}\mathbf{n}, \tag{3.38}$$

i el multiplicador de Lagrange es calcula a partir del quocient:

$$\mathbf{I} = \frac{N - \mathbf{n}^\top \mathbf{S}^{-1} \mathbf{t}}{\mathbf{n}^\top \mathbf{S}^{-1} \mathbf{n}} \tag{3.39}$$

Així els coeficients òptims que minimitzen al funció  $\mathbf{e}^{(2)}$  i que compleixen la restricció (3.31) venen donats per la igualtat

$$\mathbf{c} = \mathbf{S}^{-1}(\mathbf{t} + \mathbf{I}\mathbf{n}) \tag{3.40}$$

### 3.5 Funcions densitat ASA

Recentment s'ha desenvolupat un nou model teòric d'ajust de la *DF* anomenat aproximació de capes atòmiques (*Atomic Shell Approximation, ASA*).<sup>11-19</sup> La innovació que suposen les funcions *ASA* és la restricció dels coeficients de l'expansió a tenir valors positius, de manera que la *DF* resultant mantingui les propietats d'una distribució de probabilitats. Si les funcions utilitzades en l'expansió lineal estan normalitzades,

$$\int s_i(\mathbf{r} - \mathbf{r}_a) d\mathbf{r} = 1 \quad \forall i, \tag{3.41}$$

les condicions imposades als coeficients en el procés d'optimització són:

$$\sum_i c_i = N \quad \text{i} \quad c_i \geq 0 \quad \forall i \tag{3.42}$$

### 3.5.1 Mètode de mínims quadrats adaptat a les funcions ASA

El primer intent d'ajust de ASA DF va ser una variant del mètode de mínims quadrats descrit en l'apartat 3.4, amb l'afegit que el vector dels coeficients es restringeix a tenir únicament valors positius.<sup>11,12</sup> El mètode proposat assigna inicialment al sistema considerat una base pràcticament saturada de funcions ASA, amb exponents generats a partir d'una seqüència *even-tempered*.<sup>20,21</sup> Llavors l'algorisme selecciona les capes o funcions amb exponents òptims i coeficients definits positius.

Degut a que la funció error quadràtic integral té una forma quadràtica, el seu mínim  $\mathbf{c}'_0$  es pot expressar en termes d'un vector arbitrari  $\mathbf{c}$  mitjançant l'equació:

$$\mathbf{c}'_0 = \mathbf{c} - \mathbf{S}^{-1} \nabla \mathbf{e}^{(2)}(\mathbf{c}), \quad (3.43)$$

on el gradient de  $\mathbf{c}$  es calcula d'acord amb

$$\nabla \mathbf{e}^{(2)}(\mathbf{c}) = 2(\mathbf{S}\mathbf{c} - \mathbf{t}'). \quad (3.44)$$

Escollint el punt arbitrari  $\mathbf{c}$  amb tots els components positius, i prenent la direcció  $\mathbf{p}$ ,

$$\mathbf{p} = \mathbf{S}^{-1} \nabla \mathbf{e}^{(2)}(\mathbf{c}), \quad (3.45)$$

com el vector d'aproximació més proper entre el punt  $\mathbf{c}$  i el mínim  $\mathbf{c}'_0$ , és possible definir un nou punt  $\mathbf{c}'_1$  conforme a:

$$\mathbf{c}'_1 = \mathbf{c} - \xi \mathbf{p}. \quad (3.46)$$

El paràmetre  $\xi \in [0,1]$  és l'increment més gran que es pot donar en sentit descendent cap al mínim considerant els coeficients positius. Un cop determinat  $\xi$ ,<sup>11</sup> les funcions amb coeficient zero o amb un pendent positiu en el punt  $\mathbf{c}'_1$  són rebutjades. És així perquè són les funcions que tindrien coeficients negatius en el desplaçament diferencial en la direcció al mínim partint de  $\mathbf{c}'_1$ . Després es calcula un nou vector d'aproximació,

$$\mathbf{p}_r = \mathbf{S}_r^{-1} \nabla_r \mathbf{e}^{(2)}(\mathbf{c}'_{1,r}). \quad (3.47)$$

El subíndex  $r$  en l'expressió (3.47) indica que la dimensió del problema s'ha reduït. A partir d'aquest punt se segueix un procés iteratiu fins a trobar el mínim restringit que compleix l'equació:

$$\mathbf{c}'_{0,r} = \mathbf{S}_r^{-1} \mathbf{t}'_r. \quad (3.48)$$

### 3.6 Funcions densitat PASA

La principal limitació de qualsevol mètode d'ajust de la densitat electrònica és la necessitat de calcular prèviament la *DF ab initio*. En àtoms no hi ha problema, però quan s'estudien sistemes moleculars amb un nombre gran de partícules hom està condicionat per la capacitat de càlcul de les computadores. Si a les limitacions esmentades s'afegeix que en els estudis de *MQSM* aplicats a anàlisis *QSAR* normalment hi intervenen conjunts moleculars extensos, i és necessari l'optimització de la posició relativa de tots els parells de compostos involucrats, es conclou que és inviable dur a terme càlculs a nivell *ab initio*. Aleshores, per agilitar els càlculs de les *MQSM*, s'ha desenvolupat una aproximació promolecular que genera la *DF* d'una molècula com una suma *DF* atòmiques independents,<sup>22</sup> que es coneix amb les sigles *PASA* (*Promolecular Atomic Shell Approximation*).<sup>12-18</sup> La construcció de la densitat electrònica promolecular només requereix de les coordenades dels àtoms i d'una base ajustada de funcions atòmiques. Malgrat les limitacions de qualsevol aproximació promolecular, especialment en la descripció dels enllaços, la utilització de *PASA DF* en càlculs de *MQSM* és admissible, com ho demostren els exemples inclosos en els articles adjunts en aquest capítol, així com les anàlisis *QSAR/QSPR* presentades en els capítols 6 i 7.

Mitjançant l'aproximació *PASA*, la *DF* d'una molècula  $A$  es descriu per mitjà de simples sumes de *DF* dels àtoms  $\{a\}$  que la formen:

$$\mathbf{r}_A^{PASA}(\mathbf{r}) = \sum_a P_a \mathbf{r}_a^{ASA}(\mathbf{r}). \quad (3.49)$$

$P_a$  és la càrrega total sobre l'àtom  $a$ , normalment definida com el nombre atòmic.

Mentre que les densitats atòmiques  $\mathbf{r}_a^{ASA}(\mathbf{r})$ , es construeixen a través de l'expressió

$$\mathbf{r}_a^{ASA}(\mathbf{r}) = \sum_{i \in a} w_i |s_i(\mathbf{r} - \mathbf{r}_a)|^2, \quad (3.50)$$

on prenent un conjunt de funcions Gaussians 1s normalitzades:

$$|s_i(\mathbf{r} - \mathbf{r}_a)|^2 = \left( \frac{2z_i}{\pi} \right)^{3/2} \exp(-2z_i |\mathbf{r} - \mathbf{r}_a|^2) \quad (3.51)$$

de manera que

$$\int |s_i(\mathbf{r} - \mathbf{r}_a)|^2 d\mathbf{r} = 1 \quad \forall i, \quad (3.52)$$

llavors els coeficients ASA han de complir les condicions de convexitat:

$$\left\{ w_i \in \mathbf{R}^+ \quad \forall i \quad \wedge \quad \sum_{i \in a} w_i = 1 \right\}. \quad (3.53)$$

En l'equació (3.50) s'utilitzen les funcions  $s_i$  al quadrat per analogia amb la densitat electrònica *ab initio*, definida com el producte de la funció d'ona al quadrat. A la pràctica equival a multiplicar l'exponent de la funció  $s_i$  per dos.

Substituint  $\mathbf{r}_a^{ASA}(\mathbf{r})$  en l'equació (3.49) per l'expressió (3.50) i fent la integral en tot l'espai de la *PASA DF*,

$$\int \mathbf{r}_A^{PASA}(\mathbf{r}) d\mathbf{r} = \sum_{a \in A} P_a \sum_{i \in a} w_i \int |s_i(\mathbf{r} - \mathbf{r}_a)|^2 d\mathbf{r} = \sum_{a \in A} P_a \sum_{i \in a} w_i = \sum_{a \in A} P_a = N, \quad (3.54)$$

s'obté el nombre d'electrons total de la molècula, i per tant es compleix la condició de normalització (3.31). En els ajustos atòmics que es descriuran més endavant, s'ha normalitzat les funcions ASA i *ab initio* a u, tal com indica la segona restricció de l'equació (3.53).

### 3.7 Funció error quadràtic integral

Degut a les variacions introduïdes en la definició de la funció ASA respecte a la fórmula general (3.28), a partir d'ara s'adoptaran uns canvis de nomenclatura. Així, la funció error quadràtic integral, en la seva notació matricial, es defineix com:

$$\mathbf{e}^{(2)} = \mathbf{Z} + \mathbf{w}^\top \mathbf{Z} \mathbf{w} - 2\mathbf{b}^\top \mathbf{w} \quad (3.55)$$

A diferència de l'equació (3.32), no s'ha inclòs el multiplicador de Lagrange. A més, s'utilitza  $\mathbf{w}$  en lloc de  $\mathbf{c}$  per descriure el vector de coeficients, i es defineix la matriu  $\mathbf{Z}$ , els elements de la qual són

$$Z_{ij} = \int |s_i(\mathbf{r})|^2 |s_j(\mathbf{r})|^2 d\mathbf{r}, \quad (3.56)$$

que corresponen a la definició de la MQSM de tipus solapament donada en l'equació (2.13) entre dues capes  $i \in a$  i  $j \in b$ :

$$Z_{ij} = \left( \frac{2\mathbf{z}_i \mathbf{z}_j}{\pi(\mathbf{z}_i + \mathbf{z}_j)} \right)^{3/2} \exp \left( -\frac{2\mathbf{z}_i \mathbf{z}_j}{\mathbf{z}_i + \mathbf{z}_j} R_{ab}^2 \right). \quad (3.57)$$

Si es realitza un ajust de la densitat electrònica d'un àtom, llavors la distància interatòmica  $R_{ab}$  és zero i l'element  $Z_{ij}$  se simplifica d'acord amb l'expressió

$$Z_{ij} = \left( \frac{2\mathbf{z}_i \mathbf{z}_j}{\pi(\mathbf{z}_i + \mathbf{z}_j)} \right)^{3/2}. \quad (3.58)$$

Un altre canvi de nomenclatura és la utilització del vector  $\mathbf{b}$  per descriure les integrals de recobriment entre la DF *ab initio* i les funcions ASA:

$$b_i = \int |s_i(\mathbf{r})|^2 \mathbf{r}(\mathbf{r}) d\mathbf{r}, \quad (3.59)$$

on substituint  $\mathbf{r}(\mathbf{r})$  per l'expressió (3.25) es defineix

$$b_i = \sum_{\mathbf{m}} D_{\mathbf{m}} \int |s_i(\mathbf{r})|^2 \mathbf{c}_{\mathbf{m}}^*(\mathbf{r}) \mathbf{c}_{\mathbf{n}}(\mathbf{r}) d\mathbf{r}. \quad (3.60)$$

### 3.8 Optimització dels coeficients ASA mitjançant rotacions de Jacobi

La tècnica de rotacions de Jacobi (*Elementary Jacobi Rotations, EJR*)<sup>23</sup> és un procediment que produeix variacions en un vector de dimensió  $n$  amb la característica que es conserva la seva norma. Inicialment es va desenvolupar com un mètode de diagonalització de matrius, però amb el temps s'ha aplicat en altres dominis. Així per exemple, en l'àmbit de la química quàntica, la tècnica *EJR* s'ha emprat en l'optimització directa de l'energia electrònica,<sup>24-31</sup> i en la localització d'orbitals moleculars.<sup>32,33</sup> També, més recentment, s'ha proposat la transformació *EJR* per solucionar el problema associat amb l'ajust restringit de la densitat electrònica utilitzant funcions ASA. Inicialment es va desenvolupar com un mètode d'ajust d'àtoms,<sup>14,15,17,18</sup> però posteriorment també s'ha aplicat en sistemes moleculars.<sup>19</sup>

L'algorisme dissenyat té dos trets bàsics que el diferencien de la resta. En primer lloc es defineix un nou vector de coeficients,  $\mathbf{x}$ , que serveix per generar els coeficients ASA mitjançant la igualtat:  $w_i = |x_i|^2 \quad \forall i$ . Conseqüentment, els elements del vector  $\mathbf{w}$  seran positius. A més, s'elegeix com a punt inicial de l'optimització un vector  $\mathbf{x}$  amb norma igual a 1:

$$\langle \mathbf{x} | \mathbf{x} \rangle = \mathbf{x}^+ \mathbf{x} = 1 \quad \rightarrow \quad \sum_i x_i^2 = \sum_i w_i = 1 \quad (3.61)$$

de manera que es compleixen les dues condicions de convexitat detallades en l'equació (3.53). L'error quadràtic integral s'escriu en funció dels nous coeficients  $\mathbf{x}$  d'acord amb l'expressió

$$\mathbf{e}^{(2)} = Z + \sum_{i,j} x_i^2 x_j^2 Z_{ij} - 2 \sum_i x_i^2 b_i \quad (3.62)$$

El segon aspecte fa referència a l'optimització del vector de coeficients mitjançant la tècnica de rotacions de Jacobi. L'aplicació de la transformació ortogonal *EJR*,  $\mathbf{J}_{pq}(\alpha)$ , sobre parells d'elements del vector  $\mathbf{x}$ ,  $\{x_p, x_q\}$ , es pot descriure mitjançant les relacions

$$\begin{aligned}\dot{x}_p &\leftarrow c x_p - s x_q \\ \dot{x}_q &\leftarrow s x_p + c x_q ,\end{aligned}\quad (3.63)$$

on únicament els elements  $p$  i  $q$  del vector  $\mathbf{x}$  són modificats. Els símbols  $c$  i  $s$  de l'equació (3.63) fan referència al cosinus i al sinus de l'angle  $\alpha$  de rotació de Jacobi.

$\mathbf{e}^{(2)}$  es pot expressar en funció dels coeficients  $\{x_p, x_q\}$ :

$$\begin{aligned}\mathbf{e}^{(2)} &= \mathbf{Z} + x_p^4 \mathbf{Z}_{pp} + x_q^4 \mathbf{Z}_{qq} + 2x_p^2 x_q^2 \mathbf{Z}_{pq} + 2x_p^2 \sum_{i \neq p,q} x_i^2 \mathbf{Z}_{pi} + 2x_q^2 \sum_{i \neq p,q} x_i^2 \mathbf{Z}_{iq} \\ &+ \sum_{i \neq p,q} \sum_{j \neq p,q} x_i^2 x_j^2 \mathbf{Z}_{ij} - 2x_p^2 b_p - 2x_q^2 b_q - 2 \sum_{i \neq p,q} x_i^2 b_i .\end{aligned}\quad (3.64)$$

Aplicant la rotació  $\mathbf{J}_{pq}(\alpha)$  sobre l'equació (3.64) s'obté la variació de  $\mathbf{e}^{(2)}$  respecte als coeficients actius  $\{x_p, x_q\}$ :

$$\begin{aligned}d\mathbf{e}^{(2)} &= d x_p^4 \mathbf{Z}_{pp} + d x_q^4 \mathbf{Z}_{qq} + 2d(x_p^2 x_q^2) \mathbf{Z}_{pq} + 2d x_p^2 \sum_{i \neq p,q} x_i^2 \mathbf{Z}_{pi} + 2d x_q^2 \sum_{i \neq p,q} x_i^2 \mathbf{Z}_{iq} \\ &- 2b_p d x_p^2 - 2b_q d x_q^2\end{aligned}\quad (3.65)$$

La deducció de les variacions dels coeficients degudes a la transformació (3.63),

$$\begin{aligned}d x_p^2 &= (\dot{x}_p^2 - x_p^2) , & d x_q^2 &= (\dot{x}_q^2 - x_q^2) , & d(x_p^2 x_q^2) &= (\dot{x}_p^2 \dot{x}_q^2 - x_p^2 x_q^2) \\ d x_p^4 &= (\dot{x}_p^4 - x_p^4) , & d x_q^4 &= (\dot{x}_q^4 - x_q^4) ,\end{aligned}\quad (3.66)$$

està àmpliament detallada en l'article 3.1 agregat en la pàgina 59 d'aquest capítol. En últim terme,  $d\mathbf{e}^{(2)}$  es pot expressar com una equació d'ordre quatre relativa al sinus:

$$d\mathbf{e}^{(2)} = E_{04}s^4 + E_{13}cs^3 + E_{02}s^2 + E_{11}cs .\quad (3.67)$$

Les fórmules dels paràmetres  $E_{04}$ ,  $E_{13}$ ,  $E_{02}$  i  $E_{11}$  també es troben especificades en l'article 3.1. Aplicant la condició de mínim en l'equació (3.67) s'obté

$$\frac{d d\mathbf{e}^{(2)}}{ds} = -c(T_1 t^2 - 2T_2 t - T_3) = 0 ,\quad (3.68)$$



on  $t = \frac{s}{c}$ , i

$$T_1 = (E_{13}s^2 + E_{11}); T_2 = (2E_{04}s^2 + E_{02}) \text{ i } T_3 = (3E_{13}s^2 + E_{11}) \quad (3.69)$$

L'angle òptim de rotació es troba solucionant l'equació de segon grau definida en l'expressió (3.68). Inicialment es va proposar un algorisme iteratiu per resoldre-la perquè els termes  $T_1$ ,  $T_2$  i  $T_3$  són funció del sinus al quadrat. Posteriorment s'han simplificat les fórmules expressant el sinus i el cosinus en sèries de Taylor relatives a l'angle  $\alpha$ , com es mostrarà en l'apartat 3.10.

### ***3.8.1 Esquema computacional de l'algorisme EJR***

Tot seguit es detallen les principals subrutines dissenyades per modelar els coeficients de les funcions *ASA* mitjançant l'algorisme *EJR* en forma de pseudo codi.

**Determinació dels coeficients *ASA* inicials.** S'escull en qualitat de vector inicial  $\mathbf{x}$  del procés *EJR* el vector propi de la matriu  $\mathbf{Z}$  que dona el valor més petit de la funció  $\mathbf{e}^{(2)}$ . Una propietat intrínseca dels vectors propis és precisament la condició (3.61). En la taula 3.1 es descriu la subrutina emprada en la selecció del vector  $\mathbf{x}$  inicial.

### Subrutina COEFICIENTS INICIALS

- ✓ Donat el valor de l'auto semblança *ab initio*,  $Z$
- ✓ Donat el nombre de funcions ASA,  $n$
- ✓ Donat el vector  $\mathbf{b}(n \times 1)$  i la matriu de mesures de semblança  $\mathbf{Z}(n \times n)$
- Calcula la descomposició espectral:  $\mathbf{V}^+ \mathbf{Z} \mathbf{V} = \mathbf{L}$ , on  $\mathbf{V}$  és la matriu dels vectors propis i  $\mathbf{L}$  una matriu diagonal amb els valors propis
- Inicialitza  $k=1$  i  $\mathbf{e}_{op}^{(2)} \approx \infty$
- Per tot  $k \in n$ 
  - Iguala  $\mathbf{x} = \mathbf{v}_k$ , on  $\mathbf{v}_k = \{V_{ik} \mid \forall i \in n\}$
  - Calcula els elements de  $\mathbf{w}$ :  $w_i = |x_i|^2 \forall i$
  - Calcula  $\mathbf{e}_k^{(2)} = \mathbf{Z} + \mathbf{w}^T \mathbf{Z} \mathbf{w} - 2\mathbf{b}^T \mathbf{w}$
  - Si  $\mathbf{e}_k^{(2)} < \mathbf{e}_{op}^{(2)} \rightarrow \mathbf{x}^* = \mathbf{x}$ ;  $\mathbf{w}^* = \mathbf{w}$  i  $\mathbf{e}_{op}^{(2)} = \mathbf{e}_k^{(2)}$
- Fi per tot  $k$
- ❖ Retorna els vectors  $\mathbf{x}^*$  i  $\mathbf{w}^*$ , i el valor de la funció  $\mathbf{e}_{op}^{(2)}$

**Taula 3.1** Algorisme per obtenir els coeficients ASA inicials

**Càlcul del sinus EJR òptim.** L'angle EJR òptim s'obté resolent l'equació de segon grau,

$$T_1 t^2 - 2T_2 t - T_3 = 0, \tag{3.70}$$

que apareix en l'expressió (3.68). S'ha de comprovar quina de les dues possibles solucions:

$$t_{\pm} = \left( T_2 \pm \sqrt{T_2^2 + T_1 T_3} \right) T_1^{-1}, \tag{3.71}$$

dóna un valor de  $d \mathbf{e}^{(2)}$  més petit. A més, s'ha de resoldre de forma iterativa perquè els termes  $T_1$ ,  $T_2$  i  $T_3$  definits en l'equació (3.69) depenen de  $s^2$ .

En la taula 3.2 es detalla l'algorisme iteratiu seguit en la determinació l'angle  $EJR$  òptim. Inicialment es fa una estimació de l'angle  $EJR$  mitjançant una tècnica de Fibonacci.<sup>34</sup> És un mètode indicat per a problemes d'una dimensió on no es coneix el gradient. Es busca el valor de la variable  $t$  que fa mínim la funció (3.70) en l'interval de l'angle  $\mathbf{a} \in \left(-\mathbf{P}/2, \mathbf{P}/2\right)$ . La subrutina FIBONACCI retorna el valor inicial  $s_0^2$ .

---

#### Subrutina SINUS

- ✓ Donat els valors  $E_{04}, E_{13}, E_{02}$  i  $E_{11}$
- Demana **subrutina FIBONACCI** Necessita:  $E_{04}, E_{13}, E_{02}$  i  $E_{11}$ . Retorna:  $s_0^2$
- Inicialitza  $k=1$
- Per tot  $k <$  màxim nombre de cicles SINUS
  - Calcula els termes  $T_1(s_{k-1}^2)$ ,  $T_2(s_{k-1}^2)$  i  $T_3(s_{k-1}^2)$  mitjançant l'equació (3.69)
  - Calcula els dos possibles valors de la tangent:  $t_+$  i  $t_-$  mitjançant l'equació (3.71)
  - Calcula  $s_+^2 = t_+^2 (1 + t_+^2)^{-1}$ ,  $s_+ = (\text{signe } t_+) \sqrt{s_+^2}$  i  $c_+ = \sqrt{1 - s_+^2}$
  - Calcula  $de_+^{(2)} = E_{04}s_+^4 + E_{13}c_+s_+^3 + E_{02}s_+^2 + E_{11}c_+s_+$
  - Calcula  $s_-^2 = t_-^2 (1 + t_-^2)^{-1}$ ,  $s_- = (\text{signe } t_-) \sqrt{s_-^2}$  i  $c_- = \sqrt{1 - s_-^2}$
  - Calcula  $de_-^{(2)} = E_{04}s_-^4 + E_{13}c_-s_-^3 + E_{02}s_-^2 + E_{11}c_-s_-$
  - Si  $de_+^{(2)} < de_-^{(2)}$  llavors
    - Defineix  $s_k = s_+$ ;  $c_k = c_+$  i  $s_k^2 = s_+s_+$
  - En cas contrari
    - Defineix  $s_k = s_-$ ;  $c_k = c_-$  i  $s_k^2 = s_-s_-$
  - Fi de la condició
  - Si  $s_k \notin [-1, 1] \rightarrow s_k = s_{k-1}$ ;  $c_k = c_{k-1}$ ; Fi per tot  $k$
  - Si  $|s_k^2 - s_{k-1}^2| < \text{tolerància} \rightarrow$  Fi per tot  $k$
- Fi per tot  $k$
- ❖ Retorna el valor òptim del sinus i del cosinus:  $s_k$  i  $c_k$

---

**Taula 3.2** Algorisme d'optimització del sinus  $EJR$

**Aplicació de la tècnica EJR.** En la taula 3.3 es descriu la seqüència d'instruccions que s'executa quan s'optimitza el vector de coeficients  $\mathbf{x}$ . Donat un vector inicial de coeficients que compleixi les condicions de convexitat detallades en l'equació (3.53), es transforma mitjançant successives rotacions  $\mathbf{J}_{pq}(\alpha)$ ,  $\forall p, q \in n$ , amb l'objectiu de minimitzar la funció error quadràtic integral. Les condicions de convexitat es conserven al llarg de tot el procés perquè les transformacions aplicades als coeficients són ortogonals, i per tant la norma del vector no varia. El resultat final és un vector amb els coeficients  $\mathbf{w}$  òptims.

---

**Subrutina EJR**

- ✓ Donat el valor de l'auto semblança *ab initio*,  $Z$ , i les integrals  $\mathbf{Z}$  i  $\mathbf{b}$
- ✓ Donat el nombre de funcions ASA:  $n$
- ✓ Donat el vector dels coeficients inicials:  $\mathbf{x}$ ,  $\mathbf{w}$
- ✓ Donat el valor inicial de la funció error quadràtic:  $\mathbf{e}_0^{(2)}$
- Inicialitza  $k=1$
- Per tot  $k < \text{màxim nombre de cicles EJR}$ 
  - Per tot  $p \in n$ 
    - Per tot  $q \in n \mid q > p$ 
      - Calcula els valors  $E_{04}$ ,  $E_{13}$ ,  $E_{02}$  i  $E_{11}$
      - Demana **subrutina SINUS**. Necessita:  $E_{04}$ ,  $E_{13}$ ,  $E_{02}$  i  $E_{11}$ . Retorna:  $s$  i  $c$
      - Calcula els nous coeficients:  $\left\{ \begin{array}{l} \dot{x}_p = c x_p - s x_q \\ \dot{x}_q = s x_p + c x_q \end{array} \right\} \dot{w}_p = \dot{x}_p^2$  i  $\dot{w}_q = \dot{x}_q^2$
      - Construeix el vector  $\dot{\mathbf{w}} = (w_1, w_2, \dots, \dot{w}_p, \dots, \dot{w}_q, \dots, w_n)^T$
      - Calcula  $\mathbf{e}^{(2)} = \mathbf{Z} + \dot{\mathbf{w}}^T \mathbf{Z} \dot{\mathbf{w}} - 2\mathbf{b}^T \dot{\mathbf{w}}$
      - Si  $\mathbf{e}^{(2)} < \mathbf{e}_k^{(2)} \rightarrow w_p = \dot{w}_p; w_q = \dot{w}_q; x_p = \dot{x}_p; x_q = \dot{x}_q; \mathbf{e}_k^{(2)} = \mathbf{e}^{(2)}$
    - Fi per tot  $q$
  - Fi per tot  $p$
  - Si  $|\mathbf{e}_k^{(2)} - \mathbf{e}_{k-1}^{(2)}| < \text{tolerància} \rightarrow$  Fi per tot  $k$
- Fi per tot  $k$
- ❖ Retorna el valor de  $\mathbf{x}$  i  $\mathbf{w}$  òptims, i el valor de la funció  $\mathbf{e}^{(2)}$

---

**Taula 3.3** Algorisme d'optimització dels coeficients ASA amb la tècnica EJR

### 3.9 Ajustos atòmics

Els primers exemples d'aplicació del mètode d'ajust de les funcions ASA emprant *EJR* han estat sistemes atòmics. Però abans de descriure el càlcul d'alguna de les bases atòmiques parametritzades, és convenient explicar com s'obtenen els exponents de les funcions ASA. Una possibilitat, comentada en l'apartat 3.5.1, és generar els exponents ASA a l'inici del procés mitjançant una seqüència *even-tempered* i no modificar-los durant l'ajust. Se sol començar amb una sèrie molt gran de funcions esfèriques, que gairebé saturen l'espai on estan definides, i l'algorisme d'ajust s'encarrega de triar-ne un subconjunt amb els exponents òptims. La filosofia seguida en l'algorisme que ara es proposa és diferent. La idea és partir d'un nombre molt reduït de funcions, i optimitzar-ne tant els exponents com els coeficients fins a assolir uns valors de l'error quadràtic similars als que s'obtidrien saturant l'espai de funcions ASA.

#### 3.9.1 Optimització dels exponents ASA

Els exponents de les funcions ASA es determinen mitjançant un mètode de Newton. Això ha estat possible perquè s'han resolt les primeres i segones derivades de la funció error quadràtic integral referides als exponents de les funcions ASA. En l'article 3.1 es detallen les expressions involucrades en les derivades dels elements del vector  $\mathbf{b}$  i de la matriu  $\mathbf{Z}$ . La deducció de les derivades s'ha fet per un cas general de funcions Gaussianes  $ns$ , malgrat que únicament s'han utilitzat les fórmules simplificades corresponents a funcions  $1s$ . Tot el desenvolupament que segueix a continuació és exclusiu dels àtoms.

**Mètode d'optimització de Newton.** S'ha de minimitzar la funció  $\mathbf{e}^{(2)}(\mathbf{F})$ , que depèn del vector d'exponents ASA,  $\mathbf{F} = (\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_n)^T$ , i de la qual es coneix el seu vector gradient,  $\tilde{\mathbf{N}}\mathbf{e}^{(2)} = \mathbf{g} = (\partial\mathbf{e}^{(2)}/\partial\mathbf{z}_1, \partial\mathbf{e}^{(2)}/\partial\mathbf{z}_2, \dots, \partial\mathbf{e}^{(2)}/\partial\mathbf{z}_n)^T$ , i la matriu Hessiana,  $\mathbf{H} = \{h_{ij} | h_{ij} = \partial^2\mathbf{e}^{(2)}/\partial\mathbf{z}_i\partial\mathbf{z}_j\}$ . El procés d'optimització és iteratiu i es fonamenta en la

propietat que en la regió propera a l'actual punt de la recerca  $\mathbf{F}_k$ , la funció  $\mathbf{e}^{(2)}(\mathbf{F})$  es pot escriure com:<sup>34</sup>

$$\mathbf{e}^{(2)}(\mathbf{F}) = \mathbf{e}^{(2)}(\mathbf{F}_k) + (\mathbf{F} - \mathbf{F}_k)^\top \mathbf{g}_k + \frac{1}{2}(\mathbf{F} - \mathbf{F}_k)^\top \mathbf{H}_k (\mathbf{F} - \mathbf{F}_k). \quad (3.72)$$

Els punts estacionaris d'aquesta funció quadràtica s'obtenen derivant-la respecte a les variables  $\mathbf{F}$  i igualant a zero,

$$\mathbf{H}_k (\mathbf{F} - \mathbf{F}_k) = -\mathbf{g}_k. \quad (3.73)$$

Suposant que existeix la inversa de la matriu  $\mathbf{H}_k$ , el següent punt en el procés d'optimització,  $\mathbf{F} = \mathbf{F}_{k+1}$ , es calcula com

$$\mathbf{F}_{k+1} = \mathbf{F}_k - \mathbf{H}_k^{-1} \mathbf{g}_k. \quad (3.74)$$

Abans d'aplicar l'increment i calcular el nou punt, s'ha de comprovar si és definit positiu o negatiu, per determinar si el punt  $\mathbf{F}_k$  és proper a un mínim o a un màxim local.

En la taula 3.4 s'esquematitza el mètode de Newton que s'ha programat. Donat un conjunt inicial de funcions ASA, amb exponents  $\mathbf{F}_0$  i coeficients  $\mathbf{w}$ , s'optima el vector  $\mathbf{F}$  de manera iterativa. Per a cada cicle  $k$  es calcula el vector del gradient,  $\mathbf{g}_{k-1}$ , i la matriu Hessiana,  $\mathbf{H}_{k-1}$ . Tot seguit es determina la inversa de l'Hessiana,  $\mathbf{H}_{k-1}^{-1}$ , i l'increment de la variable  $\mathbf{F}$  en la direcció de l'òptim local més proper:  $\mathbf{p} = -\mathbf{H}_{k-1}^{-1} \mathbf{g}_{k-1}$ . S'ha de fer un test per comprovar si es va en sentit positiu o negatiu al mínim, canviant, si s'escau, el signe del vector  $\mathbf{p}$ . L'algorisme segueix amb un bucle, on es calcula la funció error quadràtic del nou punt en la direcció del mínim,  $\mathbf{e}^{(2)}(\mathbf{F}_{k-1} + \sigma \mathbf{p})$ , essent  $\sigma$  un escalar amb valor inicial u. Si el valor calculat de la funció  $\mathbf{e}^{(2)}$  no és inferior al determinat en el cicle anterior  $k-1$ ,  $\mathbf{e}_{k-1}^{(2)}$ , aleshores es divideix  $\sigma$  per dos i es torna a calcular  $\mathbf{e}^{(2)}$ . El procés es repeteix fins a aconseguir un valor òptim de la funció  $\mathbf{e}^{(2)}$  o bé que la variable  $\sigma$  sigui menor a un límit preestablert. L'optimització s'atura quan la diferència de valors de la funció  $\mathbf{e}^{(2)}$  entre dues iteracions,  $k$  i  $k-1$ , és inapreciable. El vector  $\mathbf{w}$  amb els coeficients ASA, necessari per calcular la funció error quadràtic integral, es manté invariant durant l'execució de la subrutina NEWTON.

---

## Subrutina NEWTON

- ✓ Donat el valor de l'auto semblança *ab initio*,  $Z$ , i un conjunt de funcions de base  $\{\mathbf{c}_m\}$
- ✓ Donades  $n$  funcions ASA, amb coeficients  $\mathbf{w}$  i exponents  $\mathbf{F}_0$
- ✓ Donat el valor inicial de la funció error quadràtic:  $\mathbf{e}_0^{(2)}$ 
  - Inicialitza  $k=1$  i  $\mathbf{F}_{op} = \mathbf{F}_0$
  - Per tot  $k <$  màxim nombre de cicles NEWTON
    - Calcula el vector del gradient  $\mathbf{g}_{k-1}$  i la matriu de l'Hessiana  $\mathbf{H}_{k-1}$
    - Calcula la inversa de l'Hessiana  $\mathbf{H}_{k-1}^{-1}$
    - Calcula el vector increment  $\mathbf{p} = -\mathbf{H}_{k-1}^{-1} \mathbf{g}_{k-1}$
    - Calcula una estimació de la funció a optimitzar:  $\mathbf{e}_{apr.}^{(2)} = \mathbf{e}_{k-1}^{(2)} + \mathbf{p}^T \mathbf{g}_{k-1} + \frac{1}{2} \mathbf{p}^T \mathbf{H}_{k-1}^{-1} \mathbf{p}$
    - Si  $\mathbf{e}_{apr.}^{(2)} > \mathbf{e}_{k-1}^{(2)} \rightarrow \mathbf{p} = -\mathbf{p}$
    - Inicialitza  $l=1$  i  $\sigma=1.0$
    - Per tot  $l <$  màxim nombre de cicles SIGMA
      - Calcula  $\mathbf{F}_k = \mathbf{F}_{k-1} + \sigma \mathbf{p}$
      - Calcula els elements de  $\mathbf{Z}$  i  $\mathbf{b}$ : eq. (3.56) i (3.60). Per  $\mathbf{b}$  es necessita  $\{\mathbf{c}_m\}$
      - Calcula  $\mathbf{e}_k^{(2)} = Z + \mathbf{w}^T \mathbf{Z} \mathbf{w} - 2\mathbf{b}^T \mathbf{w}$
      - Si  $\mathbf{e}_k^{(2)} < \mathbf{e}_{k-1}^{(2)}$  llavors
        - Fi per tot  $l$
      - En cas contrari
        - $\sigma = \sigma/2$
        - Si  $\sigma < \sigma_{min} \rightarrow$  Fi per tot  $k$
        - Fi condicional
    - Fi per tot  $l$ 
      - $\mathbf{F}_{op} = \mathbf{F}_k$ , i guarda la matriu  $\mathbf{Z}$  i el vector  $\mathbf{b}$
      - Si  $\mathbf{e}_{k-1}^{(2)} - \mathbf{e}_k^{(2)} <$  tolerància  $\rightarrow$  Fi per tot  $k$
  - Fi per tot  $k$
  - ❖ Retorna el valor òptim de  $\mathbf{F}_{op}$  i de la funció  $\mathbf{e}^{(2)}$ , i les corresponents integrals  $\mathbf{Z}$  i  $\mathbf{b}$

---

**Taula 3.4** Algorisme d'optimització dels exponents ASA amb el mètode de Newton

### 3.9.2 Optimització conjunta de coeficients i exponents ASA

En els apartats anteriors, s'ha descrit el càlcul dels coeficients i exponents ASA per separat. La subrutina que engloba el procés global d'optimització de les funcions ASA es descriu en la taula 3.5. Es basa en un procés iteratiu on s'executa consecutivament la subrutina NEWTON i la subrutina EJR fins que la variació del valor de la funció  $e^{(2)}$  entre dos cicles és inapreciable. Els exponents inicials es calculen mitjançant una sèrie *even-tempered*,<sup>20,21</sup>

$$z_i = \alpha\beta^i \quad \forall i \in n, \tag{3.75}$$

mentre que el vector de coeficients inicials es determina amb la subrutina COEFICIENTS INICIALS descrita en la taula 3.1.

#### Subrutina OPTIMIZA FUNCIO ASA

- ✓ Donat el valor de l'autosemblança *ab initio*,  $Z$ , i un conjunt de funcions de base  $\{\mathbf{c}_m\}$
- ✓ Donat el nombre de funcions ASA,  $n$ ,
- ✓ Donat els paràmetres *even-tempered*  $\{\alpha, \beta\}$
- Calcula el vector d'exponents ASA inicial  $\mathbf{F} = \{z_i \mid z_i = \alpha\beta^i \forall i \in n\}$
- Calcula els elements de  $\mathbf{Z}$  i  $\mathbf{b}$ : eq. (3.56) i (3.60). Per  $\mathbf{b}$  es necessita  $\{\mathbf{c}_m\}$
- Demana **subrutina COEFICIENTS INICIALS**. Necessita:  $n, Z, \mathbf{Z}$  i  $\mathbf{b}$ . Retorna:  $\mathbf{w}, \mathbf{x}$  i  $e_0^{(2)}$
- Inicialitza  $k=1$
- Per tot  $k <$  màxim nombre de cicles ASA
  - Demana la **subrutina NEWTON**. Necessita:  $n, \mathbf{F}, \mathbf{w}, Z, \{\mathbf{c}_m\}$  i  $e_{k-1}^{(2)}$ . Retorna:  $\mathbf{F}_{op}, \mathbf{Z}, \mathbf{b}$  i  $e_k^{(2)}$
  - Demana la **subrutina EJR**. Necessita:  $n, Z, \mathbf{Z}, \mathbf{b}, \mathbf{x}, \mathbf{w}$  i  $e_k^{(2)}$ . Retorna:  $\mathbf{x}_{op}, \mathbf{w}_{op}$  i  $e_k^{(2)}$
  - Si  $e_{k-1}^{(2)} - e_k^{(2)} <$  tolerància  $\rightarrow$  Fi per tot  $k$ 
    - Igualta  $\mathbf{F} = \mathbf{F}_{op}, \mathbf{x} = \mathbf{x}_{op}, \mathbf{w} = \mathbf{w}_{op}$
- Fi per tot  $k$
- ❖ Retorna el valor de  $\mathbf{F}$  i  $\mathbf{w}$  òptims, i el corresponent valor de la funció  $e^{(2)}$

**Taula 3.5** Algorisme d'optimització conjunt dels coeficients i exponents ASA



El resultat de subrutina OPTIMIZA FUNCIO ASA és un conjunt de funcions ASA, amb exponents  $\mathbf{F}$  i coeficients  $\mathbf{w}$ , que minimitza la funció  $e^{(2)}$  en la zona propera als paràmetres generadors  $\{\alpha, \beta\}$  de la sèrie *even-tempered*.

**Xarxa de punts per trobar l'òptim global.** El mètode de Newton és molt sensible i sempre troba el punt estacionari més proper a la zona d'inici de la recerca. Quan la funció a optimitzar té molts d'òptims locals, com és el cas que s'estudia, el problema és trobar el mínim/màxim global. S'ha proposat construir una xarxa de punts sobre els paràmetres  $\{\alpha, \beta\}$  generadors de la sèrie *even-tempered*, i a cadascun d'ells fer la crida de la subrutina OPTIMIZA FUNCIO ASA. Així, per a cada parell de paràmetres  $\{\alpha, \beta\}$  generats, es calculen els exponents i coeficients ASA inicials sobre els quals s'aplica l'optimització de Newton dels exponents més l'ajust EJR dels coeficients amb l'objectiu de determinar el mínim local més proper a la zona del punt inicial. Repetint el càlcul en una xarxa de punts prou densa, s'obté el mínim global del problema.

### 3.9.3 Programa GATOMIC

El programa GATOMIC<sup>35</sup> és un programa d'ajust de ASA DF atòmiques. El codi aglutina totes les subrutines que s'han descrit en els apartats precedents. L'entrada de dades del programa s'ha adaptat per llegir els fitxers de sortida del programa ATOMIC<sup>36</sup> de càlcul SCF d'energies atòmiques. L'ATOMIC segueix l'estructura original SCF proposada per Roothaan i Bagus,<sup>37</sup> i implementada posteriorment en el programa dissenyat per Roos, Salez, Veillard i Clementi.<sup>38</sup> El programa pot utilitzar tant orbitals de tipus Slater com de tipus Gaussià en qualitat de funcions de base dels orbitals SCF. Calcula les energies RHF, obtenint distribucions electròniques esfèriques i simètriques. El programa GATOMIC necessita com a entrada de dades de qualsevol àtom el fitxer amb la matriu densitat RHF i el conjunt de funcions de base emprat per generar els orbitals atòmics.

El fragment de codi descrit en la taula 3.6 inclou els dos bucles necessaris per explorar la superfície descrita a través dels paràmetres  $\{\alpha, \beta\}$  generadors de la sèrie *even-tempered*. El programa requereix que s'especifiquin els valors extrems de la xarxa de punts i l'increment per a cadascuna de les variables  $\alpha$  i  $\beta$ . En tots els càlculs realitzats fins ara sobre diferents conjunts de bases de funcions atòmiques s'ha comprovat que els valors òptims dels paràmetres *even-tempered* se solen localitzar en els intervals  $\alpha \in (0,2]$  i  $\beta \in [1,6]$ , depenent del nombre de funcions i de l'àtom estudiat.

---

### Programa GATOMIC

- ✓ Donat un àtom  $a$
- ✓ Donada la matriu densitat  $D_{mm}$ , i el conjunt de funcions de base  $\{\mathbf{c}_m\}$
- ✓ Donat el nombre de funcions ASA,  $n$
- ✓ Donat els límits pels paràmetres *even-tempered*  $\{\alpha_{\min}, \alpha_{\max}\}$  i  $\{\beta_{\min}, \beta_{\max}\}$ , i els increments  $\{\Delta\alpha, \Delta\beta\}$
- Calcula el valor de l'autosemblança *ab initio*,  $Z$ , descrita en l'equació (3.33)
  - Inicialitza  $\alpha = \alpha_{\min}$  i  $\mathbf{e}_{op}^{(2)} \approx \infty$
  - Per tot  $\alpha \mid \alpha_{\min} \leq \alpha \leq \alpha_{\max}$ 
    - Inicialitza  $\beta = \beta_{\min}$
    - Per tot  $\beta \mid \beta_{\min} \leq \beta \leq \beta_{\max}$ 
      - Demana la **subrutina OPTIMITZA FUNCIO ASA**. Necessita  $n$ ,  $Z$ ,  $\{\mathbf{c}_m\}$ , i  $\{\alpha, \beta\}$ . Retorna:  $\mathbf{F}$ ,  $\mathbf{w}$  i  $\mathbf{e}^{(2)}$ 
        - Si  $\mathbf{e}^{(2)} < \mathbf{e}_{op}^{(2)}$  llavors
          - $\mathbf{e}_{op}^{(2)} = \mathbf{e}^{(2)}$ ,  $\mathbf{F}_{op} = \mathbf{F}$ ,  $\mathbf{w}_{op} = \mathbf{w}$
        - Fi del condicional
          - $\beta = \beta + \Delta\beta$
      - Fi per tot  $\beta$ 
        - $\alpha = \alpha + \Delta\alpha$
    - Fi per tot  $\alpha$
  - ❖ Sortida del programa: coeficients i exponents ASA òptims,  $\{\mathbf{F}_{op}, \mathbf{w}_{op}\}$ , per a l'àtom  $a$

---

**Taula 3.6** Esquema global del programa GATOMIC

### 3.9.4 Ajust ASA atòmic d'una base de funcions 3-21G

Amb el programa GATOMIC s'han determinat diferents bases de funcions ASA atòmiques. El primer exemple fa referència al càlcul de les funcions ASA per als àtoms H–Kr que millor ajusten la densitat HF obtinguda a partir d'una base atòmica 3-21G.<sup>39-41</sup> En l'article 3.1 es mostren els paràmetres de l'ajust per als diferents àtoms. En concret, es donen els valors de la funció error quadràtic integral i l'error relatiu produït en el càlcul de la *QS-SM*, definit com:  $\%Z_{aa} = 100(Z_{aa}^{\text{HF}} - Z_{aa}^{\text{ASA}})/Z_{aa}^{\text{HF}}$ . Amb referència al valor de  $e^{(2)}$ , s'ha de tenir present que l'ajust dels àtoms s'ha fet amb les dues *DF*, HF i ASA, normalitzades a 1. A més s'han descrit els diferents àtoms emprant més d'un conjunt de funcions ASA: per l'hidrogen i l'heli s'utilitza una sola funció Gaussiana 1s, mentre que pels àtoms Li–Ar s'han determinat conjunts de 3, 4 i 5 funcions, i per la sèrie K–Kr únicament 5 funcions. En la pàgina electrònica [42] estan llistats els coeficients i exponents de totes les funcions ASA.

**Construcció de PASA DF.** La base parametritzada de funcions atòmiques s'ha utilitzat en la construcció de *PASA DF*, corresponent a l'equació (3.49), i posterior aplicació en mesures de semblança quàntica. La generació de *PASA DF* es detalla en la taula 3.7.

---

#### Programa/Subrutina PASA DF

- ✓ Donada una molècula *A*, amb nombres atòmics  $\mathbf{P}_A = \{P_a\}$  i nombre d'àtoms  $m_A$
  - ✓ Donat un conjunt de funcions ASA atòmiques:  $\{\mathbf{F}_c, \mathbf{w}_c\} \forall c \in \{\text{H, He, ...}\}$
  - ✓ Inicialitza  $n_A=0$
  - Per tot  $a \in A$ 
    - Busca en la base de funcions ASA l'àtom amb nombre atòmic  $P_a$ :  $\{n_{Pa}, \mathbf{F}_{Pa}, \mathbf{x}_{Pa}, \mathbf{w}_{Pa}\}$
    - $n_A = n_A + n_{Pa}$  ;  $\mathbf{F}_A = \mathbf{F}_A + \mathbf{F}_{Pa}$  ;  $\mathbf{x}_A = \mathbf{x}_A + \mathbf{x}_{Pa}$  ;  $\mathbf{w}_A = \mathbf{w}_A + \mathbf{w}_{Pa}$
  - Fi per tot  $a$
  - ❖ Sortida del programa/subrutina: nombre de funcions, exponents i coeficients ASA de la molècula *A*:  $\{n_A, \mathbf{F}_A, \mathbf{x}_A, \mathbf{w}_A\}$
- 

Taula 3.7 Esquema del Programa/Subrutina PASA DF

Únicament es necessita conèixer quins àtoms formen la molècula i especificar una base de funcions ASA atòmiques. El principal avantatge de les funcions promoleculares és l'omissió del càlcul de la  $DF$  molecular a nivell *ab initio*.

**Exemples moleculars emprant l'aproximació PASA.** El funcionament de l'algorisme PASA i la precisió de les densitats resultants en el càlcul de mesures de semblança s'ha analitzat a través de varis exemples numèrics presentats en l'article 3.1. El primer és la representació dels mapes de semblança molecular quàntica dels sistemes HCCH/Ne i NNO/Ne. Els mapes de semblança són una representació bidimensional de la superfície definida per les mesures de semblança que s'obtenen quan es trasllada un àtom en un pla definit en l'entorn d'una molècula.<sup>13</sup> En l'article 3.1 han servit per comparar les mesures PASA amb les *ab initio*. Si s'analitzen els mapes de semblança presentats s'observa que en les zones properes als nuclis atòmics es produeix menys error entre les mesures PASA i *ab initio*, mentre que en els enllaços hi ha més diferències, tal com caldria esperar d'una aproximació promolecular. El segon exemple d'utilització de les funcions PASA el formen nou derivats fluorats i clorats del metà. És un conjunt de molècules que ha servit de test en precedents metodologies d'ajust de la  $DF$ .<sup>10-12</sup> El primer estadi és el càlcul de la superposició molecular òptima dels 36 parells de compostos involucrats mitjançant el mètode de màxima semblança que es descriu en el capítol 4. El procés d'alineament es fa amb les densitats PASA. Llavors amb les geometries moleculars de la superposició òptima es fa un càlcul puntual HF/3-21G emprant el programa GAUSSIAN,<sup>43</sup> i amb les  $DF$  resultants es calcula la MQSM *ab initio*. En l'article 3.1 es mostra la matriu de mesures de semblança i la d'índex de Carbó, definit en l'equació (2.19), per les dues densitats electròniques, PASA i *ab initio*. Dels resultats obtinguts destaca la gran similitud entre els índexs de Carbó d'ambdues mesures. L'últim exemple que es presenta el componen vuit molècules extretes de la Cambridge Structural Database,<sup>44</sup> amb estructures molt diverses i que tenen àtoms pesants com Cr, Zn, As i Br. En aquest cas només s'ha calculat el percentatge d'error en la mesura d'autosemblança  $Z_{AA}$  emprant les densitats PASA i exacta.

**Article 3.1**

---

**Autors:** *Lluís Amat, Ramon Carbó-Dorca.*

**Títol:** *Quantum similarity measures under atomic self approximation: first order density fitting using elementary Jacobi rotations*

**Revista:** *Journal of Computational Chemistry*

**Volum:** 18      **Pàgines, inicial:** 2023   **final:** 2039   **Any:** 1997

---

---

# Quantum Similarity Measures under Atomic Shell Approximation: First Order Density Fitting Using Elementary Jacobi Rotations

---

LLUÍS AMAT, RAMON CARBÓ-DORCA

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

*Received 2 May 1997; accepted 30 July 1997*

---

**ABSTRACT:** The elementary Jacobi rotations technique is proposed as a useful tool to obtain fitted electronic density functions expressed as linear combinations of atomic spherical shells, with the additional constraint that all coefficients are kept positive. Moreover, a Newton algorithm has been implemented to optimize atomic shell exponents, minimizing the quadratic error integral function between *ab initio* and fitted electronic density functions. Although the procedure is completely general, as an application example both techniques have been used to compute a *1S-type* Gaussian basis for atoms H through Kr, fitted from a 3-21G basis set. Subsequently, molecular electronic densities are modeled in a *promolecular* approximation, as a simple sum of parameterized atomic contributions. This simple molecular approximation has been employed to show, in practice, its usefulness to some computational examples in the field of molecular quantum similarity measures. © 1997 John Wiley & Sons, Inc. *J Comput Chem* **18**: 2023–2039, 1997

**Keywords:** atomic shell approximation; Carbó Index; elementary Jacobi rotations; promolecular densities; quadratic error integral function; quantum similarity measures

*Correspondence to:* R. Carbó-Dorca

Contract grant sponsor: CICYT; contract grant number:  
SAF 96-0158

Contract grant sponsor: Spanish Ministerio de Educación y  
Cultura

## Introduction

The elementary Jacobi rotations (EJR) technique as a source of metric-preserving orthogonal transformations has been used in many fields of quantum chemistry. Besides the task of electronic energy optimization,<sup>1</sup> EJR's have been employed in other areas such as the localization of orbitals<sup>2</sup> and the variational determination of orbitals in molecular pair wave functions.<sup>3</sup> In our laboratory, a large contribution of EJR to direct electronic energy optimization has been developed<sup>4</sup> as well as some earlier applications related to quantum similarity.<sup>5</sup> Also, work has been performed on the many aspects of the Jacobi diagonalization algorithm,<sup>6</sup> proposing a new parallelizable procedure, constituting a practical computational scheme,<sup>7</sup> able to deal with large matrices and a chosen set of eigenvalues and eigenvectors. Recently, an application of EJR to the study of approximate full CI wave functions has also been reported.<sup>8</sup>

The present article will deal, in a general manner, with the proposal of using the EJR technique to solve the problem of fitting the usual first order electronic density functions to a linear combination of *nS-type* spherical functions, as has been done by several investigators.<sup>9,10</sup> However, this work demands the fulfillment of the additional fitting constraint, which is that all coefficients have to be positive.<sup>11</sup> This is in order to preserve the *statistical meaning* of the density function,<sup>12</sup> which has to be considered as a probability distribution and thus it should be positive definite *everywhere*. The fitted approximate density employed in this work has been modulated using a superposition of several atomic spherical function squared modules, the so-called atomic shell approximation (ASA).<sup>11</sup> In this work only atomic density fittings have been performed and subsequent application to molecular quantum similarity measures (QSM) evaluation<sup>13-15</sup> will be carried out using a *promolecular* approximation.<sup>16</sup> The main advantage of fitting *nS-type* functions for practical application purposes is the low computational costs compared to *ab initio* calculations. In this way, the feasibility of QSAR project studies involving big sets of large sized molecules of biological interest can be envisaged.<sup>17</sup>

To achieve this objective, the present study will first study the description of QSM and ASA-type

functions. The constrained least-squares fitting problem, consisting of minimizing the quadratic error integral function between *ab initio* and ASA electronic density functions, will be studied next. Apart from the least-squares EJR technique, a Newton algorithm will also be constructed to find the optimal ASA exponents. The corresponding computational algorithms will be detailed and some illustrative examples discussed. Finally, using the *promolecular* approximation some molecular QSM will be presented as an application example.

## Quantum Similarity Measures

A QSM is defined as the integral involving at least two density functions,  $\{\rho_A(\mathbf{r}_1), \rho_B(\mathbf{r}_2)\}$ , and a given positive definite operator  $\Omega(\mathbf{r}_1, \mathbf{r}_2)$ , where the vector  $\mathbf{r}$  contains the coordinates of the involved electrons. The QSM integral may be formally written as:

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

When both compared systems are the same, then the integral is denoted as  $Z_{AA}$ , and referred to as a quantum self-similarity measure (QS-SM), and when the quantum systems studied are molecules, the literature speaks of molecular QSM (MQSM). The first MQSM defined, and also the most common, is the so-called overlap type.<sup>14a</sup> This choice corresponds to the replacement of  $\Omega(\mathbf{r}_1, \mathbf{r}_2)$  by a Dirac delta function, transforming eq. (1) into:

$$Z_{AB} = \int \rho_A(\mathbf{r}) \rho_B(\mathbf{r}) d\mathbf{r} \quad (2)$$

Accurate MQSM may be obtained using a LCAO approach to describe molecular electronic density functions, corresponding to the expression:

$$\rho_A = \sum_{\mu, \nu} D_{\mu\nu} \chi_\mu(\mathbf{r}) \chi_\nu(\mathbf{r}) \quad (3)$$

where  $D_{\mu\nu}$  is the charge-bond order matrix, and  $\{\chi_\mu\}$  the atomic orbital basis set functions. Using this approximation, four-center integrals have to be computed in MQSM. Besides these computational difficulties, there is the additional problem of the maximization of MQSM:<sup>18</sup> the function,  $Z_{AB}$ , depends on the relative position of molecules *A* and *B*, and the similarity between them is usually defined as the maximum. It is not surprising,

therefore, that the major disadvantage of *ab initio* calculations is the fact that quantum similarity integral computation is cumbersome. For this reason computational algorithms have been developed to calculate approximate density functions.<sup>9–11</sup>

QSM theory is an ideal mathematical construction, logically placed into the structure of quantum mechanics,<sup>13b,c</sup> which may be used whenever it is necessary to compare two or more density functions. In fact, this may be seen from a more general point of view, associating the density functions in eq. (1) with square summable definite positive functions. From the mathematical point of view, a similarity integral defined as  $Z_{AB}$ , can be interpreted as positive-valued, weighted scalar products.

## Atomic Shell Approximation

The construction of ASA-like density functions can be traced up to the initial studies on QSM,<sup>14a</sup> where a CNDO-like<sup>19</sup> approach was invoked to deal with the computational problem of the evaluation of the quantum similarity integrals,  $Z_{AB}$ , for molecules. From this initial viewpoint, the concept of approximating a given density function, has led to the consideration of ASA as a general superposition of spherical *nS-type*, STO, or GTO, functions.

The density function in ASA form may be written in terms of atomic function set contributions  $\{\sigma_a\}$ :

$$\rho_A^{\text{ASA}}(\mathbf{r}) = \sum_{a \in A} \sigma_a(\mathbf{r}) \quad (4)$$

where the sum runs over all the atoms present in molecule *A*. At the same time, the atomic set of functions  $\{\sigma_a\}$ , may be constructed using a set of atomic *shells*  $\{s_i\}$ , using the sum:

$$\sigma_a(\mathbf{r}) = \sum_{i \in a} s_i(\mathbf{r}) \quad (5)$$

and the sum is over all the atomic *shells* of the *ath* atom. Finally, the spherical function set  $\{s_i\}$ , can be defined as:

$$s_i(\mathbf{r}) = \sum_{k \in i} c_k \varphi_k(\mathbf{r}) \quad (6)$$

where the sum is carried out over the entire positive definite function set  $\{\varphi_k\}$ , belonging to shell *i*. The coefficients  $\{c_k\}$  must be *positive in all cases* to keep the distribution structure of the approximated function,  $\rho_A^{\text{ASA}}$ , positive definite.

The above ASA partition is equivalent to writing eq. (4) in a more compact notation, as a linear combination of a positive definite function set  $\{\theta_i\}$ :

$$\rho_A^{\text{ASA}}(\mathbf{r}) = \sum_{i \in A} w_i \theta_i(\mathbf{r}) \quad (7)$$

where the sum is performed over all the basis function set  $\{\theta_i\}$ , and the set of positive coefficients  $\{w_i\}$  must be determined.

The most interesting case is constituted by the ASA approximation of first order density functions, but other high level density function forms may be considered as well. In any case, both the exact and the ASA density functions may be supposed to be normalized to one particle, by dividing by the appropriate particle number factors, and in eq. (7), considering the basis functions normalized as follows:

$$\int \theta_i(\mathbf{r}) d\mathbf{r} = 1, \forall i \in A \quad (8)$$

then, necessarily the set of ASA coefficients  $\{w_i\}$ , apart from the condition:

$$w_i > 0, \forall i \in A \quad (9)$$

must fulfill the additional constraint:

$$\sum_{i \in A} w_i = 1 \quad (10)$$

Although the second condition may be easily taken into account with a Lagrange multiplier technique,<sup>20</sup> the first one, as expressed in eq. (9), could not be so easily introduced into the computational algorithm.<sup>11</sup> It will be shown here that both conditions can be maintained through the optimization process by using the appropriate tools.

## PROMOLECULAR ATOMIC SHELL APPROXIMATION

Recently, similarity measures applied to QSAR studies<sup>21</sup> have gained more importance in computational chemistry. When QSMs are used in QSAR studies, ASA will represent a considerable saving in computation time. However, this approach is still insufficient when a vast number of large molecules are studied. This is due to the necessity of obtaining both the *ab initio* electronic density and the fitted functions (optimal coefficients and exponents) for each molecule.<sup>11</sup> A further theoretical and computational simplification is achieved



when the so-called *promolecular* approximation is considered, and then QSAR studies can be performed<sup>17</sup> without much effort. Within this approach, the total molecular electronic density is calculated as a sum of atomic electronic density contributions:

$$\rho_A^{\text{ASA}}(\mathbf{r}) = N_A^{-1} \sum_a N_a \rho_a^{\text{ASA}}(\mathbf{r}) \quad (11)$$

where  $N_A$  is the total number of electrons in molecule  $A$ , and  $N_a$  the atomic number of each atom  $a$ . The sum in eq. (11) runs over all the molecular atoms.

Every atomic density function  $\rho_a^{\text{ASA}}$  is constructed with the same structure as the one given in eq. (7), but replacing  $\theta_i$  by a squared *nS-type* spherical function centered on the  $a$ th atom and keeping the coefficient convex constraints, given in eq. (9) and (10):

$$\rho_a^{\text{ASA}}(\mathbf{r}) = \sum_i w_i S_i(\mathbf{r} - \mathbf{r}_a)^2 \quad (12)$$

Using this approximation, overlap-like QSM between two atoms can be expressed by:

$$Z_{ab} = \sum_i w_i \sum_j w_j Z_{ij} \quad (13)$$

where the  $Z_{ij}$  elements belong to the similarity matrix,  $\mathbf{Z}$ , and are defined by the integral:

$$Z_{ij} = \int S_i(\mathbf{r} - \mathbf{r}_a)^2 S_j(\mathbf{r} - \mathbf{r}_b)^2 d\mathbf{r} \quad (14)$$

the similarity matrix being positive definite.

### Constrained Fitting of ASA Coefficients

The set of  $\{w_i\}$  optimal coefficients, appearing in eq. (12), have been calculated minimizing the quadratic error integral function between *ab initio* and ASA electronic density functions. The usual form of the quadratic error function is:

$$\begin{aligned} \varepsilon^{(2)} = & \int \rho_a(\mathbf{r}) - \rho_a^{\text{ASA}}(\mathbf{r})^2 d\mathbf{r} \\ & Z_{aa} \sum_{i,j} w_i w_j Z_{ij} - 2 \sum_i w_i \\ & \sum_{\mu, \nu} D_{\mu\nu} \int S_i(\mathbf{r})^2 \chi_\mu(\mathbf{r}) \chi_\nu(\mathbf{r}) d\mathbf{r} \quad (15) \end{aligned}$$

where  $Z_{aa}$  corresponds to the *ab initio* QS-SM of atom  $a$ :

$$Z_{aa} = \int \rho_a(\mathbf{r})^2 d\mathbf{r} \quad (16)$$

which is computed within the LCAO approximation, replacing electronic density,  $\rho_a(\mathbf{r})$ , by the corresponding expression described in eq. (3).

Eq. (15) may be written in matrix form as:

$$\varepsilon^{(2)} = \mathbf{Z}_{aa} \mathbf{w}^T \mathbf{Z} \mathbf{w} - 2 \mathbf{b}^T \mathbf{w} \quad (17)$$

where the elements of the vector  $\mathbf{b} = \{b_i\}$  are given by the integral:

$$b_i = \sum_{\mu, \nu} D_{\mu\nu} \int S_i(\mathbf{r})^2 \chi_\mu(\mathbf{r}) \chi_\nu(\mathbf{r}) d\mathbf{r} \quad (18)$$

and  $\mathbf{w}$  is a normalized column vector ( $\mathbf{w}^T \mathbf{w} = 1$ ) containing the ASA coefficients.

The set of positive coefficients  $\{w_i\}$  can be easily defined as:

$$w_i = x_i^2 \quad (19)$$

fulfilling condition (9). If they are taken as real, then the quadratic error integral function is simply expressed in terms of the squared coefficients as:

$$\varepsilon^{(2)} = \mathbf{Z}_{aa} \sum_{i,j} x_i^2 x_j^2 Z_{ij} - 2 \sum_i x_i^2 b_i \quad (20)$$

On the other hand, due to the fact that the similarity matrix,  $\mathbf{Z}$ , is positive definite, a unitary matrix,  $\mathbf{U}$ , can be found such as:

$$\mathbf{U} \mathbf{Z} \mathbf{U} = \mathbf{D} \quad (21)$$

where  $\mathbf{D}$  is a diagonal matrix. The first step in the procedure consists in diagonalizing the similarity matrix  $\mathbf{Z}$ , to obtain their eigenvalues and eigenvectors. Then,  $\{x_i\}$  coefficients are extracted from the most promising normalized eigenvector of the  $\mathbf{Z}$  matrix; that is, the one producing the minimal quadratic error. Consequently, the required constraint, specified in eq. (10), is fulfilled from the beginning. Starting from this initial vector, and applying orthogonal EJR over its components, the required conditions in eq. (9) and (10) will be maintained throughout the iterative procedure.

### VARIATION OF QUADRATIC ERROR INTEGRAL FUNCTION APPLYING ELEMENTARY JACOBI ROTATIONS

EJRs are good tools to obtain elementary orthogonal transformations usable over vectors or matri-

ces. The origin of such transformation matrices can be found in Jacobi's study,<sup>6</sup> dating back to the last century. Being orthogonal, EJR may also be viewed as rotation matrices over  $n$ -dimensional spaces. Applied to a given  $n$ -dimensional vector, an EJR, which is written as  $J_{pq}(\alpha)$ , transforms the corresponding  $p$  and  $q$  components, keeping the rest of them invariable. Calling  $\mathbf{x} = \{x_i\}$  the vector to be transformed by means of the EJR  $J_{pq}(\alpha)$ , the transformation is defined by the equations:

$$\begin{aligned} \dot{x}_p &\leftarrow c x_p - s x_q \\ \dot{x}_q &\leftarrow s x_p + c x_q \end{aligned} \quad (22)$$

where  $c$  and  $s$  are the cosine and sine of the rotation angle  $\alpha$ . The norm of the new vector,  $\dot{\mathbf{x}} = \{\dot{x}_i\}$ , obtained after applying the EJR  $J_{pq}(\alpha)$ , remains invariable with respect to the initial one.

ASA coefficients are obtained by an optimization procedure, which minimizes  $\varepsilon^{(2)}$ . Isolating the  $p$  and  $q$  elements of the vector  $\mathbf{x}$  from the rest, eq. (20) gives:

$$\begin{aligned} \varepsilon^{(2)} &= Z_{aa} x_p^4 Z_{pp} x_q^4 Z_{qq} + 2 x_p^2 x_q^2 Z_{pq} \\ &+ 2 x_p^2 \sum_{i \neq p,q} x_i^2 Z_{pi} + 2 x_q^2 \sum_{i \neq p,q} x_i^2 Z_{iq} \\ &+ \sum_{i \neq p,q} \sum_{j \neq p,q} x_i^2 x_j^2 Z_{ij} + 2 x_p^2 b_p + 2 x_q^2 b_q \\ &+ 2 \sum_{i \neq p,q} x_i^2 b_i \end{aligned} \quad (23)$$

Over this equation, it is easy to apply the EJR  $J_{pq}(\alpha)$ , thus the variation of  $\varepsilon^{(2)}$  respects the active pair of elements  $\{p, q\}$ , and may be expressed as:

$$\begin{aligned} \delta \varepsilon^{(2)} &= \delta x_p^4 Z_{pp} + \delta x_q^4 Z_{qq} + 2 \delta (x_p^2 x_q^2) Z_{pq} \\ &+ 2 \delta x_p^2 \sum_{i \neq p,q} x_i^2 Z_{pi} + 2 \delta x_q^2 \sum_{i \neq p,q} x_i^2 Z_{iq} \\ &+ 2 \delta x_p^2 b_p + 2 \delta x_q^2 b_q \end{aligned} \quad (24)$$

The squared transformed coefficient elements,  $\dot{x}_p^2$  and  $\dot{x}_q^2$ , necessary to obtain expressions for the variations  $\delta x_p^2$  and  $\delta x_q^2$ , are easily obtained from eq. (22), as:

$$\begin{aligned} \dot{x}_p^2 &= (c x_p - s x_q)^2 = c^2 x_p^2 - s^2 x_q^2 - 2cs x_p x_q \\ \dot{x}_q^2 &= s^2 (x_p^2 + x_q^2) + 2cs x_p x_q \\ \dot{x}_q^2 &= (s x_p + c x_q)^2 = s^2 x_p^2 + c^2 x_q^2 + 2cs x_p x_q \\ \dot{x}_p^2 &= s^2 (x_p^2 + x_q^2) - 2cs x_p x_q \end{aligned} \quad (25)$$

and their second-order variations are given by:

$$\begin{aligned} \delta x_p^2 &= \begin{pmatrix} \dot{x}_p^2 & x_p^2 \\ s^2 (x_p^2 + x_q^2) & 2cs x_p x_q \end{pmatrix} \Delta \\ \delta x_q^2 &= \begin{pmatrix} \dot{x}_q^2 & x_q^2 \\ s^2 (x_p^2 + x_q^2) & 2cs x_p x_q \end{pmatrix} \Delta \end{aligned} \quad (26)$$

The fourth power of the transformed coefficients is defined from the quadratic ones as follows:

$$\begin{aligned} \dot{x}_p^4 &= (\dot{x}_p^2)^2 = (x_p^2 + \Delta)^2 = x_p^4 + \Delta^2 + 2x_p^2 \Delta \\ \dot{x}_q^4 &= (\dot{x}_q^2)^2 = (x_q^2 + \Delta)^2 = x_q^4 + \Delta^2 + 2x_q^2 \Delta \\ \dot{x}_p^2 \dot{x}_q^2 &= (x_p^2 + \Delta)(x_q^2 + \Delta) \\ &= x_p^2 x_q^2 + (x_p^2 + x_q^2) \Delta + \Delta^2 \end{aligned} \quad (27)$$

and their corresponding variation,  $\delta x_p^4$ ,  $\delta x_q^4$ , and  $\delta(x_p^2 x_q^2)$ , are defined by:

$$\begin{aligned} \delta x_p^4 &= (\dot{x}_p^4 - x_p^4) = \Delta^2 + 2x_p^2 \Delta \\ \delta x_q^4 &= (\dot{x}_q^4 - x_q^4) = \Delta^2 + 2x_q^2 \Delta \\ \delta(x_p^2 x_q^2) &= (\dot{x}_p^2 \dot{x}_q^2 - x_p^2 x_q^2) = (x_p^2 + x_q^2) \Delta + \Delta^2 \end{aligned} \quad (28)$$

Developing the  $\delta x_p^4$  and  $\delta x_q^4$  expressions in eq. (28) gives:

$$\begin{aligned} \delta x_p^4 &= s^4 (x_p^2 - x_q^2)^2 + 4c^2 s^2 x_p^2 x_q^2 \\ &+ 4cs^3 (x_p^2 - x_q^2) x_p x_q \\ &+ 2x_p^2 [s^2 (x_p^2 + x_q^2) - 2cs x_p x_q] \\ \delta x_q^4 &= s^4 [(x_p^2 + x_q^2)^2 - 4x_p^2 x_q^2] \\ &+ 4cs^3 (x_p^2 + x_q^2) x_p x_q \\ &+ 2s^2 x_p^2 [(x_p^2 + x_q^2) - 2x_q^2] \\ &+ 4cs x_p^3 x_q \end{aligned} \quad (29)$$

and:

$$\begin{aligned} \delta x_q^4 &= s^4 (x_p^2 - x_q^2)^2 + 4c^2 s^2 x_p^2 x_q^2 \\ &+ 4cs^3 (x_p^2 - x_q^2) x_p x_q \\ &+ 2x_q^2 [s^2 (x_p^2 + x_q^2) - 2cs x_p x_q] \end{aligned}$$

$$\begin{aligned}
 & s^4 \left[ \begin{pmatrix} x_p^2 & x_q^2 \\ x_p^2 & x_q^2 \end{pmatrix}^2 \quad 4x_p^2 x_q^2 \right] \\
 & 4cs^3 \begin{pmatrix} x_p^2 & x_q^2 \\ x_p^2 & x_q^2 \end{pmatrix} x_p x_q \\
 & 2s^2 x_q^2 \left[ \begin{pmatrix} x_p^2 & x_q^2 \\ x_p^2 & x_q^2 \end{pmatrix} \quad 2x_p^2 \right] \quad 4cs x_p x_q^3
 \end{aligned} \tag{30}$$

whereas the term  $\delta(x_p^2 x_q^2)$  adopts the form:

$$\begin{aligned}
 \delta(x_p^2 x_q^2) &= s^2 \begin{pmatrix} x_p^2 & x_q^2 \\ x_p^2 & x_q^2 \end{pmatrix}^2 \quad 2cs \begin{pmatrix} x_p^2 & x_q^2 \\ x_p^2 & x_q^2 \end{pmatrix} x_p x_q \\
 & s^4 \begin{pmatrix} x_p^2 & x_q^2 \\ x_p^2 & x_q^2 \end{pmatrix}^2 \quad 4c^2 s^2 x_p^2 x_q^2 \\
 & 4cs^3 \begin{pmatrix} x_p^2 & x_q^2 \\ x_p^2 & x_q^2 \end{pmatrix} x_p x_q \\
 & s^4 \left[ \begin{pmatrix} x_p^2 & x_q^2 \\ x_p^2 & x_q^2 \end{pmatrix}^2 \quad 4x_p^2 x_q^2 \right] \\
 & 4cs^3 \begin{pmatrix} x_p^2 & x_q^2 \\ x_p^2 & x_q^2 \end{pmatrix} x_p x_q \\
 & s^2 \left[ \begin{pmatrix} x_p^2 & x_q^2 \\ x_p^2 & x_q^2 \end{pmatrix}^2 \quad 4x_p^2 x_q^2 \right] \\
 & 2cs \begin{pmatrix} x_p^2 & x_q^2 \\ x_p^2 & x_q^2 \end{pmatrix} x_p x_q
 \end{aligned} \tag{31}$$

Substituting expressions  $\delta x_p^2, \delta x_q^2, \delta x_p^4, \delta x_q^4$ , and  $\delta(x_p^2 x_q^2)$  into eq. (24), and collecting terms, one has:

$$\delta\epsilon^{(2)} = E_{04}s^4 + E_{13}cs^3 + E_{02}s^2 + E_{11}cs \tag{32}$$

where the new parameter set  $\{E_{cs}\}$  introduced in eq. (32) is defined using:

$$\begin{aligned}
 E_{04} &= (Z_{pp} \quad Z_{qq} \quad 2Z_{pq}) \left[ \begin{pmatrix} x_p^2 & x_q^2 \\ x_p^2 & x_q^2 \end{pmatrix}^2 \quad 4x_p^2 x_q^2 \right] \\
 E_{13} &= 4(Z_{pp} \quad Z_{qq} \quad 2Z_{pq}) \begin{pmatrix} x_p^2 & x_q^2 \\ x_p^2 & x_q^2 \end{pmatrix} x_p x_q \\
 E_{02} &= 4(Z_{pp} \quad Z_{qq} \quad 2Z_{pq}) x_p^2 x_q^2 \quad 2 \begin{pmatrix} x_p^2 & x_q^2 \\ x_p^2 & x_q^2 \end{pmatrix} G \\
 E_{11} &= 4x_p x_q G
 \end{aligned} \tag{33}$$

and

$$G = \sum_{i=p,q} x_i^2 (Z_{pi} \quad Z_{qi}) \quad x_p^2 Z_{pp} \quad x_q^2 Z_{qq} \quad \begin{pmatrix} x_p^2 & x_q^2 \\ x_p^2 & x_q^2 \end{pmatrix} Z_{pq} \quad b_p \quad b_q$$

The optimal sine belonging to the EJR can be chosen with the gradient condition  $\frac{d\delta\epsilon^{(2)}}{ds} = 0$ ; that is:

$$\begin{aligned}
 \frac{d\delta\epsilon^{(2)}}{ds} &= 4E_{04}s^3 + E_{13}(ts^3 + 3cs^2) \\
 & 2E_{02}s + E_{11}(ts + c) \\
 & c[(E_{13}s^2 + E_{11})t^2
 \end{aligned}$$

$$\begin{aligned}
 & 2(2E_{04}s^2 + E_{02})t + (3E_{13}s^2 + E_{11}) \\
 & c(T_1 t^2 + 2T_2 t + T_3) = 0
 \end{aligned} \tag{34}$$

where  $t = \frac{s}{c}$  and  $\frac{dc}{ds} = t$ .

The best EJR angle is found by solving the second-degree equation appearing in expression (34). The optimization is conducted by an iterative procedure, until the variation of the EJR angle or the quadratic error integral function value becomes negligible.

### Optimization of ASA Exponents

A Newton algorithm<sup>20</sup> has been used to optimize the exponents of the fitted *nS-type* functions in ASA. This method is operative when the analytic gradient vector and the Hessian matrix are available, as in the present ASA case. Newton's equation may be written as:

$$\Phi^{k+1} = \Phi^k + (\mathbf{H}^k)^{-1} \mathbf{g}^k \tag{35}$$

where  $\Phi = \{\zeta_i\}$  is a vector containing all the exponents of the fitted basis set functions. The superscripts  $k$  and  $k+1$  represent the points before and after every iteration, and  $\mathbf{g}$  and  $\mathbf{H}$  are the gradient vector and the Hessian matrix in the step  $k$ , respectively.

At every iteration the gradient vector elements are calculated in the following manner:

$$g_i = \frac{\partial\epsilon^{(2)}}{\partial\zeta_i} = x_i^4 \frac{\partial Z_{ii}}{\partial\zeta_i} + x_i^2 \sum_j x_j^2 \frac{\partial Z_{ij}}{\partial\zeta_i} + 2x_i^2 \frac{\partial b_i}{\partial\zeta_i} \tag{36}$$

and the expression of the Hessian matrix elements are given by:

$$\begin{aligned}
 h_{ii} &= \frac{\partial^2\epsilon^{(2)}}{\partial\zeta_i^2} = x_i^4 \frac{\partial^2 Z_{ii}}{\partial\zeta_i^2} + x_i^2 \sum_j x_j^2 \frac{\partial^2 Z_{ij}}{\partial\zeta_i^2} \\
 & 2x_i^2 \frac{\partial^2 b_i}{\partial\zeta_i^2} \\
 h_{ij} &= \frac{\partial^2\epsilon^{(2)}}{\partial\zeta_i \partial\zeta_j} = x_i^2 x_j^2 \frac{\partial^2 Z_{ij}}{\partial\zeta_i \partial\zeta_j}
 \end{aligned} \tag{37}$$

The Newton method is generally reliable and reasonably efficient. The optimization stops when the difference of the quadratic error integral function between two successive steps is less than a given threshold.

**FIRST AND SECOND DERIVATIVES OF QUADRATIC ERROR INTEGRAL FUNCTION**

The gradient vector and Hessian matrix, corresponding to eq. (36) and (37), should be computed to solve the Newton eq. (35). These expressions require the following derivatives:

$$\frac{\partial Z_{ii}}{\partial \zeta_i} \quad \frac{\partial Z_{ij}}{\partial \zeta_i} \quad \frac{\partial b_i}{\partial \zeta_i} \tag{38}$$

and:

$$\frac{\partial^2 Z_{ii}}{\partial \zeta_i^2} \quad \frac{\partial^2 Z_{ij}}{\partial \zeta_i^2} \quad \frac{\partial^2 b_i}{\partial \zeta_i^2} \quad \frac{\partial^2 Z_{ij}}{\partial \zeta_i \partial \zeta_j} \tag{39}$$

A normalized nS Gaussian function in the spherical polar coordinate system may be written, after integrating the angular part, as:

$$S_i(r) = N_i r^{n_i} e^{-\zeta_i r^2} \tag{40}$$

$$N_i = \left[ \frac{2^{2n_i} \pi^{1/2} \zeta_i^{n_i}}{(2n_i - 1)!! \pi^{3/2}} \right]^{1/2}$$

where  $n_i$  is the principal quantum number of the  $i$ th atomic shell and  $N_i$  the normalization constant. A factor  $(1/4\pi)^{1/2}$  is included in the normalization factor, coming from the integration of the angular part.

Substituting expression (40) into eq. (14), the  $Z_{ij}$  measure becomes:

$$Z_{ij} = 4\pi N_i^2 N_j^2 \int_0^\infty r^{2(n_i + n_j - 1)} e^{-2(\zeta_i + \zeta_j)r^2} dr \tag{41}$$

and integrating this equation yields:

$$Z_{ij} = \left(\frac{2}{\pi}\right)^{3/2} \left\{ \frac{[2(n_i + n_j - 3)]!!}{(2n_i - 1)!! (2n_j - 1)!!} \right\} \frac{\zeta_i^{n_i - 1/2} \zeta_j^{n_j - 1/2} (\zeta_i + \zeta_j)^{1/2 - n_i - n_j}}{\zeta_i^{n_i} \zeta_j^{n_j}} \tag{42}$$

In the case where  $i = j$  eq. (42) is simplified to:

$$Z_{ii} = \left\{ \frac{2^{2 - 2n_i} (4n_i - 3)!!}{\pi^{3/2} (2n_i - 1)!!^2} \right\} \zeta_i^{3/2 - 2n_i} \tag{43}$$

Differentiating  $Z_{ij}$  with respect to the exponent  $\zeta_i$  gives:

$$\frac{\partial Z_{ij}}{\partial \zeta_i} = Z_{ij} \left( \frac{n_i - 1/2}{\zeta_i} - \frac{n_i + n_j - 1/2}{\zeta_i + \zeta_j} \right) \tag{44}$$

and, in the same way,  $Z_{ii}$  measure produces:

$$\frac{\partial Z_{ii}}{\partial \zeta_i} = Z_{ii} \frac{3}{2\zeta_i} \tag{45}$$

To evaluate the  $\mathbf{b}$  vector elements, the spherical functions,  $S_i(r)$ , of expression (40) are substituted into eq. (18):

$$b_i = 4\pi N_i^2 \sum_{\mu, \nu} D_{\mu\nu} N_\mu N_\nu \int_0^\infty r^{2n_i - n_\mu - n_\nu - 2} e^{-(2\zeta_i + \zeta_\mu + \zeta_\nu)r^2} dr \tag{46}$$

These integrals depend on the exponent of  $r$ . Defining  $m = 2n_i - n_\mu - n_\nu - 2$ , then  $b_i$  is equal to:

$$b_i = \sum_{\mu, \nu} D_{\mu\nu} N_\mu N_\nu \left[ \frac{(2n_i - n_\mu - n_\nu - 3)!! 2^{\frac{2n_i - n_\mu - n_\nu - 3}{2}}}{(2n_i - 1)!!} \right] \frac{(\zeta_i)^{n_i - 1/2}}{(2\zeta_i + \zeta_\mu + \zeta_\nu)^{\frac{2n_i - n_\mu - n_\nu - 1}{2}}} \tag{47}$$

if  $m$  is an even number, and if it is odd:

$$b_i = \sum_{\mu, \nu} D_{\mu\nu} N_\mu N_\nu \left\{ \frac{[1 - 2(2n_i - n_\mu - n_\nu - 3)]! 2^{2n_i - 1/2}}{(2n_i - 1)!!} \right\} \frac{(\zeta_i)^{n_i - 1/2}}{(2\zeta_i + \zeta_\mu + \zeta_\nu)^{\frac{2n_i - n_\mu - n_\nu - 1}{2}}} \tag{48}$$

In both solutions, the  $b_i$  integral may be written using an auxiliary function  $B$ , which depends on the  $i$ th atomic shell and the variable  $m$ :

$$b_i = \sum_{\mu, \nu} B_{\mu\nu}(i, m) \tag{49}$$

Using this notation, the derivative of  $b_i$  with respect to the  $\zeta_i$  exponent, is given by:

$$\frac{\partial b_i}{\partial \zeta_i} = \sum_{\mu, \nu} \left( \frac{n_i - 1/2}{\zeta_i} - \frac{2n_i - n_\mu - n_\nu - 1}{2\zeta_i + \zeta_\mu + \zeta_\nu} \right) B_{\mu\nu}(i, m) \tag{50}$$

Inserting eq. (44), (45), and (50) into eq. (36), the following expression can be obtained:

$$g_i = x_i^4 Z_{ii} \frac{3}{2\zeta_i} x_i^2 \sum_j x_j^2 Z_{ij} \left[ \left( \frac{n_i \ 1 \ 2}{\zeta_i} \quad \frac{n_i \ n_j \ 1 \ 2}{\zeta_i \ \zeta_j} \right) 2x_i^2 \left[ \sum_{\mu, \nu} \left( \frac{n_i \ 1 \ 2}{\zeta_i} \quad \frac{2n_i \ n_\mu \ n_\nu \ 1}{2\zeta_i \ \zeta_\mu \ \zeta_\nu} \right) B_{\mu\nu}(i, m) \right] \right] \quad (51)$$

The second derivatives needed to calculate the  $h_{ii}$  Hessian matrix elements are:

$$\frac{\partial^2 Z_{ii}}{\partial \zeta_i^2} = Z_{ii} \frac{3}{4\zeta_i^2} \quad (52)$$

and:

$$\frac{\partial^2 Z_{ij}}{\partial \zeta_i^2} = Z_{ij} \left[ \left( \frac{n_i \ 1 \ 2}{\zeta_i} \quad \frac{n_i \ n_j \ 1 \ 2}{\zeta_i \ \zeta_j} \right)^2 \frac{n_i \ 1 \ 2}{\zeta_i^2} \quad \frac{n_i \ n_j \ 1 \ 2}{(\zeta_i \ \zeta_j)^2} \right] \quad (53)$$

Second derivatives, with respect to the  $b_i$  integral, are obtained as:

$$\frac{\partial^2 b_i}{\partial \zeta_i^2} = \sum_{\mu, \nu} \left[ \left( \frac{n_i \ 1 \ 2}{\zeta_i} \quad \frac{2n_i \ n_\mu \ n_\nu \ 1}{2\zeta_i \ \zeta_\mu \ \zeta_\nu} \right)^2 \frac{n_i \ 1 \ 2}{\zeta_i^2} \quad \frac{2n_i \ n_\mu \ n_\nu \ 1}{(2\zeta_i \ \zeta_\mu \ \zeta_\nu)^2} \right] B_{\mu\nu}(i, m) \quad (54)$$

Substitution of these results into eq. (37) yields the Hessian matrix elements; for the diagonal part:

$$h_{ii} = x_i^4 Z_{ii} \frac{3}{4\zeta_i^2} x_i^2 \sum_j x_j^2 Z_{ij} \left[ \left( \frac{n_i \ 1 \ 2}{\zeta_i} \quad \frac{n_i \ n_j \ 1 \ 2}{\zeta_i \ \zeta_j} \right)^2 \frac{n_i \ 1 \ 2}{\zeta_i^2} \quad \frac{n_i \ n_j \ 1 \ 2}{(\zeta_i \ \zeta_j)^2} \right]$$

$$2x_i^2 \sum_{\mu, \nu} \left[ \left( \frac{n_i \ 1 \ 2}{\zeta_i} \quad \frac{2n_i \ n_\mu \ n_\nu \ 1}{2\zeta_i \ \zeta_\mu \ \zeta_\nu} \right)^2 \frac{n_i \ 1 \ 2}{\zeta_i^2} \frac{2n_i \ n_\mu \ n_\nu \ 1}{(2\zeta_i \ \zeta_\mu \ \zeta_\nu)^2} \right] B_{\mu\nu}(i, m) \quad (55)$$

whereas the expression for the off-diagonal elements is:

$$h_{ij} = \frac{\partial^2 Z_{ij}}{\partial \zeta_i \partial \zeta_j} Z_{ij} \left[ \left( \frac{n_i \ 1 \ 2}{\zeta_i} \quad \frac{n_i \ n_j \ 1 \ 2}{\zeta_i \ \zeta_j} \right) \left( \frac{n_j \ 1 \ 2}{\zeta_j} \quad \frac{n_i \ n_j \ 1 \ 2}{\zeta_i \ \zeta_j} \right) \frac{n_i \ n_j \ 1 \ 2}{(\zeta_i \ \zeta_j)^2} \right] \quad (56)$$

---

## Computational Scheme

A program called GATOMIC<sup>22</sup> has been constructed to compute fitted atomic shells using *nS-type* Gaussian functions. Figure 1 describes the main steps to obtain the coefficients and exponents of the ASA functions.

The initial  $\zeta_i$  exponents are taken from an *even-tempered* geometric sequence:<sup>23</sup>

$$\zeta_i = \frac{1}{n_i} \alpha \beta^i, \quad i = 1, 2, \dots, N \quad (57)$$

where  $n_i$  is the principal quantum number of the  $i$ th spherical shell, and  $N$  describes the number of used fitting functions. To perform a global search of the exponents, a grid method is employed to explore the surface described by the *even-tempered* set generating the parameters  $\{\alpha, \beta\}$ . Then, the computational procedure described in Figure 1 is carried out over every point of this grid.

As previously indicated, in the first optimization step, the similarity matrix,  $\mathbf{Z}$ , is diagonalized and the  $\{x_i\}$  coefficient set is extracted from the eigenvector producing the minimal  $\varepsilon^{(2)}$  value. Then, using EJR, as described previously, ASA

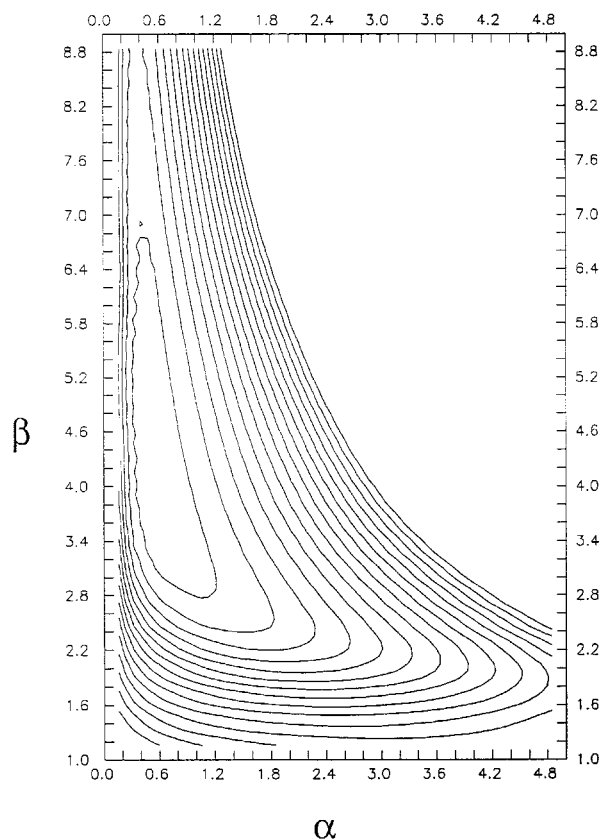
```

DATA: N = NUMBER OF FITTING FUNCTIONS
a) COMPUTE  $\zeta_i$  EXPONENTS FROM EVEN-TEMPERED SEQUENCE
   CALCULATE  $Z_{ab}$ ,  $b$  AND  $Z$ 
b) DIAGONALIZE THE MATRIX  $Z$  AND OBTAIN THE EIGENVALUES AND EIGENVECTORS ( $v$ )
   DO FOR  $j=1$  TO  $N$ 
      $x_i = v_j \wedge w_i = x_i^2, i=1, \dots, N$ 
      $\epsilon^{(2)} = \min(\epsilon^{(2)}, Z_{ab} + w^T Z w - 2b^T w)$ 
   END DO FOR  $j$ 
    $\epsilon^{(2)} = \epsilon^{(2)}$ 
c) COEFFICIENTS OPTIMIZATION USING EJR
   DO WHILE (.true.)
     DO FOR  $p=1$  TO  $N-1$ 
       DO FOR  $q=p+1$  TO  $N$ 
         COMPUTE  $E_{04}$ ,  $E_{15}$ ,  $E_{22}$  AND  $E_{11}$  FROM EQUATION (33)
         GRID SEARCH TO FIND INITIAL SQUARE SINE:  $s^2$ 
         DO WHILE2 (.true.)
           COMPUTE  $T_1$ ,  $T_2$  AND  $T_3$  FROM EQUATION (34)
            $t_1 = \frac{T_2 - \sqrt{T_2^2 + T_1 T_3}}{T_1} \rightarrow s_1, c_1 \wedge \delta \epsilon_1^{(2)} = E_{04} s_1^4 + E_{15} c_1 s_1^2 + E_{22} s_1^2 + E_{11} c_1 s_1$ 
            $t_2 = \frac{T_2 - \sqrt{T_2^2 + T_1 T_3}}{T_1} \rightarrow s_2, c_2 \wedge \delta \epsilon_2^{(2)} = E_{04} s_2^4 + E_{15} c_2 s_2^2 + E_{22} s_2^2 + E_{11} c_2 s_2$ 
            $\delta \epsilon^{(2)} = \text{Min}(\delta \epsilon_1^{(2)}, \delta \epsilon_2^{(2)}) \rightarrow s^2$ 
           IF  $|s^2 - s_1^2| < \text{TOLE}_2$  THEN
              $x_p \leftarrow -c x_p - s x_q$ 
              $x_q \leftarrow -s x_p + c x_q$ 
              $w_p = x_p^2 \wedge w_q = x_q^2$ 
           EXIT DO WHILE2
         END IF
          $s^2 = s^2$ 
       END DO WHILE2
     END DO FOR  $q$ 
   END DO FOR  $p$ 
    $\epsilon^{(2)} = Z_{ab} + w^T Z w - 2b^T w$ 
   IF  $|\epsilon^{(2)} - \epsilon^{(2)}| < \text{TOLE}_1$ , EXIT DO WHILE;
    $\epsilon^{(2)} = \epsilon^{(2)}$ 
END DO WHILE;
d) EXPONENTS OPTIMIZATION
   DO WHILE (.true.)
     CALL NEWTON SEARCH: NEEDS THE COMPUTATION OF  $g$  AND  $H$  (EQUATION (35))
     CALCULATE  $b$  AND  $Z$ 
     CALL EJR (SECCION (c))
      $\epsilon^{(2)} = Z_{ab} + w^T Z w - 2b^T w$ 
     IF  $|\epsilon^{(2)} - \epsilon^{(2)}| < \text{TOLE}$  EXIT DO WHILE
      $\epsilon^{(2)} = \epsilon^{(2)}$ 
   END DO WHILE
    
```

**FIGURE 1.** Computational flowchart describing the optimization of ASA coefficients and exponents.

coefficients are optimized solving the second degree equation, appearing in expression (34), using an iterative procedure (section c in Fig. 1). Next, the optimization of  $\zeta_i$  exponents begins. After the Newton algorithm a new EJR process is again carried out to obtain the most accurate fitted positive coefficients, and an iterative process is repeated until the convergence criterion is fulfilled.

To produce an example to depict the variation of  $\{\alpha, \beta\}$  even-tempered parameters, Figure 2 shows the surface obtained for the chlorine atom when five 15 Gaussian functions are used ( $N = 5$ ). A 30-point grid for each parameter  $\alpha$  and  $\beta$  is per-



**FIGURE 2.** Surface described by the variation of  $\alpha$  and  $\beta$  even-tempered parameters for the chlorine atom.

formed to explore the surface in the interval  $\alpha \in [0.01, 5.0]$  and  $\beta \in [1.0, 9.0]$ . Over every point of the grid, EJR optimization of the  $\{x_i\}$  coefficients is performed, keeping the ASA exponents invariable. The minimum value of  $\epsilon^{(2)}$  is 0.0143 a.u. and has been obtained at  $\alpha = 0.1734$  and  $\beta = 5.1379$  parameter values. As will be seen in the next section, this  $\epsilon^{(2)}$  value may be improved if exponents are optimized and the exponent even-tempered sequence abandoned.

## Results

Two types of results will be presented. First, to evaluate the fitting procedure, atomic QS-SM are computed using ASA functions and are compared with the *ab initio* ones. A following analysis will verify the behavior of the ASA atomic basis set density functions in molecular calculations, using the *promolecular* approximation, as described in eq. (11).

## ATOMIC CALCULATIONS

Table I lists *ab initio* QS-SM ( $Z_{aa}$ ) using the 3-21G basis set<sup>24</sup> for atoms Li to Kr. HF *ab initio* atomic density functions have been calculated with the ATOMIC-95 program.<sup>25</sup> Any GTO basis set could have been chosen. The present calculation uses the 3-12G level because it is available in the Gaussian-94 program<sup>26</sup> from H to Kr. Systematic calculations on other basis sets are under study.

Different 1S Gaussian basis sets have been computed with the GATOMIC program, which uses as the input data the output data of the ATOMIC

program. Table I also gives the quadratic error integral function and the relative error produced in the QS-SM, using different numbers of functions per atom. Table I does not show results for heavy atoms using three and four functions (corresponding to columns  $N = 3$  and  $N = 4$ ), because it is not possible to achieve a good fitting with fewer functions. Hydrogen and helium atoms are not included in this table because their contributions to MQSM are negligible (see the QSM map HCCH—Ne in the next subsection), and only one function is used to describe these two atoms. Coeffi-

**TABLE I.**  
Results Using a 3-21G Basis Set<sup>24</sup> for Li to Kr.<sup>a</sup>

Atom (electronic state)	$Z_{aa}$ (a.u.)	$N$	$\varepsilon^{(2)}$	% $Z_{aa}$ error	$N$	$\varepsilon^{(2)}$	% $Z_{aa}$ error	$N$	$\varepsilon^{(2)}$	% $Z_{aa}$ error
Li( <sup>2</sup> S)	0.34440	5	0.0000023	0.0218	4	0.000016	0.0111	3	0.00085	0.0316
Be( <sup>1</sup> S)	0.51914	5	0.0000064	0.0330	4	0.000009	0.0095	3	0.00111	0.0775
B( <sup>2</sup> P)	0.69161	5	0.0000075	0.0137	4	0.000027	0.0641	3	0.00141	0.0035
C( <sup>3</sup> P)	0.87083	5	0.0000174	0.0021	4	0.000074	0.1214	3	0.00181	0.0904
N( <sup>4</sup> S)	1.05761	5	0.0000179	0.0023	4	0.000164	0.1938	3	0.00239	0.1991
O( <sup>3</sup> P)	1.25220	5	0.0000195	0.0179	4	0.000345	0.2861	3	0.00316	0.3427
F( <sup>2</sup> P)	1.45639	5	0.0000185	0.0079	4	0.000602	0.3855	3	0.00413	0.5049
Ne( <sup>1</sup> S)	1.67148	5	0.0000245	0.0133	4	0.001046	0.5552	3	0.00532	0.6790
Na( <sup>2</sup> S)	1.90067	5	0.0001318	0.0922	4	0.002114	0.3195	3	0.00724	1.3363
Mg( <sup>1</sup> S)	2.14464	5	0.0002249	0.0351	4	0.002733	0.0990	3	0.01172	2.0284
Al( <sup>2</sup> P)	2.38804	5	0.0002282	0.0154	4	0.003055	0.0347	3	0.01888	2.6856
Si( <sup>3</sup> P)	2.63649	5	0.0002436	0.0038	4	0.003159	0.0641	3	0.02900	3.2867
P( <sup>4</sup> S)	2.88964	5	0.0001657	0.0128	4	0.003197	0.0672	3	0.04829	2.8037
S( <sup>3</sup> P)	3.14729	5	0.0001382	0.0148	4	0.003237	0.0659	3	0.07193	1.9657
Cl( <sup>2</sup> P)	3.40948	5	0.0001097	0.0167	4	0.003261	0.0566	3	0.07578	1.9987
Ar( <sup>1</sup> S)	3.67627	5	0.0000906	0.0186	4	0.003282	0.0516	3	0.07941	1.9736
K( <sup>2</sup> S)	3.94005	5	0.0001216	0.0430						
Ca( <sup>1</sup> S)	4.21566	5	0.0002768	0.0808						
Sc( <sup>2</sup> D)	4.48734	5	0.0003525	0.0933						
Ti( <sup>3</sup> F)	4.76039	5	0.0004101	0.1030						
V( <sup>4</sup> F)	5.03508	5	0.0004908	0.1182						
Cr( <sup>5</sup> D)	5.30329	5	0.0005757	0.1316						
Mn( <sup>6</sup> S)	5.59114	5	0.0006672	0.1438						
Fe( <sup>5</sup> D)	5.87170	5	0.0007898	0.1604						
Co( <sup>4</sup> F)	6.15446	5	0.0009223	0.1717						
Ni( <sup>3</sup> F)	6.44012	5	0.0010698	0.1948						
Cu( <sup>2</sup> D)	6.72772	5	0.0012540	0.2118						
Zn( <sup>1</sup> S)	7.01766	5	0.0014076	0.2365						
Ga( <sup>2</sup> P)	7.31617	5	0.0022071	0.2992						
Ge( <sup>3</sup> P)	7.61659	5	0.0034518	0.3805						
As( <sup>4</sup> S)	7.97677	5	0.0043724	0.4565						
Se( <sup>3</sup> P)	8.22790	5	0.0060097	0.5303						
Br( <sup>2</sup> P)	8.55327	5	0.0080829	0.6061						
Kr( <sup>1</sup> S)	8.84739	5	0.0106579	0.6765						

<sup>a</sup>*Ab initio* QS-SM ( $Z_{aa}$ ), quadratic error integral function and relative error in QS-SM using different numbers ( $N$ ) of 1S Gaussian functions per atom.

cients and exponents for all these basis sets of ASA functions are available for downloading at a WWW site.<sup>27</sup>

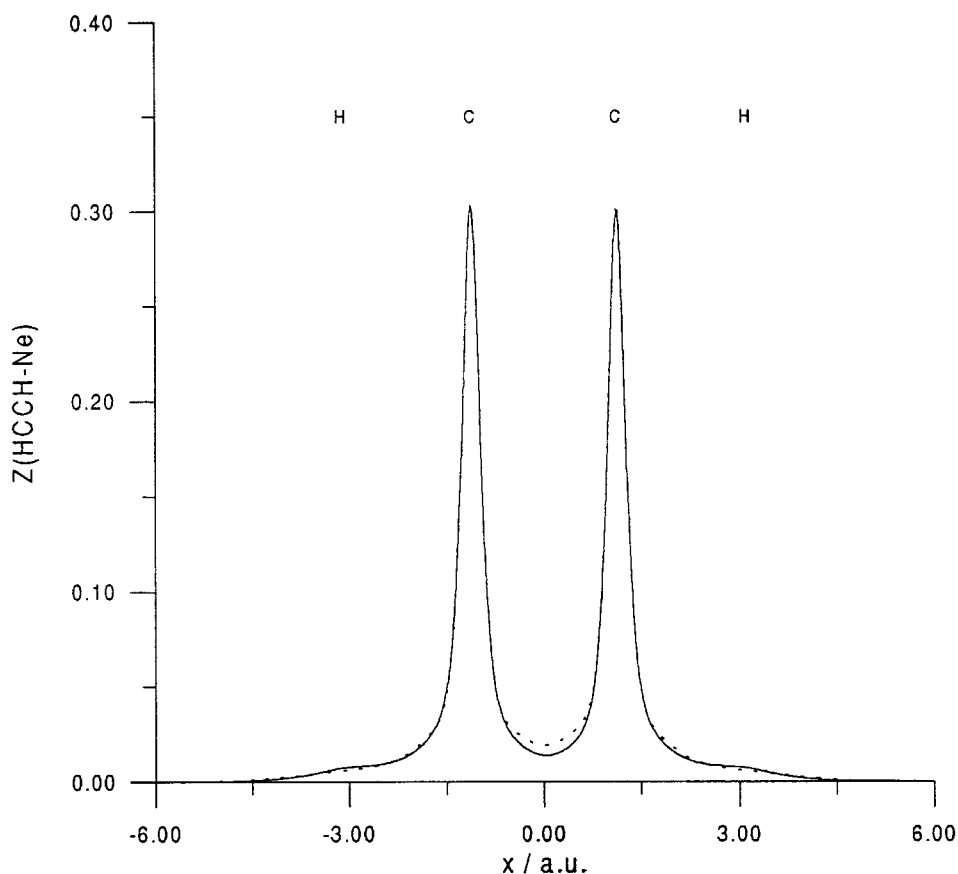
### SOME MOLECULAR EXAMPLES

QSM maps<sup>28</sup> constitute a very useful tool for observing the behavior of the ASA functions with respect to the *ab initio* ones.<sup>11b</sup> A QSM map between a molecule, *A*, and an atom, *b*, can be defined as the integral:

$$Z_{Ab}(\mathbf{R}) = \iint \rho_A(\mathbf{r}_1) \Omega(\mathbf{r}_1, \mathbf{r}_2) \rho_b(\mathbf{r}_2 - \mathbf{R}) d\mathbf{r}_1 d\mathbf{r}_2 \quad (58)$$

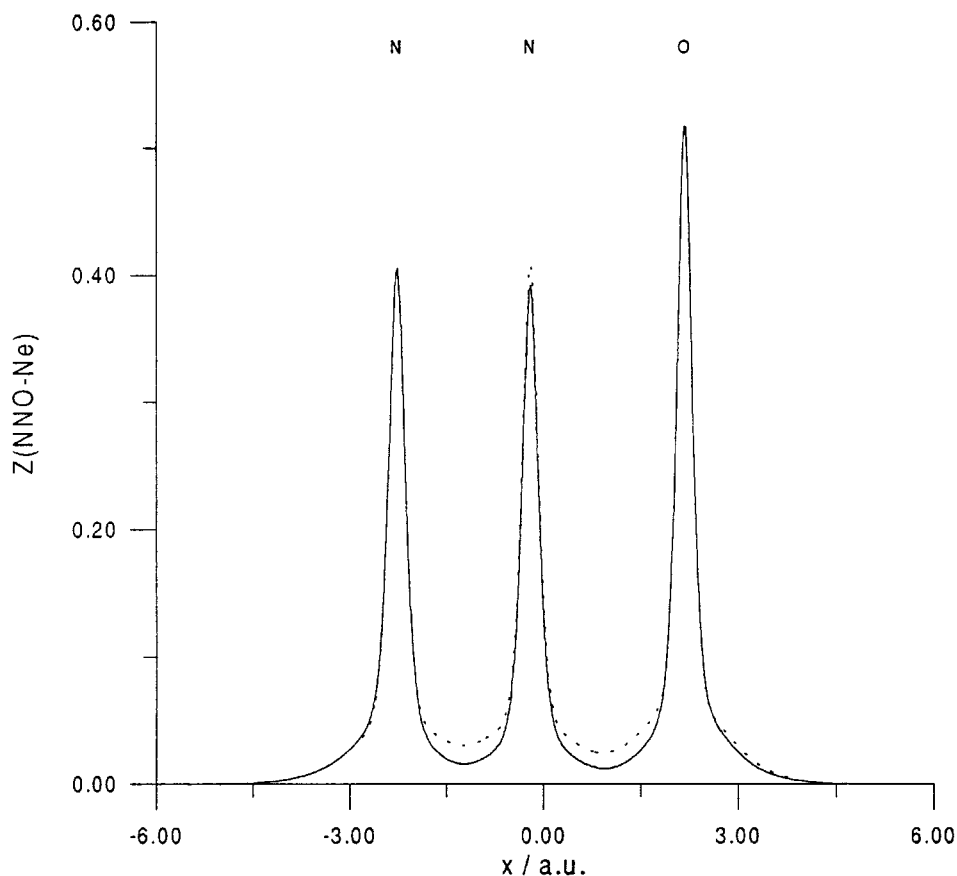
where it is made explicit that the measure  $Z_{Ab}(\mathbf{R})$  depends on the atom position  $\mathbf{R}$ . Two examples are presented of QSM maps using overlap-like measures; both are obtained when a neon atom is moved along the molecular axis defined by the

atoms of a linear molecule. Density functions and molecular geometries are obtained using the Gaussian-94 program and the HF/3-21G basis set. In the first example, a resulting QSM map for the acetylene molecule is shown. In Figure 3, the solid line shows the *ab initio* QSM map, whereas the dashed line corresponds to ASA computations, using one function for H and three functions for C and Ne atoms (corresponding to column *N* = 3 of Table I). Maximal error between *ab initio* and ASA measures is 0.006 a.u. and is located in the C—C bonding region. The second example, presented in Figure 4, consists of a QSM map of nitrous oxide, NNO. In the same manner as the above example, *ab initio* calculations are drawn in solid line and ASA measures, using three fitted functions for N, O, and Ne atoms (dashed). Bonding regions N—N and N—O are the areas where ASA functions produce maximal error with a maximum value of 0.016 a.u. This effect is due to the *promolecular* approximation itself, which obviously produces a



**FIGURE 3.** QSM map for the system acetylene–neon. Solid line corresponds to *ab initio* HF/3-21G calculation. Dashed line is the ASA calculation.





**FIGURE 4.** QSM map for the system nitrous oxide–neon. Solid line corresponds to *ab initio* HF/3-21G calculation. Dashed line is the ASA calculation.

less accurate density representation in the bonding regions of molecules, but correctly describes atom nuclei. As is evidenced in the following examples, bonding regions do not have so great an influence in MQSM, and ASA results may be considered as possessing sufficient accuracy for the practical purposes of QSM.

The optimized ASA functions have been ap-

plied to a molecular set consisting of nine fluoro- and chlorosubstituted methanes. Tables II and III show *ab initio* MQSM and Carbó indices<sup>14a</sup> for all pairs of molecules. Optimized geometries have been obtained with the Gaussian-94 program using the HF 3-21G 6D 10F level of theory, and optimal molecular superpositions have been found using the method described in ref. 18. This molec-

**TABLE II.**  
*Ab initio* MQSM for Fluoro- and Chlorosubstituted Methanes Using HF/3-21G Densities.

	CH <sub>4</sub>	CFH <sub>3</sub>	CF <sub>2</sub> H <sub>2</sub>	CF <sub>3</sub> H	CF <sub>4</sub>	CClH <sub>3</sub>	CCl <sub>2</sub> H <sub>2</sub>	CCl <sub>3</sub> H	CCl <sub>4</sub>
CH <sub>4</sub>	0.31481	0.32213	0.22321	0.17084	0.13841	0.54534	0.33762	0.24449	0.19162
CFH <sub>3</sub>		0.46106	0.31639	0.23657	0.18794	0.66778	0.41364	0.29967	0.23495
CF <sub>2</sub> H <sub>2</sub>			0.39548	0.27592	0.20845	0.46638	0.28836	0.20889	0.16388
CF <sub>3</sub> H				0.33334	0.26499	0.35777	0.22114	0.16046	0.12593
CF <sub>4</sub>					0.28533	0.29006	0.17914	0.13001	0.10246
CClH <sub>3</sub>						1.50383	0.92818	0.67035	0.52526
CCl <sub>2</sub> H <sub>2</sub>							1.13480	0.71621	0.45926
CCl <sub>3</sub> H								0.88792	0.63259
CCl <sub>4</sub>									0.72536

**TABLE III.**  
***Ab initio* Carbó Indices for Fluoro- and Chlorosubstituted Methanes Using HF / 3-21G Densities.**

	CH <sub>4</sub>	CFH <sub>3</sub>	CF <sub>2</sub> H <sub>2</sub>	CF <sub>3</sub> H	CF <sub>4</sub>	CClH <sub>3</sub>	CCl <sub>2</sub> H <sub>2</sub>	CCl <sub>3</sub> H	CCl <sub>4</sub>
CH <sub>4</sub>	1	0.84553	0.63260	0.52739	0.46181	0.79258	0.56486	0.46243	0.40100
CFH <sub>3</sub>		1	0.74094	0.60345	0.51815	0.80196	0.57185	0.46836	0.40628
CF <sub>2</sub> H <sub>2</sub>			1	0.75994	0.62054	0.60476	0.43044	0.35251	0.30597
CF <sub>3</sub> H				1	0.85926	0.50531	0.35956	0.29494	0.25610
CF <sub>4</sub>					1	0.44281	0.31482	0.25829	0.22522
CClH <sub>3</sub>						1	0.71051	0.58011	0.50292
CCl <sub>2</sub> H <sub>2</sub>							1	0.71350	0.50620
CCl <sub>3</sub> H								1	0.78824
CCl <sub>4</sub>									1

ular set has been used in previous studies.<sup>11,14f</sup> as a benchmark test to verify new fitting algorithms and approximations.

Table IV resumes the results obtained using different sets of ASA functions described in Table I. Main error scores for the molecular set studied are presented in Table IV. The data provided in the table consist of the following parameters: greatest error, error arithmetic mean, and standard deviation produced when relative error for all pairs of molecules is analyzed. As was expected, the accuracy achieved increases when the number of ASA functions per atom are augmented. Another aspect that arises from Table IV corresponds to the fact that ASA Carbó indices agree with *ab initio* values, in a better way than MQSM. This characteristic has already been observed in previous studies,<sup>11, 14f</sup> and may be an effect caused by the normalization of the measure, performed when Carbó indices are computed.

The present results are the best obtained in our laboratory using a *promolecular* approximation up to now. This is shown when they are compared

with those computed in ref. 11b, where 1S function exponents were fitted to reproduce atomic QS-SM. Another evaluation may be obtained when the present results of *promolecular* ASA are compared with those of molecular ASA.<sup>11a</sup> In ref. 11a, the set of nine fluoro- and chlorosubstituted methanes have also been studied, although at a different level of theory (HF 6-31G\*). This last remark does not become an inconvenience at all, if ref. 11a error and standard deviation means are compared with those of Table IV. In the molecular ASA computations found in ref. 11a, the percent error arithmetic mean between *ab initio* and fitted MQSM is 0.056%, with a standard deviation of 0.05%. In contrast, within Carbó index calculations of the same source, the relative error arithmetic mean is 0.066% and the standard deviation becomes 0.06%. The accuracy characteristics of these quoted results are comparable with the five-function-level *promolecular* measures of the present work (see Table IV) and the effect is more pronounced within the Carbó index figures.

To show an example of the ASA results, in

**TABLE IV.**  
**Relative Errors with Associated Greatest Error Value, Arithmetic Mean and Standard Deviation Produced by Comparing *Ab Initio* Values (Tables II and III) with ASA MQSM and Carbó Indices for Fluoro- and Chlorosubstituted Methanes.**

H	Number of shells			Measure	Relative errors			
	C	F	Cl		Greatest error (%) / pair of molecules		Mean of errors	Standard deviation
1	3	3	4	Z <sub>AB</sub>	1.421	(CH <sub>4</sub> — CH <sub>4</sub> )	0.534	0.380
				C <sub>AB</sub>	1.282	(CFH <sub>3</sub> — CClH <sub>3</sub> )	0.576	0.537
1	4	4	4	Z <sub>AB</sub>	1.432	(CH <sub>4</sub> — CH <sub>4</sub> )	0.350	0.355
				C <sub>AB</sub>	0.521	(CFH <sub>3</sub> — CClH <sub>3</sub> )	0.173	0.189
1	5	5	5	Z <sub>AB</sub>	1.277	(CH <sub>4</sub> — CH <sub>4</sub> )	0.294	0.320
				C <sub>AB</sub>	0.244	(CH <sub>4</sub> — CCl <sub>4</sub> )	0.052	0.071

**TABLE V.**  
**MQSM for Fluor- and Chlorosubstituted Methanes Using Five-Function Level ASA Fitted Densities.**

	CH <sub>4</sub>	CFH <sub>3</sub>	CF <sub>2</sub> H <sub>2</sub>	CF <sub>3</sub> H	CF <sub>4</sub>	CClH <sub>3</sub>	CCl <sub>2</sub> H <sub>2</sub>	CCl <sub>3</sub> H	CCl <sub>4</sub>
CH <sub>4</sub>	0.31883	0.32487	0.22503	0.17217	0.13945	0.55023	0.34065	0.24669	0.19335
CFH <sub>3</sub>		0.46273	0.31742	0.23708	0.18803	0.66842	0.41406	0.30002	0.23527
CF <sub>2</sub> H <sub>2</sub>			0.39648	0.27668	0.20896	0.46703	0.28873	0.20916	0.16412
CF <sub>3</sub> H				0.33405	0.26549	0.35822	0.22139	0.16070	0.12613
CF <sub>4</sub>					0.28590	0.29041	0.17934	0.13019	0.10265
CClH <sub>3</sub>						1.50494	0.92893	0.67097	0.52583
CCl <sub>2</sub> H <sub>2</sub>							1.13538	0.71670	0.45975
CCl <sub>3</sub> H								0.88831	0.63296
CCl <sub>4</sub>									0.72567

Tables V and VI, MQSM and Carbó indices of the set of fluoro- and chlorosubstituted methanes are presented in five-function-level calculation.

To test the behavior of the present ASA functions in calculations involving large molecules with heavy atoms, a sample of eight molecules has been chosen. These molecular structures have been taken from the Cambridge Structural Database (CSD)<sup>29</sup> and are shown in Table VII.

Table VIII lists the results of molecular QS-SM using *ab initio* and *promolecular* ASA density functions. From the CSD geometries, molecular electronic density functions have been calculated with the Gaussian-94 program using the HF 3-21G 6D 10F level. ASA QS-SM were computed from the atomic basis presented in Table I. The rule was to use one function for the hydrogen atom, three functions for all atoms belonging to the second period, four functions for the atoms of the third period, and five functions for the atoms of the fourth period. This procedure gives the best results with the minimum number of atomic ASA functions.

The results given in Table VIII can be considered satisfactory, in view of the fact that the relative errors made in fitted QS-SM are less than 0.81% for all the molecules studied. *Promolecular* functions produce an electron density excess around the zone of the bonding regions, as is shown in Figures 3 and 4. This less accurate density representation can also be evidenced by noticing that fitted QS-SM in Table VIII are *always* greater than the *ab initio* exact values.

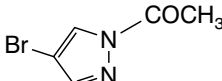
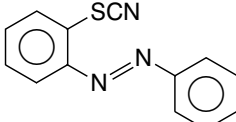
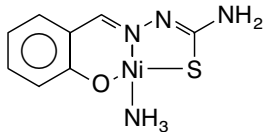
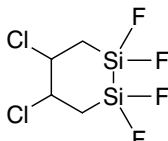
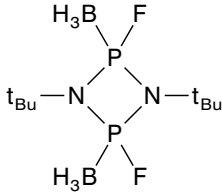
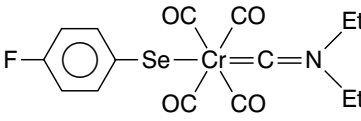
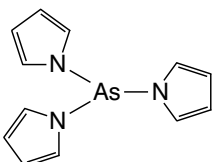
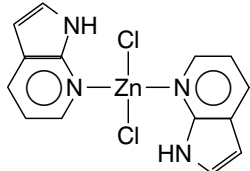
## Conclusions

The most important idea to be extracted from the present results, is that the EJR technique can be a useful methodological procedure that allows us to obtain constrained fitted density functions with positive coefficients, and in this manner preserving the statistical meaning of the fitted density functions. Also, the optimization of Gaussian scale factors, using a Newton method, gives even more accurate ASA functions. Within these techniques,

**TABLE VI.**  
**Carbó Indices for Fluoro- and Chlorosubstituted Methanes Using Five-Function Level ASA Fitted Densities.**

	CH <sub>4</sub>	CFH <sub>3</sub>	CF <sub>2</sub> H <sub>2</sub>	CF <sub>3</sub> H	CF <sub>4</sub>	CClH <sub>3</sub>	CCl <sub>2</sub> H <sub>2</sub>	CCl <sub>3</sub> H	CCl <sub>4</sub>
CH <sub>4</sub>	1	0.84579	0.63291	0.52758	0.46188	0.79434	0.56619	0.46355	0.40198
CFH <sub>3</sub>		1	0.74107	0.60301	0.51696	0.80098	0.57125	0.46795	0.40600
CF <sub>2</sub> H <sub>2</sub>			1	0.76025	0.62064	0.60462	0.43034	0.35245	0.30596
CF <sub>3</sub> H				1	0.85909	0.50523	0.35949	0.29500	0.25619
CF <sub>4</sub>					1	0.44274	0.31477	0.25835	0.22537
CClH <sub>3</sub>						1	0.71064	0.58031	0.50317
CCl <sub>2</sub> H <sub>2</sub>							1	0.71364	0.50651
CCl <sub>3</sub> H								1	0.78836
CCl <sub>4</sub>									1

**TABLE VII.**  
**Cambridge Structural Database Reference Codes, Chemical Formulae, and Structures of Selected Molecules Used to Compute QS-SM.**

Molecule	CSD ref. code <sup>29</sup>	Formula	Structure
A <sub>1</sub>	ABPZOL10	C <sub>5</sub> H <sub>5</sub> BrN <sub>2</sub> O	
A <sub>2</sub>	ABSFCN	C <sub>13</sub> H <sub>9</sub> N <sub>3</sub> S	
A <sub>3</sub>	AMSCNI11	C <sub>8</sub> H <sub>10</sub> N <sub>4</sub> NiOS	
A <sub>4</sub>	BABPAV	C <sub>4</sub> H <sub>6</sub> Cl <sub>2</sub> F <sub>4</sub> Si <sub>2</sub>	
A <sub>5</sub>	BAKNAC	C <sub>8</sub> H <sub>24</sub> B <sub>2</sub> F <sub>2</sub> N <sub>2</sub> P <sub>2</sub>	
A <sub>6</sub>	BAMCOH	C <sub>15</sub> H <sub>14</sub> CrFNO <sub>4</sub> Se	
A <sub>7</sub>	BEYCIR	C <sub>12</sub> H <sub>12</sub> AsN <sub>3</sub>	
A <sub>8</sub>	BIFXAP	C <sub>14</sub> H <sub>12</sub> Cl <sub>2</sub> N <sub>4</sub> Zn	

**TABLE VIII.**  
**QS-SM ( $Z_{AA}$ ) Using HF/3-21G and ASA Densities**  
**and Relative Error for the Molecular Set in Table VII.**

	$Z_{AA}^{\text{HF}}$ (a.u.)	$N^a$	$Z_{AA}^{\text{ASA}}$ (a.u.)	% $Z_{AA}$
A <sub>1</sub>	1.27877	34	1.28601	0.5659
A <sub>2</sub>	0.08890	61	0.08932	0.4733
A <sub>3</sub>	0.33386	58	0.33646	0.7776
A <sub>4</sub>	0.21973	46	0.22001	0.1278
A <sub>5</sub>	0.09287	74	0.09317	0.3287
A <sub>6</sub>	0.30662	87	0.30817	0.5040
A <sub>7</sub>	0.48412	62	0.48642	0.4734
A <sub>8</sub>	0.25129	79	0.25331	0.8025

<sup>a</sup> $N$  is the number of shells per molecule (see text for number of shells per atom).

function exponents and coefficients are completely optimized. In this context it has been demonstrated that an excellent ASA Gaussian spherical basis set can be obtained.

Moreover, the present work shows how *pro-molecular* electronic densities describe molecular densities with sufficient accuracy for some MQSM simplified computational purposes. In subsequent MQSM studies, it will only be necessary to know the molecular coordinates, and then, with a set of ASA functions, as parameterized in this work, the molecular electronic density distribution can be built automatically without further effort.

## Acknowledgments

The authors thank the referees for their advice on improving the contents and structure of the present work.

## References

- (a) K. J. Miller and K. Ruedenberg, *J. Chem. Phys.*, **48**, 3414 (1968); (b) R. C. Raffanetti and K. Ruedenberg, *Int. J. Quant. Chem.*, **35**, 625 (1970); (c) D. K. Hoffman, R. C. Raffanetti and K. Ruedenberg, *J. Math. Phys.*, **13**, 528 (1972).
- (a) C. Edmiston and K. Ruedenberg, *Rev. Mod. Phys.*, **35**, 457 (1963); (b) R. C. Raffanetti, K. Ruedenberg, C. L. Janssen and H. F. Schaefer, *Theor. Chim. Acta*, **86**, 149 (1993).
- D. M. Silver, E. L. Mehler and K. Ruedenberg, *J. Chem. Phys.*, **52**, 1206 (1970).
- (a) R. Carbó, L. Domingo and J. J. Peris, *Adv. in Quant. Chem.*, **15**, 215 (1982); (b) R. Carbó, J. Miró, L. Domingo and J. J. Novoa, *Adv. in Quant. Chem.*, **20**, 375 (1989); (c) R. Carbó, L. Domingo, J. J. Peris and J. J. Novoa, *J. Mol. Struct.*, **93**, 15 (1983); (d) R. Carbó and B. Calabuig, *Comput. Phys. Commun.*, **52**, 345 (1989).
- R. Carbó and L. Domingo, *Int. J. Quant. Chem.*, **23**, 517 (1987).
- C. G. J. Jacobi, *J. Reine Angew. Math.*, **30**, 51 (1846).
- R. Carbó, L. Molino and B. Calabuig, *J. Comput. Chem.*, **13**, 155 (1992).
- R. Carbó-Dorca and E. Besalú, *J. Math. Chem.*, **20**, 263 (1997).
- J. Mestres, M. Solà, M. Duran and R. Carbó, *J. Comput. Chem.*, **15**, 1113 (1994).
- J. Cioslowski, P. Piskorz and P. Rez, *J. Chem. Phys.*, **106**, 3607 (1997).
- (a) P. Constans and R. Carbó, *J. Chem. Inf. Comput. Sci.*, **35**, 1046 (1995); (b) P. Constans, L. Amat, X. Fradera and R. Carbó-Dorca, In R. Carbó-Dorca and P. G. Mezey (Eds.), *Advances in Molecular Similarity*, Vol. 1, JAI Press, Inc., Greenwich, CT, 1996, p. 187.
- J. von Neumann, *Mathematical Foundations of Quantum Mechanics*, Princeton University Press, Princeton, N.J., 1955.
- See, for example: (a) M. A. Johnson and G. Maggiora, Eds., *Concepts and Applications of Molecular Similarity*, Wiley, New York, 1990; (b) R. Carbó, Ed., *Molecular Similarity and Reactivity: From Quantum Chemical to Phenomenological Approaches*, Kluwer Amsterdam, 1995; (c) R. Carbó-Dorca, and P. G. Mezey, Eds., *Advances in Molecular Similarity*, Vol. 1, JAI Press, Greenwich, CT, 1996.
- See, for example: (a) R. Carbó, M. Arnau and L. Leyda, *Int. J. Quantum Chem.*, **17**, 1185 (1980); (b) R. Carbó and B. Calabuig, *Comput. Phys. Commun.*, **55**, 117 (1989); (c) R. Carbó and B. Calabuig, *Int. J. Quantum Chem.*, **42**, 1681 (1992); (d) R. Carbó, E. Besalú and L. Vera, *Adv. in Quantum Chem.*, **25**, 253 (1994); (e) M. Solà, J. Mestres, R. Carbó and M. Duran, *J. Am. Chem. Soc.*, **116**, 5909 (1994); (f) E. Besalú, R. Carbó, J. Mestres and M. Solà, *Topics Curr. Chem.*, **173**, 31 (1995); (g) J. Mestres, M. Solà, R. Carbó, F. J. Luque and M. Orozco, *J. Phys. Chem.*, **100**, 606 (1996).
- See, for example: (a) D. L. Cooper and N. L. Allan, *J. Chem. Soc. Faraday Trans.*, **83**, 449 (1987); (b) D. L. Cooper and N. L. Allan, *J. Am. Chem. Soc.*, **114**, 4773 (1992); (c) J. Cioslowski and E. D. Fleischmann, *J. Am. Chem. Soc.*, **113**, 64 (1991); (d) J. V. Ortiz and J. Cioslowski, *Chem. Phys. Lett.*, **185**, 270 (1991); (e) R. Ponec and M. Strnad, *J. Phys. Org. Chem.*, **4**, 701 (1991); (f) R. Ponec and M. Strnad, *Int. J. Quant. Chem.*, **42**, 501 (1992); (g) E. E. Hodgkin and W. G. Richards, *Int. J. Quant. Chem.*, **14**, 105 (1987); (h) P. G. Mezey, *J. Comp. Chem.*, **8**, 462 (1987).
- (a) K. Ruedenberg and W. H. E. Schwarz, *J. Chem. Phys.*, **92**, 4956 (1990); (b) P. Coppens, In *International Tables for Crystallography*, Vol. B, Kluwer, Amsterdam, 1992, p. 10; (c) P. Coppens and J. Becker, In *International Tables for Crystallography*, Vol. C, Kluwer, Amsterdam, 1992, p. 628.
- X. Fradera, L. Amat, E. Besalú and R. Carbó-Dorca, *Quant. Struct.-Activ. Relat.*, **16**, 25 (1977).
- (a) P. Constans, L. Amat and R. Carbó-Dorca, *J. Comput. Chem.*, **18**, 826 (1997); (b) L. Amat, P. Constans and R. Carbó, *Sci. Gerund*, **22**, 109 (1996).
- J. A. Pople and D. L. Beveridge, *Approximate Molecular Orbital Theory*, McGraw-Hill, New York, 1970.
- See, for example: D. A. Pierre, *Optimization Theory with Applications*. Wiley, New York, 1969.
- See, for example: (a) H. Kubinyi, Ed., *3D QSAR in Drug Design: Theory Methods and Applications*, ESCOM, Leiden,

- 1993; (b) P. M. Dean, Ed. *Molecular Similarity in Drug Design*, Blackie, London, 1995; (c) A. C. Good, S. S. So and W. G. Richards, *J. Med. Chem.*, **36**, 433 (1993).
22. L. Amat and R. Carbó-Dorca, *GATOMIC Program*, Institute of Computational Chemistry, University of Girona, Girona, Spain, 1997.
23. S. Huzinaga and M. Klobukowski, *J. Mol. Struct. (Theorchem)*, **167**, 1 (1988).
24. (a) J. S. Binkley, J. A. Pople and W. J. Hehre, *J. Am. Chem. Soc.*, **102**, 939 (1980); (b) M. S. Gordon, J. S. Binkley, J. A. Pople, W. J. Pietro and W. J. Hehre, *J. Am. Chem. Soc.*, **104**, 2797 (1982); (c) W. J. Pietro, M. M. Francl, W. J. Hehre, D. J. DeFrees, J. A. Pople and J. S. Binkley, *J. Am. Chem. Soc.*, **104**, 5039 (1982).
25. *ATOMIC Program 1995*, by R. Carbó-Dorca, from *A General Program for Calculation of SCF Orbitals by the Expansion Method*, B. Roos, C. Salez, A. Veillard and E. Clementi, IBM Research RJ518(#10901), 1968.
26. M. J. Frisch, G. W. Trucks, H. B. Schlegel, P. M. W. Gill, B. G. Johnson, M. A. Robb, J. R. Cheeseman, T. A. Keith, G. A. Petersson, J. A. Montgomery, K. Raghavachari, Al-Laham, M. A., V. G. Zakrzewski, J. V. Ortiz, J. B. Foresman, J. Cioslowski, B. B. Stefanov, A. Nanaykkara, M. Challacombe, C. Y. Peng, P. Y. Ayala, W. Chen, M. W. Wong, J. L. Andres, E. S. Replogle, R. Comperts, R. L. Martin, D. J. Fox, H. S. Binkley, D. J. Defrees, H. Baker, J. J. P. Stewart, M. Head-Gordon, C. Gonzalez and J. A. Pople, *Gaussian-94*, (Revision A.1), Gaussian, Inc., Pittsburgh, PA, 1995.
27. ASA coefficients and exponents can be seen and downloaded from the WWWsite: [http://iqc.udg.es/cat/similarity/ASA\\_funcset.html](http://iqc.udg.es/cat/similarity/ASA_funcset.html)
28. (a) R. Carbó-Dorca, E. Besalú, L. Amat and X. Fradera, In R. Carbó-Dorca and P. G. Mezey, Eds., *Advances in Molecular Similarity*, Vol. 1, JAI Press, Greenwich, CT, 1996, p. 1; (b) L. Amat, X. Fradera and R. Carbó, *Sci. Gerund.*, **22**, 97 (1996).
29. F. H. Allen and O. Kennard, *Chem. Design Automat. News*, **8**, 31 (1993).



### 3.10 Milliores en l'algorisme d'optimització dels coeficients ASA

Un dels processos que s'executa repetidament en l'ajust de les funcions ASA és el càlcul del sinus *EJR* descrit en la taula 3.2. Amb l'objectiu d'accelerar el procés global s'ha simplificat l'expressió de la variació de l'error quadràtic (3.67) substituint el sinus i el cosinus pels seus corresponents desenvolupaments en sèries de Taylor:

$$\sin(\mathbf{a}) = \sum_{k=0}^{\infty} \frac{(-1)^k \mathbf{a}^{2k+1}}{(2k+1)!} = \mathbf{a} - \frac{\mathbf{a}^3}{3!} + \frac{\mathbf{a}^5}{5!} - \frac{\mathbf{a}^7}{7!} + \dots \quad (3.76)$$

i

$$\cos(\mathbf{a}) = \sum_{k=0}^{\infty} \frac{(-1)^k \mathbf{a}^{2k}}{(2k)!} = 1 - \frac{\mathbf{a}^2}{2!} + \frac{\mathbf{a}^4}{4!} - \frac{\mathbf{a}^6}{6!} + \dots \quad (3.77)$$

Truncant les sèries en els termes d'ordre tres,

$$\begin{aligned} c = \cos(\mathbf{a}) &= 1 - \frac{1}{2}\mathbf{a}^2 + \mathbf{q}(\mathbf{a}^3) \\ s = \sin(\mathbf{a}) &= \mathbf{a} \left( 1 - \frac{1}{6}\mathbf{a}^2 \right) + \mathbf{q}(\mathbf{a}^4) \end{aligned} \quad (3.78)$$

i considerant nul·les les potències de l'angle  $\mathbf{a}$  superiors a tres, es poden definir les variables:

$$s^2 \approx \mathbf{a}^2, \quad s^3 \approx \mathbf{a}^3, \quad s^4 \approx 0, \quad cs \approx \mathbf{a} \left( 1 - \frac{2}{3}\mathbf{a}^2 \right), \quad cs^3 \approx \mathbf{a}^3, \quad (3.79)$$

que substituïdes en l'equació (3.67) resulta una expressió molt simple de la variació de la funció error quadràtic integral referida a l'angle de rotació  $\mathbf{a}$ :

$$d\mathbf{e}^{(2)} = \mathbf{a}^3 A + \mathbf{a}^2 B + \mathbf{a} C, \quad (3.80)$$

on  $A = E_{13} - 2E_{11}/3$ ,  $B = E_{02}$  i  $C = E_{11}$ . Aplicant la condició de mínim:

$$\frac{d\mathbf{e}^{(2)}}{d\mathbf{a}} = 3\mathbf{a}^2 A + 2\mathbf{a} B + C = 0, \quad (3.81)$$



s'obté una equació de segon grau, on la solució

$$\mathbf{a}^* = \mathbf{a}_+ = \frac{1}{3A} \left[ -B + (B^2 - 3AC)^{\frac{1}{2}} \right] \quad (3.82)$$

és la que dóna un valor positiu en la segona derivada. Finalment, el sinus òptim es calcula d'acord amb la igualtat

$$s = \mathbf{a}_+ \left( 1 - \frac{1}{6} \mathbf{a}_+^2 \right), \quad (3.83)$$

mentre que el cosinus es determina a partir del valor del sinus per així assegurar que es farà una rotació ortogonal. Els valors de  $c$  i  $s$  s'utilitzen per determinar un nou parell de valors  $\{x_p, x_q\}$  a partir de l'equació (3.63).

La simplificació de l'algorisme evita el càlcul iteratiu necessari per obtenir el sinus i el cosinus de l'angle de la rotació  $\mathbf{J}_{pq}(\alpha)$  descrit en la subrutina SINUS de la taula 3.2. El resultat ha estat la millora de l'algorisme matemàtic, accelerant-ne el temps de computació. Però l'aproximació en sèries de Taylor només és vàlida quan el procés d'optimització és proper a l'òptim.

### ***3.10.1 Esquema computacional del càlcul del sinus EJR***

En la taula 3.8 es mostra l'esquema del nou algorisme d'optimització del sinus *EJR*. Respecte al programa GATOMIC únicament suposa canviar la línia de comandes on es sol·licita la subrutina SINUS, dins l'algorisme EJR, per la subrutina SINUS/TAYLOR. Com s'ha comentat, el desenvolupament en sèries de Taylor únicament és eficaç en zones pròximes al punt extrem. És per això que en el codi es comprova si el valor calculat del sinus pertany a l'interval  $[-1,1]$ . En cas contrari s'executa la subrutina SINUS descrita en la taula 3.2.

**Subrutina SINUS/TAYLOR**

- ✓ Donat els valors  $E_{04}$ ,  $E_{13}$ ,  $E_{02}$  i  $E_{11}$
- Calcula els termes  $A = E_{13} - 2 E_{11}/3$ ,  $B = E_{02}$  i  $C = E_{11}$
- Calcula  $\mathbf{a}_+ = \left[ -B + (B^2 - 3AC)^{1/2} \right] (3A)^{-1}$
- Calcula  $s = \mathbf{a}_+ [1 - \mathbf{a}_+^2/6]$
- Si  $s \in [-1,1]$  llavors
  - Calcula  $c = \sqrt{1 - s^2}$
- En cas contrari
  - Demana **subrutina SINUS**. Necessita:  $E_{04}$ ,  $E_{13}$ ,  $E_{02}$  i  $E_{11}$ . Retorna:  $s$  i  $c$
- Fi condicional
- ❖ Retorna el valor òptim del sinus i el cosinus:  $s$  i  $c$

**Taula 3.8** Algorisme d'optimització del sinus *EJR* mitjançant sèries de Taylor

### 3.10.2 Ajust atòmic d'una base de funcions d'Huzinaga

En l'article 3.2 es descriuen les innovacions introduïdes en el programa GATOMIC, així com el càlcul d'un nou conjunt de funcions de base atòmiques. En concret s'han ajustat els coeficients i exponents ASA a un conjunt de funcions de base d'Huzinaga per a la sèrie d'àtoms H fins al Rn.<sup>45</sup> La qualitat de la nova base de funcions ASA atòmiques s'analitza amb els paràmetres  $\mathbf{e}^{(2)}$  i  $\%Z_{aa}$ , a més del percentatge d'error en el càlcul de l'energia potencial d'atracció electró-nucli. Emprant l'aproximació ASA, el potencial monoelectrònic d'un àtom es defineix com:

$$V(\mathbf{r}) = -\int \frac{\mathbf{r}_a^{ASA}(\mathbf{r})}{|\mathbf{r}|} d\mathbf{r} = -\sum_{i \in a} w_i \int \frac{1}{|\mathbf{r}|} |s_i(\mathbf{r})|^2 d\mathbf{r}. \quad (3.84)$$

Un dels objectius d'ajustar la base d'Huzinaga és disposar de funcions ASA corresponents a àtoms amb nombre atòmic superior al criptó. En l'article 3.2 es mostra un exemple del comportament de les funcions PASA en molècules on hi intervenen àtoms pesants. En el cas pràctic s'utilitzen les MQSM per determinar quin mètode de càlcul teòric reproduïx millor les geometries experimentals del compost cis-diaminediclor platí (II). En primer lloc es fa una recerca d'estructures determinades experimentalment per aquest complex en la *Cambridge Structural Database*.<sup>44</sup> Llavors, partint d'una geometria experimental del compost cis-platin es calcula la conformació

de mínima energia amb diferents mètodes teòrics. S'ha utilitzat l'aproximació HF, el funcional de la densitat B3LYP, i diferents nivells Møller-Plesset: MP2, MP3, MP4(DQ), MP4(SDQ). Tots els càlculs *ab initio* s'han fet emprant un conjunt de funcions de base que simulen l'efecte dels electrons més interns mitjançant un potencial. Amb les geometries òptimes de cada metodologia, més les estructures experimentals, es generen les densitats *PASA* i es determina la superposició molecular òptima de tots els possibles parells de complexos cis-platin. La matriu d'índexs de Carbó resultant s'il·lustra en la *Table V* de l'article 3.2. Analitzant els valors de  $C_{AB}$  es pot concloure que els diferents nivells MP donen resultats semblants entre ells, i són millors que les geometries obtingudes amb els mètodes HF i B3LYP.

### Article 3.2

---

**Autors:** Lluís Amat, Ramon Carbó-Dorca.

**Títol:** *Fitted electronic density functions from H to Rn for use in quantum similarity measures: cis-diammine-dichloroplatinum(II) complex as an application example*

**Revista:** *Journal of Computational Chemistry*

**Volum:** 20      **Pàgines, inicial:** 911    **final:** 920    **Any:** 1999

---

---

# Fitted Electronic Density Functions from H to Rn for Use in Quantum Similarity Measures: *cis*-Diammine-Dichloroplatinum(II) Complex as an Application Example

---

LLUÍS AMAT, RAMON CARBÓ-DORCA

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

*Received 7 December 1998; accepted 28 January 1999*

---

**ABSTRACT:** A consistent set of fitted electronic density functions was generated for the elements from hydrogen to radon using an algorithm based on the elementary Jacobi rotations (EJR) technique. The main distinguishing attribute of this fitting procedure is the production of approximated electronic density functions with positive definite expansion coefficients; in this way, the statistical meaning of the probability distribution is preserved. The methodology, which was fully described previously, was modified in this work to improve and accelerate the fitting procedure. This variation concerns the optimization method employed to obtain the optimal angle of the EJR, implementing an algorithm based on a Taylor series expansion. Additionally, a new *1S-Type* Gaussian basis set for atoms H to Rn is presented, that was fitted from a primitive basis set of Huzinaga. Fitted density functions facilitate theoretical calculations over large molecules and may be employed in many areas of computational chemistry, for example, in quantum similarity measures (QSM). To verify the basis set, a sound example related to QSM applications is given. This corresponds to the comparison of experimental structures obtained from X-ray determination for *cis*-diamminedichloroplatinum(II) complex with

*Correspondence to:* Prof. R. Carbó-Dorca; e-mail: director@iqc.udg.es

Contract grant sponsor: CICYT; contract grant number: SAF 96-0158

Contract grant sponsor: Fundació Maria Francisca de Roviralta

Contract grant sponsor: EC; contract grant number: ENV4-CT97-0508

optimized molecular geometries using several theoretical methods to quantify the differences between the analyzed levels of theory. © 1999 John Wiley & Sons, Inc. J Comput Chem 20: 911–920, 1999

**Keywords:** elementary Jacobi rotations (EJR); atomic shell approximation (ASA); quantum similarity measures (QSM); *cis*-diamminedichloroplatinum; promolecular density functions

## Introduction

Nowadays, precise theoretical *ab initio* studies of large molecular systems or transition metal complexes are usually limited by the number and kind of atoms involved, which is due to the fact that computational requirements clearly increase with the number of basis functions. The main *ab initio* calculation hindrance is generally located in the computation of the cumbersome four-center integrals. These integrals, which appear in the calculation of two-electron repulsion energies, as well as in quantum similarity measures (QSM),<sup>1–5</sup> can be readily evaluated if approximated electronic density functions are used instead of *ab initio*. These simplified electron density clouds, commonly constructed as linear combinations of Gaussian-type functions, are generally obtained from a fitting procedure consisting of optimizing the coefficients of the linear expansion by minimizing the quadratic error integral between *ab initio* and the approximated function. In the literature one can find different fitting algorithms<sup>6–11</sup> for first-order electronic density functions,  $\rho(\mathbf{r})$ , but not all of them take into account the conditions needed to obtain a definite positive  $\rho(\mathbf{r})$ . Recently, the atomic shell approximation (ASA)<sup>10,11</sup> was described, consisting of a linear expansion of 1S GTO functions to the  $\rho(\mathbf{r})$ , where a set of convex conditions are imposed to the expansion coefficients, yielding a fitted density function with the suitable features of a probability distribution.

In a previous article<sup>11</sup> the atomic density fitting of the 3-21G basis set for atoms H to Kr was examined in complete detail using a robust algorithm based on the elementary Jacobi rotation (EJR) technique.<sup>12</sup> This first work demonstrated that an EJR algorithm based on norm conserving orthogonal transformations provides an accurate method for obtaining fitted atomic density functions with the additional attribute that the expansion coefficients remain positive definite within the proce-

dure. Furthermore, in subsequent studies the usefulness of the fitted basis set to be applied in QSM<sup>13–18</sup> was confirmed, employing a *promolecular* method to construct the molecular electronic density functions. To further improve the atomic basis set fitting and permit QSM calculations over large atoms, a Huzinaga atomic basis set available for atoms H to Rn<sup>19,20</sup> was considered in the present study. On the basis of the fitting procedure put forward in the previous work,<sup>11</sup> where a general methodology was established, minimal modifications, involving the use of Taylor series expansions, were implemented in the program to improve the efficiency of the search for the optimal EJR rotation angle.

## ASA FIRST-ORDER DENSITY FUNCTIONS

Accurate analytical representations of electron densities are obtained using the LCAO-MO approximation, corresponding to the expression

$$\rho_A(\mathbf{r}) = \sum_{\mu, \nu} D_{\mu\nu} \chi_{\mu}(\mathbf{r}) \chi_{\nu}(\mathbf{r}), \quad (1)$$

where  $\{D_{\mu\nu}\}$  are the elements of the charge-bond order matrix and  $\{\chi_{\mu}\}$  is the atomic orbital basis set. To reduce computational requirements, simplified density functions have been proposed.<sup>6–11</sup> One of these is ASA electron density,<sup>10,11</sup> which is constructed by means of a linear combination of spherical Gaussian functions

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

where  $\{w_i\}$  represents the ASA coefficients and  $\{\varphi_i\}$  denotes the set of the corresponding 1S-type functions. Considering electron density as a probability distribution which may be definite positive *everywhere*, ASA coefficients of the linear expansion (2) have to be positive to preserve the statistical meaning of the density function.<sup>21</sup> Within this context, a set of convex conditions is imposed on the  $\{w_i\}$

coefficients, corresponding to

$$\left\{ w_i \quad R \quad \forall i \quad \sum_{i \in A} w_i \quad 1 \right\}, \quad (3)$$

which provides a normalized description of the electron density:

$$\int \rho_A^{\text{ASA}}(\mathbf{r}) d\mathbf{r} = 1.$$

Furthermore, a *promolecular* approximation can be used in order to avoid the computation of *ab initio* molecular density functions, consisting of building the molecular electronic distribution as a sum of individual atomic densities.<sup>11,13–18</sup>

### FITTING ALGORITHM

In a recent article<sup>11</sup> a broad description of ASA fitting was carried out, giving exhaustive details of the whole procedure. Fundamentally the method is based on two parts: generation of ASA exponents using *even-tempered* geometric sequences<sup>22</sup> and optimization of coefficients and exponents using an EJR technique and a Newton method, respectively. Here, on the basis of the established general procedure,<sup>11</sup> only some parts of the algorithm will be attended to, corresponding to some improvements made on the optimization procedure of ASA coefficients. Basically, the optimal ASA coefficients are obtained minimizing the quadratic error integral function between *ab initio* and ASA electronic density functions:

$$\varepsilon^{(2)} = \int \rho_A(\mathbf{r}) - \rho_A^{\text{ASA}}(\mathbf{r})^2 d\mathbf{r}, \quad (4)$$

subjected to the constraints described in eq. (3). Substituting the density functions  $\rho_A(\mathbf{r})$  and  $\rho_A^{\text{ASA}}(\mathbf{r})$  with the expressions (1) and (2), respectively, and developing the square term appearing in eq. (4), a simple matrix notation of the function  $\varepsilon^{(2)}$  can be obtained:

$$\varepsilon^{(2)} = Z_{\text{AA}} \mathbf{w} \mathbf{Z} \mathbf{w} - 2\mathbf{b} \mathbf{w}, \quad (5)$$

where  $Z_{\text{AA}}$  denotes an *ab initio* overlaplike quantum self-similarity measure (QS-SM),<sup>1–5</sup> defined by the integral

$$Z_{\text{AA}} = \int \rho_A(\mathbf{r}) \rho_A(\mathbf{r}) d\mathbf{r} = \sum_{\mu, \nu \in A} D_{\mu\nu} \sum_{\lambda, \sigma \in A} D_{\lambda\sigma} \int \chi_\mu(\mathbf{r}) \chi_\nu(\mathbf{r}) \chi_\lambda(\mathbf{r}) \chi_\sigma(\mathbf{r}) d\mathbf{r}, \quad (6)$$

whereas the vector  $\mathbf{w} = \{w_i\}$  contains the set of ASA coefficients, the matrix  $\mathbf{Z} = \{Z_{ij}\}$  corresponds to the ASA similarity measures between the atomic shells  $i$  and  $j$ ,

$$Z_{ij} = \int \varphi_i(\mathbf{r})^2 \varphi_j(\mathbf{r})^2 d\mathbf{r}, \quad (7)$$

and the elements of the vector  $\mathbf{b} = \{b_i\}$  involve the integrals between *ab initio* density function and each  $i$ th ASA function

$$b_i = \int \varphi_i(\mathbf{r})^2 \rho_A(\mathbf{r}) d\mathbf{r}. \quad (8)$$

The set of positive definite real coefficients  $\{w_i\}$  can be substituted by a complex coefficient set,  $\mathbf{x} = \{x_i\}$ , defined as  $w_i = x_i^2$ . In this way, the first restriction described in eq. (3) is fulfilled, and the function  $\varepsilon^{(2)}$  is transformed to the expression

$$\varepsilon^{(2)} = Z_{\text{AA}} \sum_{i,j \in A} x_i^2 x_j^2 Z_{ij} - 2 \sum_{i \in A} x_i^2 b_i. \quad (9)$$

An elegant method for optimizing the coefficients  $\{x_i\}$  is provided by the EJR technique. This approach, initially developed as a diagonalization method for symmetric matrices,<sup>12</sup> is based on norm conserving orthogonal transformations. This quality is particularly advantageous because, if the initial set of coefficients fulfill the normalization condition given in eq. (3),  $\sum x_i^2 = 1$ , then this constraint is conserved throughout all the procedure. The application of an EJR transformation,  $\mathbf{J}_{pq}(\alpha)$ , over the vector  $\mathbf{x}$  can be described by the equations

$$\begin{aligned} \dot{x}_p &\leftarrow c x_p - s x_q \\ \dot{x}_q &\leftarrow s x_p + c x_q, \end{aligned} \quad (10)$$

where only the elements  $p$  and  $q$  are modified. The symbols  $c$  and  $s$  appearing in eq. (10) determine the cosine and sine of the rotation angle  $\alpha$ . In order to compute the variation of the function  $\varepsilon^{(2)}$  with respect to the active pair of elements  $\{x_p, x_q\}$ , first it is necessary to isolate these elements from the rest in eq. (9), and then apply the EJR  $\mathbf{J}_{pq}(\alpha)$  over this equation, yielding

$$\begin{aligned} \delta \varepsilon^{(2)} &= \delta x_p^4 Z_{pp} - \delta x_q^4 Z_{qq} \\ &+ 2\delta(x_p^2 x_q^2) Z_{pq} - 2\delta x_p^2 \sum_{i \in p,q} x_i^2 Z_{pi} \\ &+ 2\delta x_q^2 \sum_{i \in p,q} x_i^2 Z_{iq} \\ &+ 2b_p \delta x_p^2 - 2b_q \delta x_q^2. \end{aligned} \quad (11)$$

The parameters  $\delta x_p^2$ ,  $\delta x_q^2$ ,  $\delta x_p^4$ ,  $\delta x_q^4$ , and  $\delta(x_p^2 x_q^2)$  are easily calculated (see ref. 11), giving as a result a quartic equation with respect to  $s$  and  $c$ :

$$\delta \mathcal{E}^{(2)} = E_{04} s^4 + E_{13} c s^3 + E_{02} s^2 + E_{11} c s. \quad (12)$$

The optimal sine,  $s^*$ , related to the EJR procedure is obtained by imposing a zero gradient condition to the function  $\delta \mathcal{E}^{(2)}$ :  $d \delta \mathcal{E}^{(2)} / ds = 0$ . As described in the first article of this series,<sup>11</sup> the optimal EJR angle,  $\alpha^*$ , is obtained by solving the resulting second-degree equation by means of an iterative algorithm. The aim of the present work is to improve the procedure for calculating  $\alpha^*$ . In this way, some modifications have been introduced in the algorithm, consisting in replacing the sine and cosine expressions of eq. (10) by a Taylor series expansion,<sup>23</sup> as may be done in parent SCF procedures<sup>24</sup>:

$$\begin{aligned} c &= \cos(\alpha) = 1 - \frac{1}{2} \alpha^2 + \theta(\alpha^3) \\ s &= \sin(\alpha) = \alpha \left( 1 - \frac{1}{6} \alpha^2 \right) + \theta(\alpha^4). \end{aligned} \quad (13)$$

From this reduced notation of the sine and cosine it is possible to define the subsequent product of variables up to third-order on  $\alpha$ :

$$\begin{aligned} s^2 &= \alpha^2 \\ s^3 &= \alpha^3 \\ s^4 &= 0 \\ cs &= \alpha \left( 1 - \frac{2}{3} \alpha^2 \right) \\ cs^3 &= \alpha^3. \end{aligned} \quad (14)$$

Consequently, by substituting these variables into eq. (12) a simpler expression is obtained:

$$\delta \mathcal{E}^{(2)} = \alpha^3 a + \alpha^2 b + \alpha c, \quad (15)$$

where  $a = E_{13} - (2E_{11} - 3)$ ,  $b = E_{02}$ , and  $c = E_{11}$ . When a stationary point condition is taken on (15), the following equation is obtained:

$$\frac{d \delta \mathcal{E}^{(2)}}{d \alpha} = 3 \alpha^2 a + 2 \alpha b + c = 0 \quad (16)$$

and the second derivative determines the minimum condition

$$\frac{d^2 \delta \mathcal{E}^{(2)}}{d \alpha^2} = 6 \alpha a + 2b. \quad (17)$$

The optimal angle is then

$$\alpha^* = \alpha = \frac{1}{3a} \left[ -b \pm (b^2 - 3ac)^{1/2} \right], \quad (18)$$

so the optimal cosine and sine are given by

$$\begin{aligned} c^* &= 1 - \frac{1}{2} \alpha^2 \\ s^* &= \alpha \left( 1 - \frac{1}{6} \alpha^2 \right). \end{aligned} \quad (19)$$

Substituting eq. (19) in eq. (10), a simple expression for the coefficient set variation is obtained:

$$\begin{aligned} \dot{x}_p &\leftarrow c^* x_p - s^* x_q \\ x_p &= \alpha x_q + \frac{1}{2} \alpha^2 \left( -x_p + \frac{1}{3} \alpha x_q \right) \\ \dot{x}_q &\leftarrow s^* x_p + c^* x_q \\ &= \alpha x_p + x_q + \frac{1}{2} \alpha^2 \left( \frac{1}{3} \alpha x_p - x_q \right), \end{aligned} \quad (20)$$

which gives the optimal EJR transformation. Evidently, the main advantage of the approach presented here is the replacement of an iterative method by a simpler determination of  $\alpha^*$ . This will provide a faster algorithm.

### ATOMIC ASA DENSITY FITTING OF AB INITIO HUZINAGA BASIS SET

As an application example of the EJR algorithm new development, an atomic basis set was fitted for atoms H to Rn. *Ab initio* calculations for the fitting procedure were obtained from a Huzinaga basis set<sup>19,20</sup> using the ATOMIC program. Among the multiple basis set schemes provided by refs. 19 and 20, the set of primitive functions listed in Table I was chosen, which is described using the original Huzinaga's contraction scheme notation. The fitting process was performed using the GATOMIC program,<sup>26</sup> which includes the modifications on the EJR algorithm described in the previous section. First a minimal ASA basis set for the atoms H to Rn was calculated, corresponding to the least number of fitted functions needed to obtain a value of the function  $\mathcal{E}^{(2)}$  inferior to 0.01 au. In a similar way, as in the previous study for the 3-21G basis set,<sup>11</sup> the number of atomic shells per atom varies with respect to the row of the periodic table from two functions for H and He atoms to seven functions for the elements of the sixth row. Recent studies demonstrated that this

**TABLE I.**  
**Notation for Contracted Gaussian Primitive Basis Set.**

Atomic Symbol	Huzinaga Notation <sup>a</sup>
H, He	3
Li, Be	33
B, C, N, O, F, Ne	33 / 3
Na, Mg	432 / 3
Al, Si, P, S, Cl, Ar	432 / 42
K, Ca	4322 / 42
Sc, Ti, V, Cr, Mn, Fe, Co,	4322 / 42 / 3
Ni, Cu, Zn	
Ga, Ge, As, Se, Br, Kr	4322 / 422 / 3
Rb, Sr	43222 / 422 / 3
Y, Zr, Nb, Mo, Tc, Ru, Rh,	43222 / 422 / 33
Pd, Ag, Cd	
In, Sn, Sb, Te, I, Xe	43222 / 4222 / 42
Cs, Ba	432222 / 4222 / 42
La	432222 / 4222 / 42 / 3
Ce, Pr, Nd, Pm, Sm, Eu, Gd,	432222 / 4222 / 42 / 4
Tb, Dy, Ho, Er, Tm, Yb	
Lu, Hf, Ta, W, Re, Os, Ir,	432222 / 4222 / 423 / 3
Pt, Au, Hg	
Tl, Pb, Bi, Po, At, Rn	432222 / 42222 / 422 / 3

<sup>a</sup>The expansion pattern ( $K_{s1}, K_{s2}, \dots / K_{p1}, K_{p2}, \dots / K_{d1}, K_{d2}, \dots / K_{f1}, \dots$ ) is used as the notation to specify the number of terms in the expansion of the atomic basis set.

minimal atomic basis set is sufficient to obtain excellent results for QSM applications.<sup>13–18</sup> The main results for the ASA fitting are presented in Table II, where the values for the Hartree–Fock (HF) energy, normalized *ab initio* QS-SM, number of fitted functions, quadratic error integral function, and relative error produced in the QS-SM for the atoms H to Rn are summarized. In the last column of Table II the percent error in  $Z_{AA}$  evaluation is given. Because the procedure optimizes  $\varepsilon^{(2)}$  rather than  $Z_{AA}$  values, the self-similarity errors are sufficiently low and vary in a random manner. Coefficients and exponents for this basis set of ASA functions are available for downloading at a worldwide web site.<sup>27</sup>

In addition to the minimal basis set for atoms H to Rn, a detailed study was performed for several atoms to verify the correctness of the ASA fitting methodology. Table III shows the most significant parameters related to the ASA fitting for H, C, O, N, F, P, S, and Cl atoms, employing from two to eight functions. The analyzed parameters are the values of the function  $\varepsilon^{(2)}$  and the relative error performed in the calculation of ASA QS-SM and normalized one-electron potential energy,  $V(\mathbf{r})$ ,

which may be defined as

$$V(\mathbf{r}) = \int \frac{\rho_A^{\text{ASA}}(\mathbf{r})}{\mathbf{r}} d\mathbf{r} = \sum_i w_i \int \frac{1}{\mathbf{r}} S_i(\mathbf{r})^2 d\mathbf{r}. \quad (21)$$

As the number of ASA functions increases, the quadratic error integral function quickly decreases. Moreover, from Table III it is possible to appreciate a relationship between the relative error produced in  $Z_{AA}$  and  $V(\mathbf{r})$  values, giving similar variations. There is a general tendency to decrease the value of both parameters when the number of ASA functions increases, but not as pronounced as in the values of the  $\varepsilon^{(2)}$  function. In fact, the main conclusion that can be drawn from Table III is that with few ASA functions it is possible to obtain an excellent fit.

### QSM APPLICATION EXAMPLE

In many theoretical studies concerning geometry optimization and energy minimization, the selection of the appropriate methodology that is used becomes an initial and essential process, which, in many cases, is not taken into account. It is crucial therefore to use suitable 3-dimensional (3-D) structures in the majority of computational chemistry calculations, for example, in the study of path reaction, where small changes on the molecular geometry can provide incorrect conclusions and results. A method for quantitatively comparing calculated molecular geometries obtained from different approaches and levels of theory can be based on QSM, as has been demonstrated in a recent work,<sup>14</sup> which relies on determining the differences between experimental and theoretical structures. In this work, the *cis*-diammine-dichloroplatinum(II) complex (*cis*-DDP), an anti-cancer drug,<sup>28</sup> was chosen to provide an application example. This transition metal complex is a good test for the application of the fitted atomic Gaussian basis set presented in the previous section, because in it a heavy atom, such as Pt, takes part. First a search of the Cambridge Structural Database (CDS)<sup>29</sup> was performed, and two X-ray crystallographic structures for the *cis*-DDP complex were found: *cis*-DDP dimethylformamide solvate (CUKRAB) and 18-Crown-6 bis(dimethylacetamide) bis(*cis*-DDP) (CUSRAJ).

Eight different structures are compared, two of them from X-ray analysis and six from theoretical calculations. These last geometries were fully opti-



**TABLE II.**  
**Fitting Results for *Ab Initio* Huzinaga Atomic Basis Set for Atoms H to Rn: Minimal ASA Basis Set.**

Atomic Symbol	Electronic State	HF <sup>a</sup>	Z <sub>AA</sub>	n <sup>b</sup>	$\varepsilon^{(2)}$	Z <sub>AA</sub> Error (%)
H	<sup>2</sup> S	4.969792526E-01	0.03939	2	0.0001463	2.5725
He	<sup>1</sup> S	2.835679876E+00	0.18800	2	0.0010896	3.1134
Li	<sup>2</sup> S	7.378092307E+00	0.34452	3	0.0007891	0.0486
Be	<sup>1</sup> S	1.447611084E+01	0.51917	3	0.0009877	0.0430
B	<sup>2</sup> P	2.437272923E+01	0.69367	3	0.0012675	0.0120
C	<sup>3</sup> P	3.745282379E+01	0.87234	3	0.0015070	0.1105
N	<sup>4</sup> S	5.406244335E+01	1.06079	3	0.0020636	0.2125
O	<sup>3</sup> P	7.433921998E+01	1.25522	3	0.0026446	0.3606
F	<sup>2</sup> P	9.877655073E+01	1.46136	3	0.0036004	0.5171
Ne	<sup>1</sup> S	1.277187909E+02	1.67697	3	0.0046583	0.6919
Na	<sup>2</sup> S	1.614200178E+02	1.91971	4	0.00392992	1.1778
Mg	<sup>1</sup> S	1.991000990E+02	2.15931	4	0.0045979	0.0805
Al	<sup>2</sup> P	2.415557168E+02	2.41336	4	0.0046660	0.1245
Si	<sup>3</sup> P	2.884848502E+02	2.66285	4	0.0048889	0.1386
P	<sup>4</sup> S	3.402953458E+02	2.91708	4	0.0050630	0.1379
S	<sup>3</sup> P	3.970195424E+02	3.17561	4	0.0052278	0.1315
Cl	<sup>2</sup> P	4.589288466E+02	3.43832	4	0.0053933	0.1248
Ar	<sup>1</sup> S	5.261904122E+02	3.70604	4	0.0055419	0.1181
K	<sup>2</sup> S	5.984734792E+02	3.98451	5	0.0003748	0.0376
Ca	<sup>1</sup> S	6.760028693E+02	4.26335	5	0.0005072	0.0848
Sc	<sup>2</sup> D	7.588683361E+02	4.53727	5	0.0005682	0.1017
Ti	<sup>3</sup> F	8.474105511E+02	4.81189	5	0.0006353	0.1103
V	<sup>4</sup> F	9.417431646E+02	5.08829	5	0.0007111	0.1282
Cr	<sup>5</sup> D	1.042004521E+03	5.36746	5	0.0008079	0.1432
Mn	<sup>6</sup> S	1.148378100E+03	5.64709	5	0.0009094	0.1577
Fe	<sup>5</sup> D	1.260742127E+03	5.92997	5	0.0010303	0.1695
Co	<sup>4</sup> F	1.379478279E+03	6.21501	5	0.0011963	0.1955
Ni	<sup>3</sup> F	1.504675051E+03	6.50265	5	0.0013679	0.2153
Cu	<sup>2</sup> D	1.636468665E+03	6.79265	5	0.0015600	0.2344
Zn	<sup>1</sup> S	1.775056361E+03	7.08444	5	0.0017706	0.2585
Ga	<sup>2</sup> P	1.920361458E+03	7.38764	5	0.0024764	0.3258
Ge	<sup>3</sup> P	2.072336570E+03	7.69236	5	0.0034626	0.4018
As	<sup>4</sup> S	2.231077402E+03	8.00019	5	0.0047922	0.4823
Se	<sup>3</sup> P	2.396555652E+03	8.31116	5	0.0064939	0.5650
Br	<sup>2</sup> P	2.568968763E+03	8.62436	5	0.0086121	0.6487
Kr	<sup>1</sup> S	2.748411490E+03	8.94033	5	0.0093826	0.1030
Rb	<sup>2</sup> S	2.934540749E+03	9.25956	6	0.0005184	0.0001
Sr	<sup>1</sup> S	3.127572218E+03	9.58254	6	0.0006653	0.0018
Y	<sup>2</sup> D	3.327645322E+03	9.90583	6	0.0007058	0.0025
Zr	<sup>3</sup> F	3.534786249E+03	10.22872	6	0.0006814	0.0045
Nb	<sup>6</sup> D	3.749171741E+03	10.55127	6	0.0006842	0.0013
Mo	<sup>7</sup> S	3.970931775E+03	10.87764	6	0.0007322	0.0011
Tc	<sup>6</sup> S	4.200005584E+03	11.20902	6	0.0008185	0.0043
Ru	<sup>5</sup> F	4.436501714E+03	11.53459	6	0.0008571	0.0013
Rh	<sup>4</sup> F	4.680620905E+03	11.86623	6	0.0010686	0.0015
Pd	<sup>3</sup> D	4.932403048E+03	12.19779	6	0.0011365	0.0017
Ag	<sup>2</sup> S	5.191969936E+03	12.53153	6	0.0013018	0.0011
Cd	<sup>1</sup> S	5.459204578E+03	12.87027	6	0.0011572	0.0064
In	<sup>2</sup> P	5.735169884E+03	13.21283	6	0.0011237	0.0143
Sn	<sup>3</sup> P	6.017774127E+03	13.55495	6	0.0012129	0.0237
Sb	<sup>4</sup> S	6.308159013E+03	13.89854	6	0.0013923	0.0310
Te	<sup>3</sup> P	6.606280606E+03	14.24398	6	0.0015830	0.0400

TABLE II.  
(Continued)

Atomic Symbol	Electronic State	HF <sup>a</sup>	Z <sub>AA</sub>	n <sup>b</sup>	$\epsilon^{(2)}$	Z <sub>AA</sub> Error (%)
I	<sup>2</sup> P	6.912293010E + 03	14.59028	6	0.0018443	0.0480
Xe	<sup>1</sup> S	7.226259525E + 03	14.94006	6	0.0021366	0.0562
Cs	<sup>2</sup> S	7.547866175E + 03	15.29126	7	0.0016553	0.0108
Ba	<sup>1</sup> S	7.877289281E + 03	15.64351	7	0.0017217	0.0103
La	<sup>2</sup> F	8.214448257E + 03	15.98328	7	0.0016739	0.0095
Ce	<sup>3</sup> H	8.560134394E + 03	16.33152	7	0.0018155	0.0102
Pr	<sup>4</sup> I	8.914118872E + 03	16.67686	7	0.0018673	0.0103
Nd	<sup>5</sup> I	9.276534340E + 03	17.02378	7	0.0019810	0.0111
Pm	<sup>6</sup> H	9.647454065E + 03	17.37218	7	0.0019589	0.0102
Sm	<sup>7</sup> F	1.002700168E + 04	17.72103	7	0.0037966	0.0997
Eu	<sup>8</sup> S	1.041527457E + 04	18.07250	7	0.0020845	0.0104
Gd	<sup>7</sup> F	1.081201752E + 04	18.42125	7	0.0027032	0.0136
Tb	<sup>6</sup> H	1.121763004E + 04	18.77186	7	0.0023948	0.0118
Dy	<sup>5</sup> I	1.163216483E + 04	19.12155	7	0.0024998	0.0121
Ho	<sup>4</sup> I	1.205563774E + 04	19.47228	7	0.0028697	0.0134
Er	<sup>3</sup> H	1.248813204E + 04	19.82825	7	0.0024286	0.0110
Tm	<sup>2</sup> F	1.292976991E + 04	20.17961	7	0.0025990	0.0115
Yb	<sup>1</sup> S	1.338065566E + 04	20.53538	7	0.0030558	0.0139
Lu	<sup>2</sup> D	1.383819206E + 04	20.89906	7	0.0058882	0.1618
Hf	<sup>3</sup> F	1.430740505E + 04	21.26215	7	0.0029727	0.0121
Ta	<sup>4</sup> F	1.478572471E + 04	21.62753	7	0.0027151	0.0104
W	<sup>5</sup> D	1.527318958E + 04	21.99249	7	0.0031980	0.0128
Re	<sup>6</sup> S	1.576990684E + 04	22.35979	7	0.0030633	0.0120
Os	<sup>5</sup> D	1.627572807E + 04	22.73038	7	0.0032791	0.0128
Ir	<sup>4</sup> F	1.679088555E + 04	23.10037	7	0.0031468	0.0113
Pt	<sup>3</sup> F	1.731541053E + 04	23.46991	7	0.0033635	0.0125
Au	<sup>2</sup> D	1.784934901E + 04	23.84188	7	0.0034272	0.0122
Hg	<sup>1</sup> S	1.839279390E + 04	24.21301	7	0.0035502	0.0125
Tl	<sup>2</sup> P	1.894511809E + 04	24.58568	7	0.0037250	0.0120
Pb	<sup>3</sup> P	1.950699432E + 04	24.96366	7	0.0038512	0.0107
Bi	<sup>4</sup> S	2.007825555E + 04	25.34290	7	0.0039560	0.0094
Po	<sup>3</sup> P	2.065884451E + 04	25.72485	7	0.0041634	0.0091
At	<sup>2</sup> P	2.124889325E + 04	26.10476	7	0.0043268	0.0082
Rn	<sup>1</sup> S	2.184844245E + 04	26.48835	7	0.0045285	0.0080

<sup>a</sup>Hartree–Fock energy computed using the ATOMIC program.<sup>25</sup><sup>b</sup>Number of fitted atomic functions.

mized using the Gaussian 94 program<sup>30</sup> and LANL2DZ basis set, which includes relativistic effective core potential for the second and third row transition metals. The quantum chemical calculations were carried out at the HF level; second-, third-, and fourth-order Møller–Plesset (MP) approaches; and B3LYP level, which corresponds to a hybrid density functional theory (DFT). The main structural parameters for all molecular structures are presented in Table IV. Unfortunately, in both 3-D X-ray structures the *cis*-DDP complex is solvated. These environmental effects produce some minor deformations in the experimental geome-

tries of *cis*-DDP, providing a nonsymmetric and distorted 3-D structure, as can be observed in Table IV.

For the QSM computations,  $\rho_A^{\text{ASA}}(\mathbf{r})$  functions are constructed using a *promolecular* approximation<sup>13–18</sup> and the minimal basis set described in the previous section, employing three functions for N, four functions for Cl, and seven functions for Pt. Because X-ray analysis does not determine the position of hydrogen atoms, a possible alternative consists of comparing only the molecular fragment PtCl<sub>2</sub>N<sub>2</sub>. Molecular alignments were obtained using the MOLSIMIL program<sup>31</sup> and the methodol-

**TABLE III.**  
Fitting Results for *Ab Initio* Huzinaga Atomic Basis Set for Atoms H, C, N, O, F, P, S, and Cl.

	No. Fitted Atomic Functions														
	2		3		4		5		6		7		8		
H	$\varepsilon^{(2)}$	1.46	10 <sup>4</sup>	1.40	10 <sup>6</sup>	9.49	10 <sup>7</sup>	2.62	10 <sup>7</sup>	6.06	10 <sup>8</sup>	6.42	10 <sup>9</sup>	6.09	10 <sup>9</sup>
	%Z <sub>AA</sub>	2.572		0.006		0.123		0.018		0.013		0.006		0.004	
	%V(r)	0.950		0.030		0.078		0.044		0.010		0.008		0.006	
C	$\varepsilon^{(2)}$	3.08	10 <sup>2</sup>	1.51	10 <sup>3</sup>	6.39	10 <sup>5</sup>	1.29	10 <sup>5</sup>	1.11	10 <sup>5</sup>	4.54	10 <sup>6</sup>	4.22	10 <sup>6</sup>
	%Z <sub>AA</sub>	2.455		0.111		0.107		0.008		0.006		0.008		0.005	
	%V(r)	1.802		0.347		0.433		0.061		0.013		0.065		0.019	
N	$\varepsilon^{(2)}$	3.66	10 <sup>2</sup>	2.06	10 <sup>3</sup>	1.42	10 <sup>4</sup>	1.13	10 <sup>5</sup>	1.10	10 <sup>5</sup>	7.14	10 <sup>6</sup>	5.54	10 <sup>6</sup>
	%Z <sub>AA</sub>	2.246		0.212		0.191		0.003		0.003		0.014		0.011	
	%V(r)	1.676		0.502		0.566		0.002		0.017		0.085		0.071	
O	$\varepsilon^{(2)}$	4.26	10 <sup>2</sup>	2.64	10 <sup>3</sup>	2.98	10 <sup>4</sup>	2.15	10 <sup>5</sup>	1.39	10 <sup>5</sup>	7.26	10 <sup>6</sup>	6.31	10 <sup>6</sup>
	%Z <sub>AA</sub>	1.981		0.361		0.265		0.008		0.004		0.010		0.011	
	%V(r)	1.419		0.713		0.708		0.047		0.007		0.037		0.054	
F	$\varepsilon^{(2)}$	4.90	10 <sup>2</sup>	3.60	10 <sup>3</sup>	5.27	10 <sup>4</sup>	2.39	10 <sup>5</sup>	1.60	10 <sup>5</sup>	7.42	10 <sup>6</sup>	7.34	10 <sup>6</sup>
	%Z <sub>AA</sub>	1.658		0.517		0.396		0.023		0.016		0.017		0.005	
	%V(r)	1.172		0.879		0.870		0.132		0.058		0.066		0.025	
P	$\varepsilon^{(2)}$	1.35	10 <sup>1</sup>	4.37	10 <sup>2</sup>	5.06	10 <sup>3</sup>	4.57	10 <sup>4</sup>	7.87	10 <sup>5</sup>	1.16	10 <sup>5</sup>	1.02	10 <sup>5</sup>
	%Z <sub>AA</sub>	2.462		3.639		0.138		0.015		0.012		0.006		0.006	
	%V(r)	6.492		7.827		0.113		0.134		0.103		0.082		0.063	
S	$\varepsilon^{(2)}$	1.63	10 <sup>1</sup>	6.02	10 <sup>2</sup>	5.23	10 <sup>3</sup>	4.17	10 <sup>4</sup>	1.32	10 <sup>4</sup>	2.00	10 <sup>5</sup>	1.17	10 <sup>5</sup>
	%Z <sub>AA</sub>	3.023		3.983		0.131		0.011		0.012		0.006		0.005	
	%V(r)	7.344		8.507		0.141		0.114		0.137		0.091		0.081	
Cl	$\varepsilon^{(2)}$	1.94	10 <sup>1</sup>	7.83	10 <sup>2</sup>	5.39	10 <sup>3</sup>	3.72	10 <sup>4</sup>	6.18	10 <sup>5</sup>	2.30	10 <sup>5</sup>	1.94	10 <sup>5</sup>
	%Z <sub>AA</sub>	3.451		2.056		0.125		0.015		0.015		0.008		0.006	
	%V(r)	7.974		1.118		0.143		0.114		0.111		0.085		0.086	

ogy described in ref. 32. To show the structural differences of studied theoretical approaches, Carbó indices<sup>1</sup> were computed for all possible molecular pairs and are listed in Table V. The Carbó index, which can be denoted as  $C_{AB}$ , is defined as

$$C_{AB} = Z_{AB}(Z_{AA}Z_{BB})^{1/2} \quad (22)$$

and, in the present work, has been scaled to lie within the range 0–100. Although  $C_{AB}$  values dif-

fer in a small amount, the differences are sufficient to obtain information on the optimal theoretical method to study this kind of molecules. The results suggest that the second-order MP (MP2) method is the theoretical approach that presents the best agreement with the available experimental data for the *cis*-DDP complex. This conclusion is in contrast to the major studies performed until now with organoplatinum complexes, which employ DFT methods by default (see, e.g., refs. 33 and 34). However, the present study overemphasizes

**TABLE IV.**  
Principal Structural Parameters for Eight Geometries of Molecular Fragment PtCl<sub>2</sub>N<sub>2</sub>.

	CUKRAB <sup>a</sup>	CUSRAJ <sup>a</sup>	B3LYP <sup>b</sup>	HF <sup>b</sup>	MP2 <sup>b</sup>	MP3 <sup>b</sup>	MP4(DQ) <sup>b</sup>	MP4(SDQ) <sup>b</sup>
R <sub>Pt-Cl</sub>	2.32–2.31	2.29–2.28	2.410	2.415	2.406	2.411	2.411	2.414
R <sub>Pt-N</sub>	2.00–2.08	2.00–2.04	2.110	2.126	2.123	2.128	2.129	2.132
R <sub>N-H</sub>	—	—	1.023 1.033	1.005 1.010	1.027 1.034	1.025 1.031	1.027 1.033	1.027 1.033
CIPTCl	92.40	93.21	96.784	97.127	95.880	96.234	96.180	96.042
CIPTN	89.93	88.31–87.09	81.962	83.561	83.290	83.322	83.350	83.289
NPtN	90.94	91.44	99.289	95.751	97.539	97.122	97.120	97.379

<sup>a</sup>Crystallographic structures obtained from CSD.<sup>29</sup>

<sup>b</sup>Optimized geometries using the Gaussian 94 program.<sup>30</sup>

**TABLE V.**  
**Carbó Index Values for Molecular Fragment PtCl<sub>2</sub>N<sub>2</sub> Used to Compare Different Optimization Methodologies.**

	CUKRAB	CUSRAJ	B3LYP	HF	MP2	MP3	MP4(DQ)	MP4(SDQ)
CUKRAB	100	99.569	98.804	98.800	98.834	98.820	98.821	98.818
CUSRAJ	99.569	100	98.782	98.772	98.806	98.793	98.794	98.791
B3LYP	98.804	98.782	100	99.925	99.762	99.897	99.880	99.826
HF	98.800	98.772	99.925	100	99.601	99.773	99.750	99.700
MP2	98.834	98.806	99.762	99.601	100	99.941	99.954	99.943
MP3	98.820	98.793	99.897	99.773	99.941	100	99.999	99.981
MP4(DQ)	98.821	98.794	99.880	99.750	99.954	99.999	100	99.985
MP4(SDQ)	98.818	98.791	99.826	99.700	99.943	99.981	99.985	100

molecular geometry, and this can explain such comparative results. Another interesting conclusion is that the different levels of MP approach give similar results without appreciable differences between them. The addition of third- and fourth-order perturbations does not give significant changes in the 3-D structure with respect to MP2, but computational requirements increase considerably. In particular, HF appears to be comparable to B3LYP. In accordance with this result, it is possible to assert that the HF method may be useful for the geometrical study of these kinds of organoplatinum complexes.

## Conclusions

The present study corroborates the usefulness of an EJR technique for obtaining first-order ASA density functions. Some algorithm improvements related to the calculation of the EJR rotation angle using a Taylor series expansion are given. In addition, a new set of atomic ASA  $\rho(\mathbf{r})$  is presented for atoms H to Rn. A sound application example is presented relating to QSM and concerning the determination of the best theoretical approach for the geometry optimization of the *cis*-DDP complex.

## Acknowledgments

This work was carried out using the CESCA and CEPBA resources, coordinated by C<sup>4</sup>. The second author thanks Prof. S. Huzinaga for providing detailed information on the *ab initio* basis sets. The authors also thank the referees for their constructive criticism, which improved several aspects of this work.

## References

1. Carbó, R.; Leyda, L.; Arnau, M. *Int J Quantum Chem* 1980, 17, 1185.
2. Carbó, R.; Besalú, E.; Calabuig, B.; Vera, L. *Adv Quantum Chem* 1994, 25, 253.
3. Besalú, E.; Carbó, R.; Mestres, J.; Solà, M. *Topics Curr Chem* 1995, 173, 31.
4. Carbó, R., Ed. *Molecular Similarity and Reactivity: From Quantum Chemical to Phenomenological Approaches*; Kluwer: Amsterdam, 1995.
5. Carbó-Dorca, R.; Mezey, P. G. Eds., *Advances in Molecular Similarity*; JAI Press: Greenwich, CT, 1996; Vol. 1.
6. Dunlap, B. I.; Connolly, J. W. D.; Sabin, J. R. *J Chem Phys* 1979, 71, 3396.
7. Mestres, J.; Solà, M.; Duran, M.; Carbó, R. *J Comput Chem* 1994, 15, 1113.
8. Gallant, R. T.; St-Amant, A. *Chem Phys Lett* 1996, 256, 569.
9. Cioslowski, J.; Piskorz, P.; Rez, P. *J Chem Phys* 1997, 106, 3607.
10. Constans, P.; Carbó, R. *J Chem Inf Comput Sci* 1995, 35, 1046.
11. Amat, L.; Carbó-Dorca, R. *J Comput Chem* 1997, 18, 2023.
12. Jacobi, C. G. J. *J Reine Angew Math* 1846, 30, 51.
13. Lobato, M.; Amat, L.; Besalú, E.; Carbó-Dorca, R. *Quantum Struct Act Rel* 1997, 16, 465.
14. Amat, L.; Robert, D.; Besalú, E.; Carbó-Dorca, R. *J Chem Inf Comput Sci* 1998, 38, 624.
15. Amat, L.; Carbó-Dorca, R.; Ponec, R. *J Comput Chem* 1998, 19, 1575.
16. Ponec, R.; Amat, L.; Carbó-Dorca, R. *J Comput Aided Mol Design* 1999, 13, 259.
17. Ponec, R.; Amat, L.; Carbó-Dorca, R. *J Phys Org Chem* to appear.
18. Robert, D.; Amat, L.; Carbó-Dorca, R. *J Chem Inf Comput Sci* 1999, 39, 333.
19. Huzinaga, S. Ed. *Gaussian Basis Sets for Molecular Calculations*. *Physical Sciences Data* 16; Elsevier: Amsterdam, 1984.
20. Huzinaga, S. *J Chem Phys* 1965, 42, 1293.
21. von Neumann, J. *Mathematical Foundations of Quantum Mechanics*; Princeton University Press: Princeton, NJ, 1955.

22. Huzinaga, S.; Klobukowski, M. *J Mol Struct (Theochem)* 1988, 167, 1.
23. Spiegel, M. R. *Mathematical Handbook*; McGraw-Hill: New York, 1968.
24. Bofill, J. M.; Bono, H.; Rubio, J. *J Comput Chem* 1998, 19, 368.
25. Carbó-Dorca, R. *ATOMIC Program*; Institute of Computational Chemistry, University of Girona; Catalonia, Spain, 1995. This program was based on Roos, B.; Salez, C.; Veillard, A.; Clementi, E. *A General Program for Calculation of SCF Orbitals by the Expansion Method*, IBM Research Paper RJ518 (#10901); IBM: Fishkill, NY, 1968.
26. Amat, L.; Carbó-Dorca, R. *GATOMIC Program*; Institute of Computational Chemistry, University of Girona; Catalonia, Spain, 1998.
27. ASA coefficients and exponents can be seen and downloaded from the worldwide web at [http://iqc.udg.es/cat/similarity/ASA\\_func432.html](http://iqc.udg.es/cat/similarity/ASA_func432.html).
28. Sherman, S. E.; Lippard, S. J. *Chem Rev* 1987, 87, 1153.
29. Allen, F. H.; Bellard, S.; Brice, M. D.; Cartwright, B. A.; Doubleday, A.; Higgs, H.; Hummelink, T.; Hummelink-Peters, B. G.; Kennard, O.; Motherwell, W. D. S.; Rodgers, J. R.; Watson, D. G. *Acta Crystallogr* 1979, B35, 2331.
30. 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.
31. Amat, L.; Constans, P.; Besalú, E.; Carbó-Dorca, R. *MOLSIMIL 97 Program*; Institute of Computational Chemistry, University of Girona; Catalonia, Spain, 1997.
32. Constans, P.; Amat, L.; Carbó-Dorca, R. *J Comput Chem* 1997, 18, 826.
33. Cui, Q.; Musaev, D. G.; Morokuma, K. *Organometallics* 1997, 16, 1355.
34. Hill, G. S.; Puddephatt, R. J. *Organometallics* 1998, 17, 1478.

### 3.10.3 Altres ajustos atòmics

Amb la nova versió del programa GATOMIC s'han determinat altres conjunts de funcions ASA. Així s'ha calculat les funcions ASA per als àtoms H fins a l'Ar fent un ajust de la densitat *ab initio* de cada àtom en la base 6-21G.<sup>39,40</sup> Un segon exemple és l'ajust de les funcions ASA per als àtoms H fins a l'Ar en una base 6-311G,<sup>46,47</sup> mostrats en l'article 3.3, i del Sc al Kr en la mateixa base,<sup>48</sup> presentats en l'article 3.4 d'aquest mateix capítol. Els exponents i coeficients ASA de totes les bases ajustades es poden trobar en la pàgina electrònica [42].

## 3.11 Ajustos moleculars

Com ja s'ha comentat, el principal inconvenient dels ajustos moleculars és la necessitat de calcular prèviament la matriu densitat *ab initio* de la molècula analitzada. No obstant això, la darrera innovació en l'algorisme d'ajust dels coeficients de les funcions ASA mitjançant la tècnica *EJR* ha estat la seva adaptació a molècules. L'ajust de les *DF* moleculars pren com a punt de partida la funció *PASA* i únicament fa un refinament dels coeficients de les funcions per adaptar-los a l'entorn molecular considerat. S'ha emprat la nomenclatura *FMASA* per designar les *DF* resultants de l'ajust molecular. Evidentment la descripció de la densitat electrònica d'una molècula millora si es compara amb la *PASA DF*. L'estudi també ha servit per analitzar algunes propietats moleculars que no es descriuen correctament amb les *PASA DF*, com són els potencials electrostàtics.

### 3.11.1 Esquema computacional

L'algorisme d'ajust de la densitat *ab initio* d'una molècula, descrit en la taula 3.9, pren com a punt de partida la *PASA DF* i així s'aconsegueix que el procés sigui molt més ràpid. Els coeficients ASA s'optimitzen primer amb l'algorisme de mínims quadrats descrit en l'apartat 3.4, el qual minimitza la funció error quadràtic integral entre les densitats exacta i aproximada del sistema analitzat sense supeditar els coeficients a ser positius. Si s'obté algun coeficient negatiu, llavors el programa

recupera els coeficients inicials *PASA* i executa la subrutina *EJR* descrita en la taula 3.3. En els estudis moleculars, la simplificació de l'optimització del sinus *EJR* expressant les variables *s* i *c* per mitjà de desenvolupaments en sèries truncades de Taylor és molt eficaç perquè els coeficients inicials *PASA* són una molt bona aproximació als coeficients finals *FMASA*, de manera que no requereixen de modificacions ostensibles. Pel que concerneix als exponents de les funcions *ASA*, es mantenen invariants amb referència als originals de la *PASA DF*.

---

#### Programa GATOMMOL

- ✓ Donada una molècula *A*, amb coordenades atòmiques  $\mathbf{A}=(\mathbf{a}_1 \ \mathbf{a}_2 \ \dots \ \mathbf{a}_m)$ , nombres atòmics  $\mathbf{P}_A=\{P_a\}$  i el nombre d'àtoms  $m_A$
  - ✓ Donada la matriu densitat  $D_{mm}$ , i el conjunt de funcions de base  $\{\mathbf{c}_m\}$
  - Demana **subrutina PASA DF**. Necessita:  $m_A, \mathbf{P}_A$  i especificar un conjunt de funcions *ASA* de base. Retorna:  $\{n_A, \mathbf{x}_{A,0}, \mathbf{w}_{A,0}, \mathbf{F}_A\}$
  - Calcula  $Z_{AA}, \mathbf{Z}$  i  $\mathbf{b}$ . Necessita:  $D_{mm}, \{\mathbf{c}_m\}, n_A, \mathbf{F}_A$
  - Calcula  $\mathbf{e}_0^{(2)} = Z_{AA} + \mathbf{w}_{A,0}^\top \mathbf{Z} \mathbf{w}_{A,0} - 2\mathbf{b}^\top \mathbf{w}_{A,0}$
  - Demana **subrutina MÍNIMS QUADRATS**. Necessita:  $n_A, Z_{AA}, \mathbf{Z}, \mathbf{b}, \mathbf{w}_{A,0}$  i  $\mathbf{e}_0^{(2)}$ . Retorna:  $\mathbf{w}_{A,op}$  i  $\mathbf{e}^{(2)}$
  - Si  $\exists w_{i,op} < 0 \ \forall i \in A$  llavors
    - Demana **subrutina EJR**. Necessita:  $n_A, Z_{AA}, \mathbf{Z}, \mathbf{b}, \mathbf{x}_{A,0}, \mathbf{w}_{A,0}$  i  $\mathbf{e}_0^{(2)}$ . Retorna:  $\mathbf{w}_{A,op}$  i  $\mathbf{e}^{(2)}$
  - Fi condicional
  - ❖ Sortida del programa: coeficients *ASA* òptims
- 

**Taula 3.9** Esquema global del Programa GATOMMOL

### 3.11.2 Exemples d'ajustos moleculars

La metodologia proposada per l'ajust molecular es descriu en l'article 3.3, a més d'alguns exemples d'aplicació. Un dels objectius de l'estudi ha estat comprovar que l'algorisme *EJR* desenvolupat per sistemes atòmics es pot emprar en molècules, malgrat que les densitats electròniques resultants tinguin una utilització limitada en l'àmbit *QSAR/QSPR* degut a la necessitat d'haver de calcular prèviament la *DF ab initio*. L'article 3.3 també mostra un nou exemple d'ajust *ASA* atòmic sobre un conjunt de funcions de base 6-311G.<sup>46,47</sup>

Quant als ajustos moleculars, s'ha fet un estudi comparatiu de les funcions *PASA* i *FMASA* sobre un conjunt de derivats del metà. Una altra manera de determinar la qualitat de l'ajust molecular ha estat comparant els resultats amb els obtinguts amb mètode descrit en l'apartat 3.5.1 de quasi saturació de l'espai de funcions *ASA*, i que s'ha considerat com el límit *ASA*. S'ha observat que els paràmetres de l'ajust són equivalents als valors de l'aproximació límit *ASA*,<sup>11</sup> però emprant moltes menys funcions. Una altra innovació ha estat la introducció d'un nou paràmetre per qualificar els ajustos moleculars, com és l'error relatiu produït en la *MQSM* de tipus Coulomb. En l'aproximació *ASA*, la *MQSM* de tipus Coulomb entre dues capes  $i \in a$  i  $j \in b$  es defineix com:

$$Z_{ij}(\mathbf{r}_{12}^{-1}) = 2 \left( \frac{2\mathbf{z}_i \mathbf{z}_j}{\pi(\mathbf{z}_i + \mathbf{z}_j)} \right)^{1/2} F_0 \left( \frac{2\mathbf{z}_i \mathbf{z}_j}{\mathbf{z}_i + \mathbf{z}_j} R_{ab}^2 \right). \quad (3.85)$$

També s'ha comprovat l'eficàcia de les *FMASA DF* en el càlcul de potencials electrostàtics, que és un clar exemple de les limitacions de les *PASA DF*. I finalment es presenta una aplicació de les *FMASA* en un estudi de *MQSM*. A partir d'unes divergències aparegudes en la bibliografia entre diferents grups experimentals sobre un possible intermedi d'una reacció d'inhibició de la glicosidasa, es planteja una anàlisi de semblança entre productes actius, no actius i el suposat intermedi de la reacció.



### Article 3.3

---

**Autors:** *Lluís Amat, Ramon Carbó-Dorca.*

**Títol:** *Molecular electronic density fitting using elementary Jacobi rotations under atomic shell approximation*

**Revista:** *Journal of Chemical Information and Computer Sciences*

**Volum:** 40      **Pàgines, inicial:** 1188    **final:** 1198    **Any:** 2000

---

# Molecular Electronic Density Fitting Using Elementary Jacobi Rotations under Atomic Shell Approximation

Lluís Amat and Ramon Carbó-Dorca\*

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

Received March 25, 2000

Fitted electron density functions constitute an important step in quantum similarity studies. This fact not only is presented in the published papers concerning quantum similarity measures (QSM), but also can be associated with the success of the developed fitting algorithms. As has been demonstrated in previous work, electronic density can be accurately fitted using the atomic shell approximation (ASA). This methodology expresses electron density functions as a linear combination of spherical functions, with the constraint that expansion coefficients must be positive definite, to preserve the statistical meaning of the density function as a probability distribution. Recently, an algorithm based on the elementary Jacobi rotations (EJR) technique was proven as an efficient electron density fitting procedure. In the preceding studies, the EJR algorithm was employed to fit atomic density functions, and subsequently molecular electron density was built in a promolecular way as a simple sum of atomic densities. Following previously established computational developments, in this paper the fitting methodology is applied to molecular systems. Although the promolecular approach is sufficiently accurate for quantum QSPR studies, some molecular properties, such as electrostatic potentials, cannot be described using such a level of approximation. The purpose of the present contribution is to demonstrate that using the promolecular ASA density function as the starting point, it is possible to fit ASA-type functions easily to the *ab initio* molecular electron density. A comparative study of promolecular and molecular ASA density functions for a large set of molecules using a fitted 6-311G atomic basis set is presented, and some application examples are also discussed.

## INTRODUCTION

Because of the expensive computational requirements to perform *ab initio* calculations over many-electron systems, namely, for calculations involving four-center integrals, approximated methodologies have become an important device of computational chemistry. For example, many quantum mechanical procedures often employ fitted electronic densities instead of *ab initio* ones. This is the case of some density functional approaches, which use fitted densities to reduce the formal scaling of four-index two-electron integrals.<sup>1–7</sup> However, other applications of fitted electron density functions have been investigated, for instance in the quantum similarity framework.<sup>8–37</sup> In this domain, the atomic shell approximation (ASA)<sup>21–27</sup> has been used in many instances for the evaluation of quantum similarity measures (QSM) in quantum QSPR applications.<sup>28–37</sup> This ASA approach constructs the electronic distribution of a given system as a linear combination of spherical Gaussian-type orbitals, restricting expansion coefficients to be positive definite. This last condition guarantees a probabilistic interpretation of the resultant density function. Recently, a simple but powerful technique based on an elementary Jacobi rotations (EJR) technique<sup>38</sup> in conjunction with a Taylor series expansion was developed to fit electronic density functions.<sup>24</sup> The EJR technique is a norm-conserving orthogonal transformation, and in consequence avoids the inclusion of a Lagrange multiplier to guarantee the proper

normalization of electronic density in the objective function. The constraint of preserving positive definite the expansion coefficients in the optimization procedure is merely accomplished by a simple transformation, consisting of defining ASA coefficients as the square module of an underlying variable set.

Nowadays, QSM has evidenced a considerable growth, especially for quantum QSPR studies where large molecular systems are studied.<sup>28–37</sup> This development has been feasible in part because of the utilization of ASA density functions, which allow fast QSM calculations. QSM applied to 3D QSAR analysis habitually assumes that the molecular superposition is defined on the QSM maximum.<sup>16</sup> This fact requires highly repeated computation of QSM and, as a consequence, the need to use a suitable approximation for the electron density function. A promolecular approach<sup>39</sup> was employed in early work to construct molecular ASA density functions. Within this mechanism molecular densities are expressed as sums of the contributions of free atoms. Then, once the atomic coordinates and a fitted atomic basis set are known, the density function of any molecule is easily built. This approximation has been proven to give satisfactory results in QSM applications.<sup>28–37</sup> In this way, some fitted atomic basis sets have been published and presented as a data set on the World Wide Web.<sup>40</sup> Among them, the most relevant are a fitted 3-21G basis set for atoms H to Kr,<sup>23</sup> a fitted Huzinaga basis set for atoms H to Rn,<sup>24</sup> a fitted 6-21G basis set for atoms H to Ar,<sup>25</sup> and a fitted 6-311G basis set for atoms H to Ar.<sup>26</sup> The main advantage of the promolecular approach is that it avoids the need to compute *ab initio*

\* To whom correspondence should be addressed. E-mail: director@iqc.udg.es. Phone: +34 972418357. Fax: +34 972418356.

density functions for the studied molecules. But this simple technique is not sufficient to determine some molecular properties, such as total atomic charges or electrostatic potentials. Moreover, this approach gives an inexact description of chemical bonds because it considers atomic densities constant within the molecular environment.

In the present work, a previously developed EJR fitting algorithm is applied to molecules. Promolecular density functions provide initial guess high-quality molecular electronic densities. This fact permits accurate fitted molecular ASA density functions to be achieved in a fast manner. This computational procedure could be a simple alternative to the previously proposed one,<sup>21,22</sup> which, using a sophisticated methodology, fits *ab initio* electron densities to ASA densities, while satisfying the constraint that expansion coefficients have to be definite positive. The fitted molecular electronic distributions obtained in this way furnish a more realistic description of molecules, permitting the primary electronic properties to be determined with sufficient accuracy.

Here, promolecular and fitted molecular ASA density functions are compared with respect to *ab initio* ones for a representative set of molecules. This analysis is performed by contrasting values of the quadratic error integral function, relative error in the electron–nuclei Coulomb attraction expectation energy, and relative error in overlap-like and Coulomb-like quantum self-similarity measures.

## THEORY AND METHOD

**Density Functions.** The *ab initio* first-order LCAO–MO electron density of a quantum molecular system can be generally written as

$$\rho(\mathbf{r}) = \sum_{\mu,\nu} D_{\mu\nu} \chi_{\mu}^*(\mathbf{r}) \chi_{\nu}(\mathbf{r}) \quad (1)$$

where  $\{D_{\mu\nu}\}$  are the elements of the charge–bond order matrix and  $\{\chi_{\mu}\}$  the atomic orbital basis set. At the same time the form of the density function over the MO set can be written as

$$\rho(\mathbf{r}) = \sum_i c_i |\varphi_i(\mathbf{r})|^2 \quad (2)$$

where  $\{c_i\}$  are the MO occupation numbers and  $\{\varphi_i(\mathbf{r})\}$  the MOs.

The use of fitted densities is very useful to overcome the evaluation of some bottleneck *ab initio* computations, such as four-center integrals. One of these methodologies consists of the ASA framework.<sup>21–27</sup> Formally similar to the MO density defined above and on the basis of various mathematical developments,<sup>17–20</sup> the ASA first-order density function is constructed as a simple linear combination of spherical Gaussian-type functions:

$$\rho^{\text{ASA}}(\mathbf{r}) = \sum_i w_i |s_i(\mathbf{r})|^2 \quad (3)$$

with the restriction that the expansion coefficients  $\{w_i\}$  have to be positive definite. A second constraint of  $\{w_i\}$  coefficients is related to the normalization condition of  $\rho(\mathbf{r})$ :

$$\int \rho^{\text{ASA}}(\mathbf{r}) \, d\mathbf{r} = 1 \quad (4)$$

Then, taking a basis set of normalized Gaussian functions,  $\{s_i\}$ , the expansion coefficients need to fulfill the additional condition

$$\sum_i w_i = 1$$

**EJR Fitting Algorithm for ASA Coefficients.** The EJR technique was originally employed to calculate electronic energies and wave functions,<sup>41–48</sup> although it can be successfully used in other optimization procedures, such as a fitting methodology. In this respect, EJR has been shown to be an efficient and accurate method for the calculation of fitted densities. The developed fitting algorithm was described in some detail in recent work.<sup>23,24</sup> Then, only a brief review of the main attributes which constitutes the basis of the present study will be examined. Additionally to the implementation of the EJR technique, the computational process was accelerated, expressing the cosine and sine functions of the EJR rotation angle as the first terms of a Taylor series.

The vector containing the ASA coefficients,  $\mathbf{w} = \{w_i\}$ , is computed by minimizing a quadratic error integral function between the *ab initio* and ASA density functions

$$\epsilon^{(2)} = \int |\rho(\mathbf{r}) - \rho^{\text{ASA}}(\mathbf{r})|^2 \, d\mathbf{r} \quad (5)$$

which can be expressed in a matrix notation as

$$\epsilon^{(2)} = \mathbf{Z} + \mathbf{w}^T \mathbf{Z} \mathbf{w} - 2\mathbf{b}^T \mathbf{w} \quad (6)$$

In eq 6  $\mathbf{Z}$  denotes an *ab initio* overlap-like quantum self-similarity measure (QS-SM), defined by the integral

$$\mathbf{Z} = \int |\rho(\mathbf{r})|^2 \, d\mathbf{r} = \sum_{\mu,\nu} D_{\mu\nu} \sum_{\lambda,\sigma} D_{\lambda\sigma} \int \chi_{\mu}^*(\mathbf{r}) \chi_{\nu}(\mathbf{r}) \chi_{\lambda}^*(\mathbf{r}) \chi_{\sigma}(\mathbf{r}) \, d\mathbf{r} \quad (7)$$

when  $\rho(\mathbf{r})$  is substituted by eq 1. The elements of the matrix  $\mathbf{Z} = \{Z_{ij}\}$  and the vector  $\mathbf{b} = \{b_i\}$  are given, respectively, by the integrals

$$Z_{ij} = \int |s_i(\mathbf{r})|^2 |s_j(\mathbf{r})|^2 \, d\mathbf{r} \quad (8)$$

and

$$b_i = \int |s_i(\mathbf{r})|^2 \rho(\mathbf{r}) \, d\mathbf{r} = \sum_{\mu,\nu} D_{\mu\nu} \int |s_i(\mathbf{r})|^2 \chi_{\mu}^*(\mathbf{r}) \chi_{\nu}(\mathbf{r}) \, d\mathbf{r} \quad (9)$$

To interpret  $\rho^{\text{ASA}}(\mathbf{r})$  as a probability distribution function, it is necessary that all atomic shell occupancies be positive-valued. This restriction is easily accomplished if a new set of coefficients is considered,  $\mathbf{x} = \{x_i\}$ , and the old ones are defined by the generating rule  $\forall i: w_i = |x_i|^2$ . The application of an EJR transformation,  $\mathbf{J}_{pq}(\alpha)$ , over the vector  $\mathbf{x}$  can be described by the equations

$$\begin{aligned} \dot{x}_p &\leftarrow cx_p - sx_q \\ \dot{x}_q &\leftarrow sx_p + cx_q \end{aligned} \quad (10)$$

where only the elements  $p$  and  $q$  are modified. The symbols  $c$  and  $s$  appearing in eq 10 determine the cosine and sine of the EJR rotation angle  $\alpha$ . The function to be optimized

corresponds to the variation of  $\epsilon^{(2)}$  with respect to the active pair of elements  $\{x_p, x_q\}$  after the EJR  $\mathbf{J}_{pq}(\alpha)$  is applied on vector  $\mathbf{x}$ :

$$\begin{aligned} \delta\epsilon^{(2)} = & \delta x_p^4 Z_{pp} + \delta x_q^4 Z_{qq} + 2\delta(x_p^2 x_q^2) Z_{pq} + \\ & 2\delta x_p^2 \sum_{i \neq p,q} x_i^2 Z_{pi} + 2\delta x_q^2 \sum_{i \neq p,q} x_i^2 Z_{iq} - 2b_p \delta x_p^2 - 2b_q \delta x_q^2 \end{aligned} \quad (11)$$

The  $\delta x_p^2$ ,  $\delta x_q^2$ ,  $\delta x_p^4$ ,  $\delta x_q^4$ , and  $\delta(x_p^2 x_q^2)$  terms are easily calculated (see ref 23), giving as a result a quadratic equation with respect to  $s$  and  $c$ :

$$\delta\epsilon^{(2)} = E_{04}s^4 + E_{13}cs^3 + E_{02}s^2 + E_{11}cs \quad (12)$$

To accelerate the computational process, the cosine and sine functions can be expressed as a Taylor series expansion up to second-order terms:

$$\begin{aligned} c = \cos(\alpha) &= 1 - 1/2\alpha^2 + \theta(\alpha^3) \\ s = \sin(\alpha) &= \alpha(1 - 1/6\alpha^2) + \theta(\alpha^4) \end{aligned} \quad (13)$$

Such an approach is efficient only when the objective function is studied near the minimum. Otherwise a more expensive algorithm based on an iterative procedure to find stationary points on eq 12, which was described in some detail in the first paper of this series (ref 23), has to be employed to obtain the optimal EJR angle.

Substitution of eq 13 into eq 12 yields

$$\delta\epsilon^{(2)} = A\alpha^3 + B\alpha^2 + C\alpha \quad (14)$$

where  $A = E_{13} - 2E_{11}/3$ ,  $B = E_{02}$ , and  $C = E_{11}$ . Taking a stationary point condition on eq 14

$$d\delta\epsilon^{(2)}/d\alpha = 3A\alpha^2 + 2B\alpha + C = 0 \quad (15)$$

and imposing the minimum condition determined by the second derivative, the optimal angle is

$$\alpha^* = \alpha_+ = (1/3A)[-B + (B^2 - 3AC)^{1/2}] \quad (16)$$

so the optimal cosine and sine are given by

$$\begin{aligned} c^* &\approx 1 - 1/2\alpha_+^2 \\ s^* &\approx \alpha_+(1 - 1/6\alpha_+^2) \end{aligned} \quad (17)$$

Substitution of eq 17 into eq 10 yields a simple expression for the coefficient set variation:

$$\begin{aligned} \dot{x}_p \leftarrow c^*x_p - s^*x_q &= x_p - \alpha_+x_q + 1/2\alpha_+^2(-x_p + 1/3\alpha_+x_q) \\ \dot{x}_q \leftarrow s^*x_p + c^*x_q &= x_q + \alpha_+x_p - 1/2\alpha_+^2(1/3\alpha_+x_p + x_q) \end{aligned} \quad (18)$$

which provides the optimal EJR approximate transformation.

**Fitting of ASA Exponents.** The atomic basis set  $\{s_i\}$  is determined by the exponents  $\{\zeta_i\}$ , which are optimized by minimizing the quadratic error integral function,  $\epsilon^{(2)}$ , with respect to this nonlinear parameter set. This procedure is essential to obtain accurate fitted density functions with a

small number of shells. Starting from a set of  $\{\zeta_i\}$  generated by means of an even-tempered sequence, the minimization of exponents is carried out with a Newton method, employing an analytic gradient and a Hessian of  $\epsilon^{(2)}$  with respect to  $\{\zeta_i\}$ .<sup>23</sup> This computational development is only applied in the atomic fitting procedure. In the construction of molecular ASA densities, no variation of atomic exponents is considered at all.

**Computational Process.** The objective function in the present fitting algorithm is the quadratic error integral defined in eq 5. There are two ASA independent parameter sets, coefficients  $\{w_i\}$  and exponents  $\{\zeta_i\}$ ; both are optimized in such a way that the objective function  $\epsilon^{(2)}$  is minimized. ASA coefficients are subjected to two simultaneous constraints: they are forced to be positive definite and to accomplish a normalization condition chosen as 1 for atoms and as the number of electrons,  $N_e$ , for molecules. The strategy adopted to construct ASA densities is described next.

The first operation consists of fitting an atomic basis set. The essential steps of the atomic fitting algorithm follow.

(1) Set  $n$  as the number of fitted atomic shells for a given atom  $a$ .

(2) Generate the initial ASA exponents as an even-tempered sequence:  $\zeta_i = \alpha\beta^i$ ,  $i = 1, \dots, n$ .

(3) Compute matrix  $\mathbf{Z}$  and vector  $\mathbf{b}$  from eqs 8 and 9, respectively.

(4) Calculate an initial set of expansion coefficients  $\{w_i\}$ :  $\mathbf{U}^+\mathbf{Z}\mathbf{U} = \mathbf{\Lambda}$ ;  $\mathbf{w} = \mathbf{u}^*$ .

(5) Optimize the set of exponents  $\{\zeta_i\}$  using a Newton algorithm to minimize  $\epsilon^{(2)}$ .

(6) Optimize the set of coefficients  $\{w_i\}$  using the previously described EJR algorithm with an expansion Taylor series procedure for rotation sine and cosine.

(7) If between two successive iterations the condition  $\Delta\epsilon^{(2)} < 10^{-6}$  is fulfilled, then stop; otherwise, go to step 5.

This procedure is repeated for an extended number of generating values  $\{\alpha, \beta\}$  of even-tempered series, to provide a sufficiently representative grid of starting sets for the exponents  $\{\zeta_i\}$ . The set of  $\{w_i\}$  and  $\{\zeta_i\}$  providing the lowest  $\epsilon^{(2)}$  is stored.

The next step after the generation of a fitted atomic basis set is the construction of molecular ASA densities. There is a significant point that must be kept in mind when a molecular density is fitted. This procedure needs a previously known molecular *ab initio* density function. Because this fact limits the application of fitted molecular densities, for example, in the practical use of QSM in QSAR analysis, where large molecular systems are studied, a promolecular formalism<sup>39</sup> has been employed in the generation of density functions. For a given fixed nuclear configuration, the promolecular ASA electronic distribution of a molecule  $A$  is built as a sum of discrete atomic densities, each centered on its own nucleus:

$$\begin{aligned} \rho_A^{\text{ASA}}(\mathbf{r}) &= \sum_{a \in A} P_a \rho_a^{\text{ASA}}(\mathbf{r}) = \\ & \sum_{a \in A} P_a \sum_{i \in a} w_i |s_i(\mathbf{r})|^2 = \sum_{i \in A} \omega_i |s_i(\mathbf{r})|^2 \end{aligned} \quad (19)$$

The coefficients  $P_a$  can be interpreted as the total electron density on atom  $a$ , usually approximated by the atomic

number  $Z_a$ . Equation 19 formed as a sum of free atom contributions produces a reasonable approximation to the molecular density.

On the other hand, fitting of ASA functions to molecular *ab initio* electron densities could be implemented in a way similar to that in the atomic case. The proposed procedure for molecules starts from the promolecular ASA density function. This procedure is similar to some self-consistent molecular calculations based on molecular wave functions expanded in terms of spherical Gaussians, which overlaps spherical atomlike charge distributions to generate molecular densities.<sup>1,4</sup> Beginning with this initial guess high-quality  $\rho_A^{\text{ASA}}(\mathbf{r})$ , only the set of ASA coefficients  $\{\omega_i\}$  is fitted to *ab initio*  $\rho_A(\mathbf{r})$ . The first step of the developed strategy for molecular fitting consists in using an unrestricted least-squares procedure to minimize the molecular  $\epsilon^{(2)}$  function. The least-squares fitting algorithm uses a Lagrange multiplier to guarantee charge conservation,<sup>1-7,11</sup> but does not restrict the expansion coefficients to be positive definite. In the present scheme, the least-squares fitting algorithm was only implemented as a preliminary and fast adjustment. If nonpositive coefficients are obtained, then the EJR algorithm described in the previous theoretical section is carried out using as the starting point the promolecular ASA density. Normally, if the number of fitted functions is not very large, the use of a least-squares procedure is appropriate, providing the same positive definite  $\{\omega_i\}$  coefficients which could be obtained using the EJR algorithm. For molecular fitting, the general strategy of applying expansion Taylor series to describe the cosine and sine functions is generally appropriate because the initial promolecular ASA set of  $\{\omega_i\}$  is close to the minimum of  $\epsilon^{(2)}$ .

**Molecular Integrals of Molecular Properties within the ASA Approach.** The ASA density functions have been applied to compute approximate integrals of some molecular properties. The main interest of the ASA formalism is the simplicity in the computation of molecular integrals, because they are based just on superposition of 1s Gaussian functions. In the present a normalized 1s GTO of the form

$$|s_i(\mathbf{r} - \mathbf{r}_a)|^2 = (2\zeta_i/\pi)^{3/2} \exp(-2\zeta_i|\mathbf{r} - \mathbf{r}_a|^2) \quad (20)$$

and centered at  $\mathbf{r}_a$  has been used. The first integral formula presented corresponds to the so-called overlap QS-SM, which appears in the computation of  $\epsilon^{(2)}$ . For a given quantum system A, this ASA integral is expressed as

$$\begin{aligned} Z_{AA} &= \sum_{i \in A} \sum_{j \in A} \omega_i \omega_j \int |s_i(\mathbf{r} - \mathbf{r}_a)|^2 |s_j(\mathbf{r} - \mathbf{r}_b)|^2 d\mathbf{r} \\ &= \sum_{i \in A} \sum_{j \in A} \omega_i \omega_j \left( \frac{2\zeta_i \zeta_j}{\pi(\zeta_i + \zeta_j)} \right)^{3/2} \exp\left( -\frac{2\zeta_i \zeta_j}{\zeta_i + \zeta_j} R_{ab}^2 \right) \quad (21) \end{aligned}$$

But the overlap-like QS-SM integral is not the unique formula to be used as QSM. In the Results and Discussion, Coulomb-like QSM is also used in the pairwise comparison of electron density functions of different molecules. The Coulomb-like QSM between two quantum systems A and B is defined as

$$Z_{AB}(\mathbf{r}_{12}^{-1}) = \int \int \rho_A(\mathbf{r}_1) \mathbf{r}_{12}^{-1} \rho_B(\mathbf{r}_2) d\mathbf{r}_1 d\mathbf{r}_2 \quad (22)$$

which is approximated within the ASA formalism to

$$\begin{aligned} Z_{AB}(\mathbf{r}_{12}^{-1}) &= \sum_{i \in A} \sum_{j \in B} \omega_i \omega_j \int |s_i(\mathbf{r}_1 - \mathbf{r}_a)|^2 \mathbf{r}_{12}^{-1} |s_j(\mathbf{r}_2 - \mathbf{r}_b)|^2 d\mathbf{r}_1 d\mathbf{r}_2 \\ &= 2 \sum_{i \in A} \sum_{j \in B} \omega_i \omega_j \left( \frac{2\zeta_i \zeta_j}{\pi(\zeta_i + \zeta_j)} \right)^{1/2} F_0\left( \frac{2\zeta_i \zeta_j}{\zeta_i + \zeta_j} R_{ab}^2 \right) \quad (23) \end{aligned}$$

where  $F_0(x)$  is the zeroth-order incomplete  $\gamma$  function.

Other integrals have been derived and used in this work. For instance the electron-nuclei Coulomb attraction energy,  $V(\mathbf{r})$ , is defined within the ASA approach as

$$\begin{aligned} V(\mathbf{r}) &= -\sum_c Z_c \int \frac{\rho_A^{\text{ASA}}(\mathbf{r} - \mathbf{r}_c)}{|\mathbf{r} - \mathbf{r}_c|} d\mathbf{r} \\ &= -\sum_{i \in A} \omega_i \sum_c Z_c \int \frac{|s_i(\mathbf{r} - \mathbf{r}_a)|^2}{|\mathbf{r} - \mathbf{r}_c|} d\mathbf{r} \\ &= -2 \sum_{i \in A} \omega_i (2\zeta_i/\pi)^{1/2} \sum_c Z_c F_0[2\zeta_i R_{ac}^2] \quad (24) \end{aligned}$$

and in a similar way, the molecular electrostatic potential of a molecule evaluated at the point  $\mathbf{r}_{H^+}$  is defined by

$$\begin{aligned} V_A(\mathbf{r}_{H^+}) &= \sum_c \frac{Z_c}{|\mathbf{r}_c - \mathbf{r}_{H^+}|} - \sum_{i \in A} \omega_i \int \frac{|s_i(\mathbf{r} - \mathbf{r}_a)|^2}{|\mathbf{r} - \mathbf{r}_{H^+}|} d\mathbf{r} \\ &= \sum_c \frac{Z_c}{|\mathbf{r}_c - \mathbf{r}_{H^+}|} - 2 \sum_{i \in A} \omega_i \left( \frac{2\zeta_i}{\pi} \right)^{1/2} F_0[2\zeta_i R_{aH^+}^2] \quad (25) \end{aligned}$$

## RESULTS AND DISCUSSION

**Fitted 6-311G Atomic Basis Set.** Considerable care has been employed to obtain a compact set of fitted atomic densities, because they constitute the basis for subsequent molecular studies. In this way, a set of ASA functions was fitted to an *ab initio* 6-311G basis set<sup>49,50</sup> for atoms H to Ar.<sup>26</sup> This study is an extension of previous work, where ASA densities were provided for other atomic basis sets.<sup>23-25</sup> Atomic *ab initio* RHF energies and density functions have been calculated using the ATOMIC program.<sup>51</sup> From these spherically symmetric electronic distributions the fitted ASA densities for a different number of atomic shells have been computed. Table 1 lists the fitting results for the atoms involved in the molecular examples studied later on, using for the adjustment from three to seven shells per atom. To assess the quality of the calculated basis set, Table 1 presents the quadratic error integral values  $\epsilon^{(2)}$ , as well as the errors encountered in the computation of the nuclear attraction potential  $V(\mathbf{r})$  and self-similarities  $Z_{aa}$  and  $Z_{aa}(\mathbf{r}_{12}^{-1})$  values. Coefficients and exponents for this basis set of ASA functions are available for downloading at a World Wide Web site.<sup>40</sup>

As is shown in Table 1, when the number of ASA functions  $n$  is increased, the  $\epsilon^{(2)}$  value quickly decreases. The present result agrees with early findings from a fitted Huzinaga basis set,<sup>24</sup> showing a general tendency to decrease relative errors in  $Z_{aa}$  and  $V(\mathbf{r})$  values as  $n$  increases, but not



**Table 1.** Fitting Results for the 6-311G Basis Set for Atoms H to Ar<sup>a</sup>

		three fitted atomic functions	four fitted atomic functions	five fitted atomic functions	six fitted atomic functions	seven fitted atomic functions
H	$\epsilon^{(2)}$	1.55E-05	4.27E-06	3.76E-07	1.21E-07	6.47E-08
	% $Z_{aa}$ <sup>b</sup>	-0.487	-0.362	0.014	0.007	-0.006
	% $V(\mathbf{r})$ <sup>b</sup>	-0.217	-0.220	0.026	0.023	-0.014
	% $Z_{aa}(\mathbf{r}_{12}^{-1})$ <sup>b</sup>	-1.010	-0.753	0.052	0.072	-0.041
B	$\epsilon^{(2)}$	1.91E-03	2.04E-04	4.18E-05	1.45E-05	1.36E-05
	% $Z_{aa}$	0.082	-0.070	-0.068	0.002	-0.024
	% $V(\mathbf{r})$	0.111	-0.267	-0.340	0.049	-0.082
	% $Z_{aa}(\mathbf{r}_{12}^{-1})$	-2.432	-1.754	-1.602	0.098	-0.412
C	$\epsilon^{(2)}$	2.38E-03	3.06E-04	9.84E-05	1.94E-05	1.78E-05
	% $Z_{aa}$	-0.001	-0.139	-0.133	-0.007	-0.008
	% $V(\mathbf{r})$	-0.109	-0.446	-0.507	0.046	0.038
	% $Z_{aa}(\mathbf{r}_{12}^{-1})$	-3.252	-2.514	-2.325	0.098	0.159
N	$\epsilon^{(2)}$	2.96E-03	4.76E-04	2.08E-04	2.41E-05	1.40E-05
	% $Z_{aa}$	-0.127	-0.233	-0.216	0.001	-0.015
	% $V(\mathbf{r})$	-0.317	-0.624	-0.663	0.008	-0.005
	% $Z_{aa}(\mathbf{r}_{12}^{-1})$	-3.956	-3.196	-2.936	-0.068	-0.091
O	$\epsilon^{(2)}$	3.74E-03	7.77E-04	2.45E-04	3.09E-05	1.77E-05
	% $Z_{aa}$	-0.291	-0.354	-0.029	-0.011	-0.011
	% $V(\mathbf{r})$	-0.566	-0.826	-0.106	0.003	-0.011
	% $Z_{aa}(\mathbf{r}_{12}^{-1})$	-4.835	-3.982	-0.818	-0.090	-0.131
F	$\epsilon^{(2)}$	4.71E-03	1.20E-03	2.79E-04	3.59E-05	2.87E-05
	% $Z_{aa}$	-0.474	-0.491	-0.028	-0.017	-0.014
	% $V(\mathbf{r})$	-0.780	-1.012	-0.094	-0.028	-0.033
	% $Z_{aa}(\mathbf{r}_{12}^{-1})$	-5.525	-4.643	-0.760	-0.211	-0.221
S	$\epsilon^{(2)}$	7.49E-02	5.35E-03	6.64E-04	2.25E-04	5.28E-05
	% $Z_{aa}$	1.961	0.120	-0.005	-0.018	-0.008
	% $V(\mathbf{r})$	0.626	0.125	-0.105	-0.133	-0.093
	% $Z_{aa}(\mathbf{r}_{12}^{-1})$	-6.792	-1.344	-1.011	-0.847	-0.499
Cl	$\epsilon^{(2)}$	7.88E-02	5.50E-03	6.45E-04	1.83E-04	5.71E-05
	% $Z_{aa}$	2.043	0.114	-0.007	-0.020	-0.012
	% $V(\mathbf{r})$	1.084	0.122	-0.109	-0.149	-0.113
	% $Z_{aa}(\mathbf{r}_{12}^{-1})$	-4.858	-1.279	-1.003	-0.893	-0.594

<sup>a</sup> Reproduced in part with permission from ref 26. Copyright 2000 Springer. <sup>b</sup> Percent error in  $Z_{aa}$ ,  $V(\mathbf{r})$ , and  $Z_{aa}(\mathbf{r}_{12}^{-1})$ .

in the same proportion than  $\epsilon^{(2)}$  varies. With respect to  $Z_{aa}(\mathbf{r}_{12}^{-1})$  values, the deviations between ASA and *ab initio* measures are superior to the one-electron properties  $Z_{aa}$  and  $V(\mathbf{r})$ , mainly for the first columns of Table 1, where the errors are significant. But for the basis set employing six and seven fitted functions per atom, the results ameliorate, and could be considered sufficiently accurate in view of the nature of the ASA approximation.

**Molecular Fitting.** Some illustrative results of promolecular and fitted molecular ASA density functions are presented. As has been already described in the computational process section, first the promolecular ASA density is constructed by means of eq 19, and then, using this as an initial guess,  $\rho_A^{\text{ASA}}(\mathbf{r})$ , the fitted molecular density, is computed. To simplify the terminology and facilitate the understanding of the present results, ASA calculations at various approximation levels are abbreviated as PASA for *promolecular* ASA densities, and FMASA for *fitted molecular* ASA densities. Additionally, two ASA basis sets have been used in the molecular examples. The notation (3/5/6) refers to an ASA basis set which uses three functions for the H atom, five functions for B, C, N, O, and F atoms, and six functions for S and Cl atoms, whereas in the ASA basis set (4/6/7) the rule of centering four functions on H, six functions on B, C, N, O, and F, and seven functions on S and Cl is followed. This basis set is constructed with the aim to generate sufficiently accurate density functions to be applied in QSM studies with the smallest number of expanded functions. With respect to molecular structures, fully optimized geometries computed using the GAUSSIAN 98

software package<sup>52</sup> have been considered for all studied compounds.

First, a molecular series of halomethane derivatives has been analyzed. *Ab initio* electronic structure calculations have been carried out at the HF level using a 6-311G basis set.<sup>49,50</sup> Table 2 lists the fitting results for these molecules computed for (3/5/6) and (4/6/7) ASA basis sets and using PASA and FMASA densities. The examined parameters for evaluating fitted density functions are the same as those employed for atoms:  $\epsilon^{(2)}$  and errors in nuclear attraction potential  $V(\mathbf{r})$  and self-similarities  $Z_{AA}$  and  $Z_{AA}(\mathbf{r}_{12}^{-1})$ . The values of  $\epsilon^{(2)}$  presented in Table 2 are normalized by the number of electrons, i.e., divided by  $N_e^2$ , to provide results comparable to the atomic fitting.

Several features can be spotted in the fitting results shown in Table 2. First, the high quality of the PASA density is confirmed. This is evidenced by the close values of  $\epsilon^{(2)}$  to the minimum provided by FMASA density. On the other hand, the use of fitted density functions yields reasonable one-electron properties, as is demonstrated by the remarkably small errors provided by FMASA densities in the computation of  $Z_{AA}$  and  $V(\mathbf{r})$  values. For instance, the errors in  $Z_{AA}$  values for FMASA(4/6/7) calculations are lower than 0.03%, whereas the errors in  $V(\mathbf{r})$  values for the same fitted functions are less than 0.15%. With respect to the two-electron property  $Z_{AA}(\mathbf{r}_{12}^{-1})$ , the errors in FMASA(4/6/7) calculations are less than 0.6% for the set of halomethane derivatives. Although there are high errors present in the Coulomb-like QS-SM evaluation, such parameters are precise enough to be used in QSM applications, avoiding the need to calculate an

**Table 2.** Fitting Results for Halomethane Molecules

		ASA(3/5/6)		ASA(4/6/7)				ASA(3/5/6)		ASA(4/6/7)	
		PASA	FMASA	PASA	FMASA			PASA	FMASA	PASA	FMASA
CH <sub>4</sub>	<i>n</i>	17	17	22	22	CCl <sub>4</sub>	<i>n</i>	29	29	34	34
	$\epsilon^{(2)}$	2.70E-04	6.00E-05	3.24E-04	5.87E-05		$\epsilon^{(2)}$	4.77E-05	4.72E-05	2.15E-05	2.06E-05
	% $Z_{AA}$ <sup>a</sup>	-1.488	-0.026	-1.298	0.015		% $Z_{AA}$	-0.033	-0.010	-0.027	-0.009
	% $V(\mathbf{r})$ <sup>a</sup>	0.981	-0.150	1.588	-0.023		% $V(\mathbf{r})$	-0.174	-0.167	-0.128	-0.134
	% $Z_{AA}(\mathbf{r}_{12}^{-1})$ <sup>a</sup>	2.788	-0.691	4.646	-0.073		% $Z_{AA}(\mathbf{r}_{12}^{-1})$	-0.782	-0.778	-0.508	-0.524
CFH <sub>3</sub>	<i>n</i>	19	19	24	24	CCIF <sub>2</sub> H	<i>n</i>	24	24	29	28
	$\epsilon^{(2)}$	2.82E-04	2.20E-04	2.26E-04	1.45E-04		$\epsilon^{(2)}$	1.14E-04	1.06E-04	6.93E-05	5.75E-05
	% $Z_{AA}$	-0.486	0.036	-0.446	0.030		% $Z_{AA}$	-0.097	0.004	-0.087	0.001
	% $V(\mathbf{r})$	0.220	-0.065	0.479	0.010		% $V(\mathbf{r})$	-0.169	-0.092	-0.075	-0.073
	% $Z_{AA}(\mathbf{r}_{12}^{-1})$	0.644	-0.521	1.660	0.016		% $Z_{AA}(\mathbf{r}_{12}^{-1})$	-0.725	-0.552	-0.246	-0.305
CClH <sub>3</sub>	<i>n</i>	20	20	25	25	CCl <sub>2</sub> FH	<i>n</i>	25	25	30	29
	$\epsilon^{(2)}$	1.34E-04	1.08E-04	8.70E-05	5.48E-05		$\epsilon^{(2)}$	8.16E-05	7.72E-05	4.47E-05	3.83E-05
	% $Z_{AA}$	-0.077	0.0001	-0.066	0.0003		% $Z_{AA}$	-0.057	-0.002	-0.049	-0.004
	% $V(\mathbf{r})$	-0.038	-0.050	0.052	-0.042		% $V(\mathbf{r})$	-0.159	-0.092	-0.093	-0.091
	% $Z_{AA}(\mathbf{r}_{12}^{-1})$	-0.309	-0.409	0.169	-0.197		% $Z_{AA}(\mathbf{r}_{12}^{-1})$	-0.736	-0.606	-0.364	-0.373
CF <sub>2</sub> H <sub>2</sub>	<i>n</i>	21	21	26	25	CCIFH <sub>2</sub>	<i>n</i>	22	22	27	26
	$\epsilon^{(2)}$	2.16E-04	1.87E-04	1.54E-04	1.15E-04		$\epsilon^{(2)}$	1.29E-04	1.15E-04	8.11E-05	6.24E-05
	% $Z_{AA}$	-0.364	0.040	-0.339	0.024		% $Z_{AA}$	-0.088	0.004	-0.078	0.0004
	% $V(\mathbf{r})$	-0.033	-0.066	0.163	-0.037		% $V(\mathbf{r})$	-0.098	-0.069	-0.006	-0.054
	% $Z_{AA}(\mathbf{r}_{12}^{-1})$	-0.174	-0.510	0.641	-0.171		% $Z_{AA}(\mathbf{r}_{12}^{-1})$	-0.504	-0.488	-0.019	-0.236
CCl <sub>2</sub> H <sub>2</sub>	<i>n</i>	23	23	28	28	CCl <sub>2</sub> F <sub>2</sub>	<i>n</i>	27	27	32	31
	$\epsilon^{(2)}$	8.45E-05	7.72E-05	4.56E-05	3.57E-05		$\epsilon^{(2)}$	7.63E-05	7.32E-05	4.19E-05	3.75E-05
	% $Z_{AA}$	-0.051	-0.002	-0.044	-0.003		% $Z_{AA}$	-0.061	-0.007	-0.054	-0.006
	% $V(\mathbf{r})$	-0.120	-0.086	-0.058	-0.074		% $V(\mathbf{r})$	-0.203	-0.151	-0.133	-0.108
	% $Z_{AA}(\mathbf{r}_{12}^{-1})$	-0.627	-0.540	-0.263	-0.322		% $Z_{AA}(\mathbf{r}_{12}^{-1})$	-0.855	-0.728	-0.482	-0.422
CF <sub>3</sub> H	<i>n</i>	23	23	28	27	CCIF <sub>3</sub>	<i>n</i>	26	26	31	30
	$\epsilon^{(2)}$	1.71E-04	1.53E-04	1.14E-04	9.10E-05		$\epsilon^{(2)}$	1.02E-04	9.52E-05	6.08E-05	5.24E-05
	% $Z_{AA}$	-0.316	0.037	-0.295	0.025		% $Z_{AA}$	-0.104	-0.002	-0.094	0.002
	% $V(\mathbf{r})$	-0.200	-0.076	-0.031	-0.052		% $V(\mathbf{r})$	-0.239	-0.140	-0.144	-0.082
	% $Z_{AA}(\mathbf{r}_{12}^{-1})$	-0.699	-0.518	0.012	-0.231		% $Z_{AA}(\mathbf{r}_{12}^{-1})$	-0.928	-0.693	-0.463	-0.331
CCl <sub>3</sub> H	<i>n</i>	26	26	31	31	CCl <sub>3</sub> F	<i>n</i>	28	28	33	33
	$\epsilon^{(2)}$	6.11E-05	5.86E-05	2.96E-05	2.60E-05		$\epsilon^{(2)}$	5.95E-05	5.81E-05	2.98E-05	2.76E-05
	% $Z_{AA}$	-0.041	-0.005	-0.034	-0.006		% $Z_{AA}$	-0.043	-0.009	-0.037	-0.008
	% $V(\mathbf{r})$	-0.156	-0.124	-0.104	-0.107		% $V(\mathbf{r})$	-0.184	-0.160	-0.129	-0.124
	% $Z_{AA}(\mathbf{r}_{12}^{-1})$	-0.743	-0.657	-0.433	-0.440		% $Z_{AA}(\mathbf{r}_{12}^{-1})$	-0.809	-0.755	-0.495	-0.482
CF <sub>4</sub>	<i>n</i>	25	25	30	29						
	$\epsilon^{(2)}$	1.44E-04	1.30E-04	9.32E-05	7.72E-05						
	% $Z_{AA}$	-0.293	0.020	-0.275	0.031						
	% $V(\mathbf{r})$	-0.327	-0.131	-0.173	-0.043						
	% $Z_{AA}(\mathbf{r}_{12}^{-1})$	-1.070	-0.659	-0.431	-0.208						

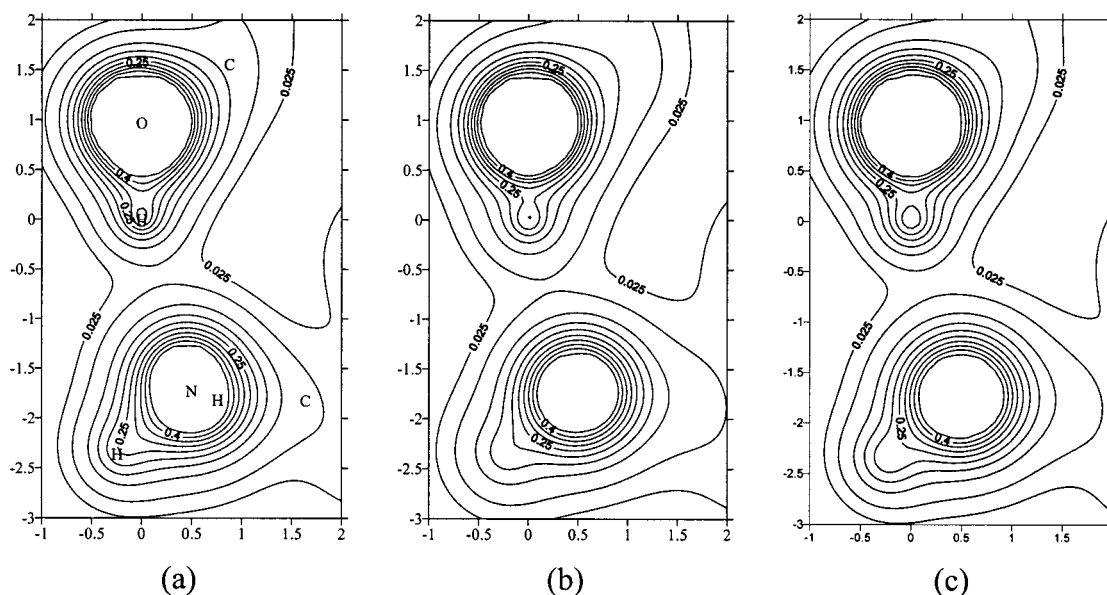
<sup>a</sup> Percent error in  $Z_{AA}$ ,  $V(\mathbf{r})$ , and  $Z_{AA}(\mathbf{r}_{12}^{-1})$ .

alternative ASA basis specifically fitted for Coulomb integrals. An other characteristic of the present results consists in that the errors in overlap-like QS-SM using FMASA densities always improve the values given by PASA. This trend confirms the high connection between  $\epsilon^{(2)}$  and  $Z_{AA}$  parameters, which is not obtained in  $V(\mathbf{r})$  and  $Z_{AA}(\mathbf{r}_{12}^{-1})$  values. Moreover, the deviations between ASA and *ab initio* measures of  $V(\mathbf{r})$  and  $Z_{AA}(\mathbf{r}_{12}^{-1})$  are generally higher than the error in  $Z_{AA}$  values. This fact is consistent with the atomic basis set results documented in Table 1.

In some of the examples presented in Table 2 for FMASA-(4/6/7) densities, the number of fitted functions has decreased with respect to the original PASA. The exclusion of some atomic shells occurs because of the fact that the associated molecular expansion coefficients become zero in the fitting process. It should be noted that the exponents of ASA density functions have been adapted for free atoms, and when these functions are employed in molecules, the saturation of the space of Gaussian functions is more quickly accomplished.

The second example presented for molecular systems concerns benzene and boron trichloride compounds. These molecules were previously fitted to ASA density functions from the *ab initio* HF/6-311G\*\* electronic distribution using

an elaborate methodology.<sup>21,22</sup> Basically this old technique could be considered as the limit of the ASA approach. This previous methodology is completely different from the one described here, because it starts from a spanned, nearly complete space of spherical functions, generated from even-tempered parameters. Then, the old algorithm<sup>21</sup> selects the optimal shells by minimizing  $\epsilon^{(2)}$ , imposing the constraint that expansion coefficients have to be definite positive. In the method of refs 21 and 22, the number of functions involved in the fitting process is very large in comparison with the present method, thus producing the highest computational cost. Table 3 documents the errors for the different approximations with respect to the *ab initio* HF/6-311G\*\* calculations. Fitting results for the nearly saturated basis set, summarized in the third column and denoted as limit ASA, are the best of all ASA approaches. On the other hand, the number of fitted functions for limit ASA is at least 2 times greater than the number of shells in the other ASA densities. Likewise, in the FMASA(4/6/7) fit of the benzene molecule, one function for each hydrogen atom has been eliminated because its expansion coefficient was zero, reducing *n* from 60 to 54. However, although the quadratic error values are improved, mainly for the BCl<sub>3</sub> molecule, the errors in self-



**Figure 1.** Isodensity contour maps for the intramolecular hydrogen bond of the GABA molecule. Density functions: (a) *ab initio*; (b) PASA(4/6/7); (c) FMASA(4/6/7).

**Table 3.** Fitting Results for Benzene and Boron Trichloride Molecules

	ASA(3/5/6)		ASA(4/6/7)		limit ASA <sup>a</sup>
	PASA	FMASA	PASA	FMASA	
C <sub>6</sub> H <sub>6</sub> <i>n</i>	48	48	60	54	132
$\epsilon^{(2)}$	8.21E-05	4.87E-05	8.99E-05	4.54E-05	3.91E-05
% Z <sub>AA</sub> <sup>b</sup>	-1.674	0.046	-1.545	0.028	0.039
BCl <sub>3</sub> <i>n</i>	23	23	27	27	66
$\epsilon^{(2)}$	5.55E-05	5.43E-05	2.24E-05	2.13E-05	8.02E-06
% Z <sub>AA</sub>	-0.049	-0.006	-0.041	-0.0014	0.0008

<sup>a</sup> From refs 21 and 22. <sup>b</sup> Percent error in Z<sub>AA</sub>.

similarity behave quite regularly between FMASA and limit ASA. Then, it is therefore plausible that the use of the present ASA densities instead of a nearly saturated basis set could be efficient enough in the computation of similarity measures.

**Isodensity Contour Maps.** In this section a representation of isodensity contour maps is shown for two molecules: the  $\gamma$ -aminobutyric acid (GABA) and the boron trichloride. This study is an extension of a comparative study of molecular density shape between *ab initio* and ASA densities which was recently reported.<sup>27</sup> The purpose is to graphically show the differences between the studied density functions.

GABA has important physiological effects because it is an inhibitory neurotransmitter in the mammalian central nervous system. The interest of the GABA molecule in this work comes from the intramolecular hydrogen bond, which could give this compound a certain conformation. Next, the description of the hydrogen bond is analyzed for different density functions. Figure 1 shows the isodensity contour maps for *ab initio* HF/6-311G, PASA(4/6/7) and FMASA(4/6/7) densities in the plane formed by the atoms O–H...N. These graphs are represented for values of the density function smaller than 0.5 au in the region of the intramolecular hydrogen bond. The main conclusion which could be deduced from Figure 1 is the high similarity between PASA and FMASA maps. On the other hand, the maximal divergences between ASA and *ab initio* densities are located

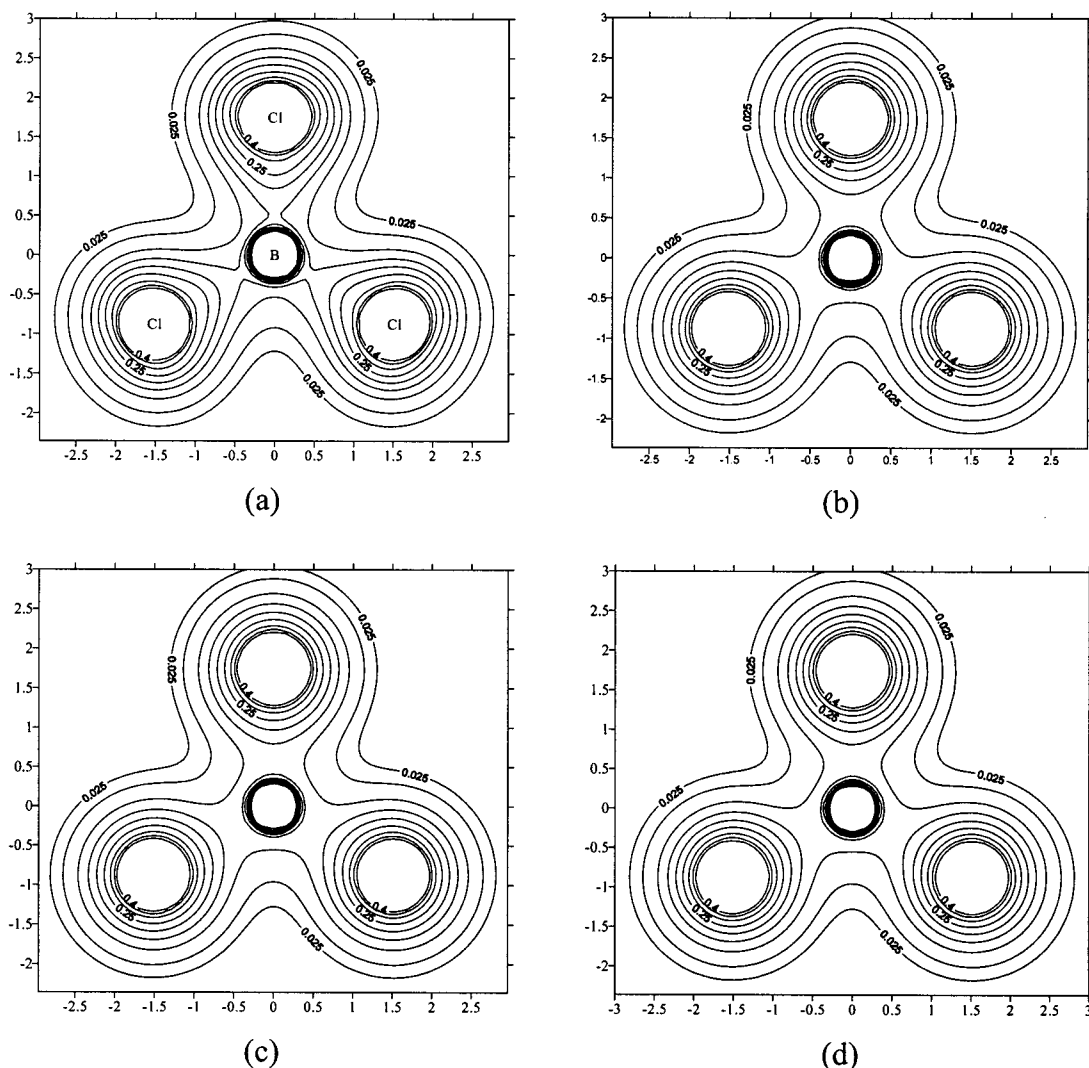
around the bond regions, in particular for the O–C and N–C bonds. This is not surprising since the PASA approach builds molecular electronic distributions from free atomic densities. Moreover, ASA densities describe inaccurately the significant regions of chemical bonds because they use only spherical functions centered on the nuclei. In contrast, it should be noted that there are not significant differences in the description of the hydrogen bond between ASA and *ab initio* graphs.

The second example of an isodensity contour map studies the boron trichloride molecule. The density adjustment for this molecule was studied in the previous section for three ASA approaches fitted to the *ab initio* HF/6-311G\*\* density. This example can help to analyze the effect of the number of fitted functions in the ASA densities. Figure 2 presents the isodensity contour maps for four densities: *ab initio*, PASA(4/6/7), FMASA(4/6/7), and limit ASA. By examining electronic density maps, the differences between the three ASA approaches are not significant. Clearly, from the present results the PASA approach could be conceived as an excellent fast computation of the fitted density function. Additionally, and as was noticed in the GABA maps, the major differences between *ab initio* and ASA densities are encountered around the bond regions.

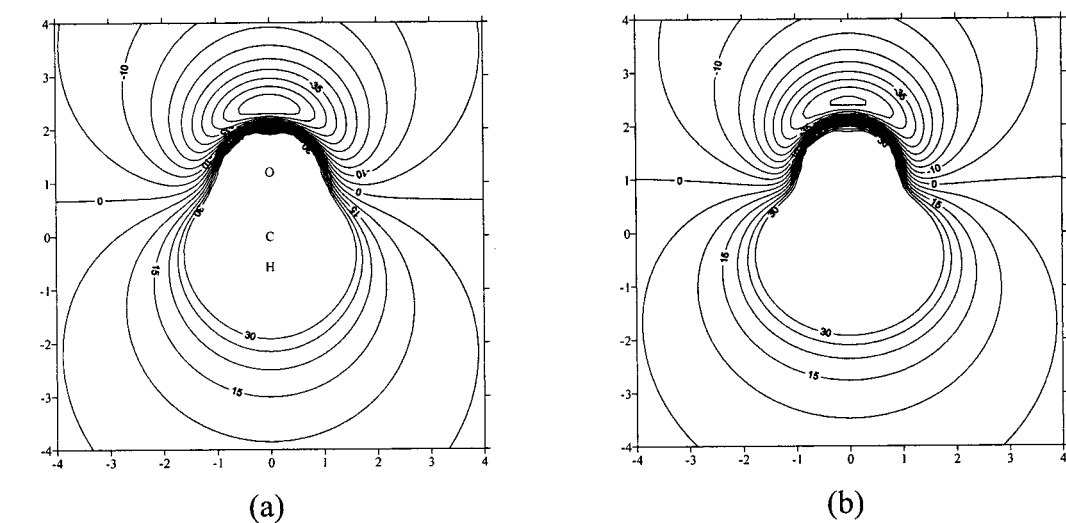
**Representation of Electrostatic Potentials in a 2D Contour Map.** Maps of electrostatic potential (MEPs)<sup>53,54</sup> permit information about molecular interactions to be qualitatively obtained, specifically to predict the most probable molecular sites susceptible to an electrophilic attack. In the present work, the electrostatic potential is presented as an example of a molecular property which is not described accurately enough by the PASA approach, and requires the use of FMASA densities to give tangible results. The simple construction of molecular densities by summing contributions of free atoms is not sufficient to describe the negative regions of the MEP, giving only zero or positive values.

A MEP for the ground state of the formaldehyde molecule is shown in Figure 3. Maps obtained from the  $V_A(\mathbf{r}_H^+)$  values and computed using *ab initio* and FMASA(4/6/7) densities





**Figure 2.** Isodensity contour maps for the boron trichloride molecule. Density functions: (a) *ab initio*; (b) PASA(4/6/7); (c) FMASA(4/6/7); (d) limit ASA.



**Figure 3.** Electrostatic potential contour maps for the  $\text{H}_2\text{CO}$  molecule. Computations for the ground state: (a) *ab initio* and (b) FMASA(4/6/7).

are represented in the plane perpendicular to the molecular plane containing the carbonyl group. Although for the ground state of the  $\text{H}_2\text{CO}$  molecule the similarity between the exact and approximated MEPs is appreciable, it must be pointed

out that ASA has some limitations in the description of MEPs. For instance, the first triplet excited state of the  $\text{H}_2\text{CO}$  molecule is poorly described even using FMASA densities. This is due to the ASA density structure based on

Chart 1. Molecular Structures of Mannosidase Inhibitors

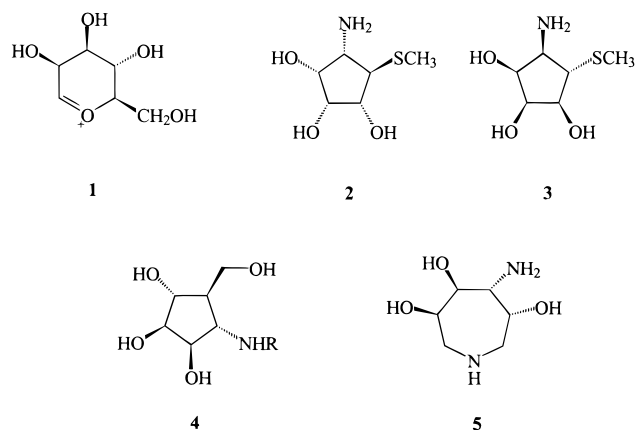


Table 4. Fitting Results for Mannosidase Inhibitors

	<i>n</i>	$\epsilon^{(2)}$	% $Z_{AA}^a$	% $Z_{AA}(\mathbf{r}_{12}^{-1})^a$
flap up mannopyranosyl cation	106	2.86E-05	0.030	-0.158
flap down mannopyranosyl cation	105	2.89E-05	0.031	-0.149
(+)-mannostatin A	115	1.84E-05	0.008	-0.219
(-)-mannostatin A	116	1.87E-05	0.008	-0.214
trihydroxycyclopentylamine	114	2.31E-05	0.024	-0.210
trihydroxyhexahydro-1H-azepine	120	2.16E-05	0.020	-0.212

<sup>a</sup> Percent error in  $Z_{AA}$  and  $Z_{AA}(\mathbf{r}_{12}^{-1})$ .

spherical symmetry functions, which impedes complete description of  $\pi$  MO contributions.

**QSM Application Example.** In this section, a simple application of QSM on molecules is presented. The example concerns some inhibitors of the  $\alpha$ -mannosidase. Glycosidase inhibitors are of interest for their diverse biological activities, for example, as antihyperglycemic compounds, inhibitors of tumor metastasis, or antivirals.<sup>55–58</sup> Chart 1 presents the structures of the studied mannosidase inhibitors. The interest to study these carbocyclic compounds by means of QSM comes from the fact that a controversy has appeared in the literature about their resemblance to the mannopyranosyl cation (**1**),<sup>55–58</sup> a proposed intermediate in the reaction catalyzed by the enzyme. Winkler and Holan<sup>55,56</sup> reported the (+)-mannostatin A (**2**), a potent inhibitor of glycoprotein processing, to have a high similarity to structure **1**. The discrepancy appeared when other authors<sup>57,58</sup> proposed the inactive mannosidase inhibitor (–)-mannostatin A (**3**) to be more similar to the mannopyranosyl cation structure than its enantiomeric form **2**. On the other hand, experimental synthesis of two additional compounds has contributed supplementary information: The trihydroxycyclopentylamine (**4**) compound was designed as a mimic of the mannopyranosyl cation structure, and was found to be a potent inhibitor.<sup>59</sup> This experimental finding validates the original proposed mannopyranosyl cation transition state. But an opposite reasoning

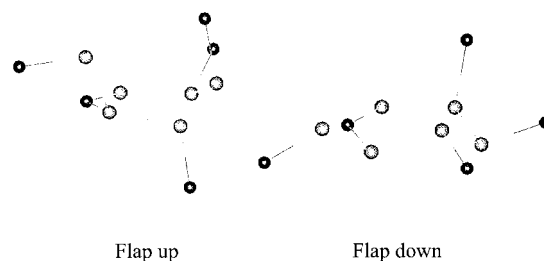


Figure 4. HF/6-311-optimized geometries for flap up and flap down forms of the mannopyranosyl cation. Hydrogen atoms have been omitted.

is evidenced with the trihydroxyhexahydro-1H-azepine (**5**) compound, which presents negligible mannosidase activity and was also synthesized as a mimic of the mannopyranosyl cation.<sup>60</sup>

Some preliminary considerations about molecular structures might be indicated prior to QSM analysis being carried out on this problem. As noted in the first paper of Winkler and Holan,<sup>55</sup> mannopyranosyl cation could present two-half-chair forms. Both conformations have been optimized here at the HF/6-311 level of theory, resulting in the finding that the “flap up” form is 4.0 kcal/mol lower in energy than the “flap down” form. Optimized geometries for both conformations are shown in Figure 4. Concerning compound **5**, the hexahydroazepine ring exhibits a high degree of flexibility, having various minimal conformational energy structures. In accordance with previous work of Farr and co-workers,<sup>60</sup> a conformation with the 4-amino group equatorial has been optimized, leading to more stable species than the one with the 4-amino group in the axial position.

The present QSM study for  $\alpha$ -mannosidase inhibitors is connected with previous work in which a possible quantification of the Hammond postulate was proposed, to evaluate the structural degree of the transition state advance with respect to the reactants and products by means of QSM.<sup>13,14</sup> Moreover, in another related work, QSM was employed to choose the optimal optimization methodology to construct molecular structures by comparing various theoretical 3D geometries with the experimental one obtained from X-ray analysis.<sup>24,29</sup> Furthermore, the present example shows a situation in which PASA densities appear unable to be used because of the difficulty in describing a delocalized cation within this approach.

After *ab initio* HF/6-311 electronic densities were computed over the studied molecular set, the adjustment of FMASA(4/6/7) densities was carried out. Table 4 documents the quadratic error integral value for the FMASA(4/6/7) densities together with errors in overlap-like and Coulomb-like QS-SM between the ASA and *ab initio* values for the  $\alpha$ -mannosidase inhibitors. Remarkable is the small  $\epsilon^{(2)}$  value for both half-chair forms of the cation **1**. Subsequent QSM analysis has been performed below using ASA measures of

Table 5. Carbó Index Values Used to Compare Mannosidase Inhibitors

	<b>1</b> (up)	<b>1</b> (down)	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>
flap up mannopyranosyl cation	1	0.890	0.849	0.876	0.889	0.882
flap down mannopyranosyl cation	0.890	1	0.883	0.863	0.888	0.911
(+)-mannostatin A	0.849	0.883	1	0.900	0.865	0.879
(–)-mannostatin A	0.876	0.863	0.900	1	0.880	0.882
trihydroxycyclopentylamine	0.889	0.888	0.865	0.880	1	0.920
trihydroxyhexahydro-1H-azepine	0.882	0.911	0.879	0.882	0.920	1

the Coulomb integral (22). The accuracy of these measures is assessed by the calculated  $Z_{AA}(\mathbf{r}_{12}^{-1})$  values, which are generally in good agreement with the *ab initio* ones, giving errors lower than 0.3% as shown in Table 4. Basically, for the usual QSM calculations this precision is sufficient.

Molecular superpositions have been optimized to obtain maximal QSM values.<sup>16</sup> The alignment search procedure is the most time-consuming part of the similarity studies, and necessarily requires the use of ASA densities. It may be worthwhile to mention that the obtained superpositions from QSM are not subjected to external manipulations nor presumptions but are given by the results of the maximal QSM algorithm. Table 5 contains the Carbó indexes, defined as  $C_{AB} = Z_{AB}(Z_{AA}Z_{BB})^{-1/2}$ , for the pairwise comparison of  $\alpha$ -mannosidase inhibitors. The Carbó index normalizes QSM in the interval {0,1}, indicating that with values closer to 1 there is a greater similarity between compared molecules. Several conclusions could be obtained from the results shown in Table 5. Basically only the first row of Table 5 needs to be analyzed, corresponding to the superpositions of mannosidase inhibitors with the lowest energy flap up half-chair form of the mannosyl cation. The best agreement with the flap up cation **1** is for the molecule **4**, confirming its synthesis as a mimic of mannosyl cation.<sup>59</sup> On the other hand, for the **2** and **3** dissenting molecules the calculated Carbó index for the inactive (–)-enantiomer,  $C_{AB} = 0.876$ , is somewhat larger than for the active (+)-enantiomer,  $C_{AB} = 0.849$ . Furthermore, (+)-mannostatin A superimposes better on the flap down mannosyl cation than on the flap up mannosyl cation. In light of the present calculations, the previous results where the (–)-mannostatin A is regarded as the enantiomer more similar to the flap up mannosyl cation can be confirmed, although it has to be pointed out that the small differences between Carbó indexes are not sufficient to discard in a strict sense the alternative hypothesis. The present results only indicate that the proposed intermediate in the hydrolysis of mannopyranosides by  $\alpha$ -mannosidase is not an adequate standard of resemblance to use as an indicator of mannosidase inhibition. This fact was demonstrated by the synthesis of compounds **4** and **5** as mimics of mannosyl cation, being active and inactive inhibitors, respectively.

## CONCLUSION

This work represents an attempt to apply the EJР technique to fit first-order molecular density functions. The capabilities of the EJР technique justify its use in electron density fitting, for both atoms and molecules. EJР provides a robust computational algorithm for fitting ASA coefficients to *ab initio* density functions. As demonstrated by the presented numerical tests, ASA densities provide quite noticeable agreement with *ab initio* HF method results.

ASA density functions have permitted the extension of QSM to real problems in pharmacological drug design. The main attribute of ASA electron density consists of the fact that expansion coefficients are maintained positive definite, preserving the statistical meaning of density function in the fitted structure. Obtained results demonstrate that ASA densities describe with high accuracy the molecular density shape without computational effort. The potential use of the fitted molecular ASA density functions in the evaluation of some molecular properties, such as electrostatic potentials,

may have several applications in computational chemistry. The promolecular approach has also been shown to provide correct density functions in several QSM analyses related to QSAR applications.

## ACKNOWLEDGMENT

The present work was supported in part by the Fundació Maria Francisca de Roviralta and a European commission contract, No. ENV4-CT97-0508. This research has been carried out using the CЕСSA and CEPBA resources, coordinated by C<sup>4</sup>.

## REFERENCES AND NOTES

- Baerends, E. J.; Ellis, D. E.; Ros, P. Self-consistent molecular Hartree–Fock–Slater calculations I. The computational procedure. *Chem. Phys.* **1973**, *2*, 41–51.
- Sambe, H.; Felton, R. H. A new computational approach to Slater's SCF- $X\alpha$  equation. *J. Chem. Phys.* **1975**, *62*, 1122–1126.
- Dunlap, B. I.; Connolly, J. W. D.; Sabin, J. R. On some approximations in applications of  $X\alpha$  theory. *J. Chem. Phys.* **1979**, *71*, 3396–3402.
- Delley, B.; Ellis, D. E. Efficient and accurate expansion methods for molecules in local density models. *J. Chem. Phys.* **1982**, *76*, 1949–1960.
- Andzelm, J.; Wimmer, E. Density functional Gaussian-type-orbital approach to molecular geometries, vibrations, and reaction energies. *J. Chem. Phys.* **1992**, *96*, 1280–1303.
- Gallant, R. T.; St-Amant, A. Linear scaling for the charge density fitting procedure of the linear combination of Gaussian-type orbitals density functional method. *Chem. Phys. Lett.* **1996**, *256*, 569–574.
- Goh, S. K.; St-Amant, A. Using a fitted electronic density to improve the efficiency of a linear combination of Gaussian-type orbitals calculation. *Chem. Phys. Lett.* **1997**, *264*, 9–16.
- Carbó, R.; Leyda, L.; 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.
- Carbó, R.; Domingo, L. LCAO-MO Similarity Measures and Taxonomy. *Int. J. Quantum Chem.* **1987**, *23*, 517–545.
- Carbó, R.; Calabuig, B. Quantum similarity measures, molecular cloud description, and structure-properties relationships. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 600–606.
- Mestres, J.; Solà, M.; Duran, M.; Carbó, R. On the calculation of *ab initio* quantum molecular similarities for large systems: fitting the electron density. *J. Comput. Chem.* **1994**, *15*, 1113–1120.
- Carbó, R.; Calabuig, B.; Vera, L.; Besalú, E. Molecular quantum similarity: theoretical framework, ordering principles, and visualization techniques. *Adv. Quantum Chem.* **1994**, *25*, 253–313.
- Solà, M.; Mestres, J.; Carbó, R.; Duran, M. Use of *ab initio* quantum molecular similarities as an interpretative tool for the study of chemical reactions. *J. Am. Chem. Soc.* **1994**, *116*, 5909–5915.
- Fradera, X.; Amat, L.; Torrent, M.; Mestres, J.; Constans, P.; Besalú, E.; Martí, J.; Simon, S.; Lobato, M.; Oliva, J. M.; Luis, J. M.; Andrés, J. L.; Solà, M.; Carbó, R.; Duran, M. Analysis of the changes on the potential energy surface of Menshutkin reactions induced by external perturbations. *J. Mol. Struct.: THEOCHEM* **1996**, *371*, 171–183.
- Besalú, E.; Carbó, R.; Mestres, J.; Solà, M. Foundations and recent developments on molecular quantum similarity. *Top. Curr. Chem.* **1995**, *173*, 31–62.
- Constans, 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.
- Carbó-Dorca, R. Tagged sets, convex sets and quantum similarity measures. *J. Math. Chem.* **1998**, *23*, 353–364.
- 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.* **1998**, *23*, 365–375.
- Carbó-Dorca, R. Fuzzy sets and Boolean tagged sets; vector semispaces and convex sets; quantum similarity measures and ASA density functions; diagonal vectors 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.
- Carbó-Dorca, R.; Besalú, E. A general survey of molecular quantum similarity. *J. Mol. Struct.: THEOCHEM* **1998**, *451*, 11–23.
- Constans, P.; Carbó, R. Atomic shell approximation: electron density algorithm restricting coefficients to positive values. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1046–1053.



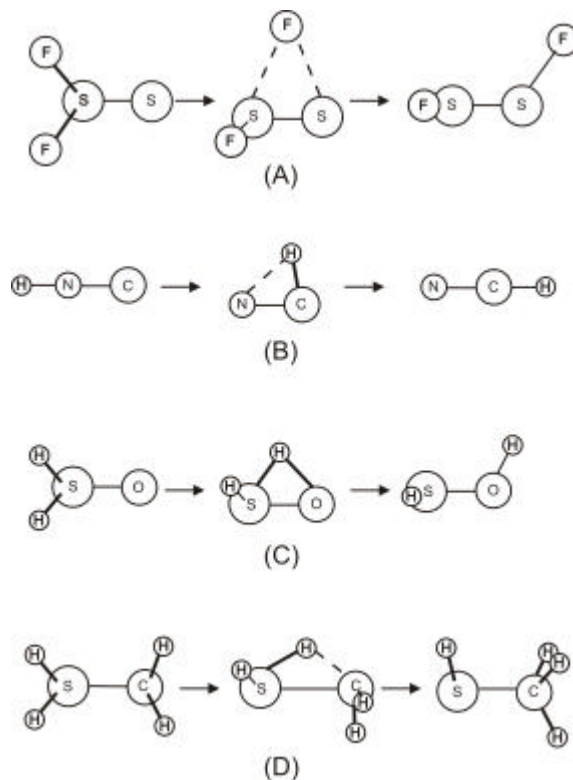
- (22) Constans, P.; Amat, L.; Fradera, X.; Carbó-Dorca, R. Quantum molecular similarity measures (QMSM) and the atomic shell approximation (ASA). In *Advances in Molecular Similarity*; Carbó-Dorca, R., Mezey, P. G., Eds.; JAI Press: Greenwich, CT, 1996; Vol. 1, pp 187–211.
- (23) Amat, L.; Carbó-Dorca, R. Quantum similarity measures under atomic shell approximation: first-order density fitting using elementary Jacobi rotations. *J. Comput. Chem.* **1997**, *18*, 2023–2039.
- (24) 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.
- (25) 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. In *The Fundamentals of Molecular Similarity*; Carbó-Dorca, R., Ed.; Kluwer Academic Press: Dordrecht, The Netherlands; in press.
- (26) Carbó-Dorca, R.; Robert, D.; Amat, L.; Gironés, X.; Besalú, E. Molecular quantum similarity in QSAR and drug design. *Lectures and Notes in Chemistry*; Springer: New York; 2000; Vol. 73.
- (27) Gironés, X.; Amat, L.; Carbó-Dorca, R. A comparative study of isodensity surfaces using *ab initio* and ASA density functions. *J. Mol. Graphics Modell.* **1998**, *16*, 190–196.
- (28) Lobato, M.; Amat, L.; Besalú, E.; 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.
- (29) 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.
- (30) Amat, L.; Carbó-Dorca, R.; Ponec, R. Molecular quantum similarity measures as an alternative to log P values in QSAR studies. *J. Comput. Chem.* **1998**, *19*, 1575–1583.
- (31) 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.
- (32) Robert, D.; Gironés, X.; Carbó-Dorca, R. Facet diagrams for quantum similarity data. *J. Comput.-Aided Mol. Des.* **1999**, *13*, 597–610.
- (33) Robert, D.; Carbó-Dorca, R. Aromatic compounds aquatic toxicity QSAR using quantum similarity measures. *SAR QSAR Environ. Res.* **1999**, *10*, 401–422.
- (34) 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*, 839–846.
- (35) Ponec, R.; Amat, L.; Carbó-Dorca, R. Molecular basis of quantitative structure-properties relationships (QSPR): a quantum similarity approach. *J. Comput.-Aided Mol. Des.* **1999**, *13*, 259–270.
- (36) Ponec, R.; Amat, L.; Carbó-Dorca, R. Quantum similarity approach to LFER: substituent and solvent effects on the acidities of carboxylic acids. *J. Phys. Org. Chem.* **1999**, *12*, 447–454.
- (37) Amat, L.; Carbó-Dorca, R.; Ponec, R. Simple linear QSAR models based on quantum similarity measures. *J. Med. Chem.* **1999**, *42*, 5169–5180.
- (38) Jacobi, C. G. J. Über ein leichtes Verfahren die in der theorie der Säcularstörungen vorkommenden gleichungen numerisch aufzulösen. *J. Reine Angew. Math.* **1846**, *30*, 51–94.
- (39) Ruedenberg, K.; Schwarz, W. H. E. Nonspherical atomic ground-state densities and chemical deformation densities from X-ray scattering. *J. Chem. Phys.* **1990**, *92*, 4956–4969.
- (40) ASA coefficients and exponents for different fitted atomic basis functions can be downloaded from the WWW site <http://iqc.udg.es/cat/similarity/ASA/basiset.html>.
- (41) Miller, K. J.; Ruedenberg, K. Electron correlation and separated-pair approximation. An application to Berylliumlike atomic systems. *J. Chem. Phys.* **1968**, *48*, 3414–3443.
- (42) Silver, D. M.; Mehler, E. L.; Ruedenberg, K. Electron correlation and separated pair approximation in diatomic molecules. I. Theory. *J. Chem. Phys.* **1970**, *52*, 1174–1180.
- (43) Mehler, E. L.; Ruedenberg, K.; Silver, D. M. Electron correlation and separated pair approximation in diatomic molecules. II. Lithium hydride and boron hydride. *J. Chem. Phys.* **1970**, *52*, 1181–1205.
- (44) Silver, D. M.; Ruedenberg, K.; Mehler, E. L. Electron correlation and separated pair approximation in diatomic molecules. III. Imidogen. *J. Chem. Phys.* **1970**, *52*, 1206–1227.
- (45) Carbó, R.; Domingo, L.I.; Peris, J. J. Elementary unitary MO transformations and SCF theory. *Adv. Quantum Chem.* **1982**, *15*, 215–265.
- (46) Carbó, R.; Domingo, L.I.; Peris, J. J.; Novoa, J. J. Energy variation and elementary Jacobi rotations. *J. Mol. Struct.: THEOCHEM* **1983**, *93*, 15–33.
- (47) Carbó, R.; Domingo, L.I.; Novoa, J. J. Multiconfigurational calculations using elementary Jacobi rotations. *J. Mol. Struct.: THEOCHEM* **1985**, *120*, 357–363.
- (48) Carbó, R.; Miró, J.; Domingo, L.I.; Novoa, J. J. Jacobi rotations: a general procedure for electronic energy optimization. *Adv. Quantum Chem.* **1989**, *20*, 375–441.
- (49) Krishnan, R.; Binkley, J. S.; Seeger, R.; Pople, J. A. Self-consistent molecular orbital methods. XX. A basis set for correlated wave functions. *J. Chem. Phys.* **1980**, *72*, 650–654.
- (50) McLean, A. D.; Chandler, G. S. Contracted Gaussian basis sets for molecular calculations. I. Second row atoms, Z=11–18. *J. Chem. Phys.* **1980**, *72*, 5639–5648.
- (51) Carbó-Dorca, R. ATOMIC Program 1995, based on Roos, B.; Salez, C.; Veillard, A.; Clementi, E. A general program for calculation of SCF orbitals by the expansion method. IBM Research/RJ518(No. 10901), 1968.
- (52) 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, A. 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, T.; Al-Laham, M. 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.
- (53) Bonaccorsi, R.; Scrocco, E.; Tomasi, J. Molecular SCF calculations for the ground state of some three-membered ring molecules: (CH<sub>2</sub>)<sub>3</sub>, (CH<sub>2</sub>)<sub>2</sub>NH, (CH<sub>2</sub>)<sub>2</sub>NH<sub>2</sub><sup>+</sup>, (CH<sub>2</sub>)<sub>2</sub>O, (CH<sub>2</sub>)<sub>2</sub>S, (CH)<sub>2</sub>CH<sub>2</sub>, and N<sub>2</sub>CH<sub>2</sub>. *J. Chem. Phys.* **1970**, *52*, 5270–5284.
- (54) See for example: Campanario, J. M.; Bronchalo, E.; Hidalgo, M. A. An effective approach for teaching intermolecular interactions. *J. Chem. Educ.* **1994**, *71*, 761–766.
- (55) Winkler, D. A.; Holan, G. Design of potential anti-HIV agents. 1. Mannosidase inhibitors. *J. Med. Chem.* **1989**, *32*, 2084–2089.
- (56) Winkler, D. A. Molecular modeling studies of “Flap Up” mannosyl cation mimics. *J. Med. Chem.* **1996**, *39*, 4332–4334.
- (57) Knapp, S.; Murali Dhar, T. G. Synthesis of the mannosidase II inhibitor mannosatin A. *J. Org. Chem.* **1991**, *56*, 4096–4097.
- (58) King, S. B.; Ganem, B. Synthetic studies on mannosatin A and its derivatives: a new family of glycoprotein processing inhibitors. *J. Am. Chem. Soc.* **1994**, *116*, 562–570.
- (59) Farr, R. A.; Peet, N. P.; Kang, M. S. Synthesis of 1S, 2R, 3S, 4R, 5R-methyl[2,3,4-trihydroxy-5-(hydroxymethyl)cyclopentyl]amine: a potent  $\alpha$ -mannosidase inhibitor. *Tetrahedron Lett.* **1990**, *31*, 7109–7112.
- (60) Farr, R. A.; Holland, A. K.; Huber, E. W.; Peet, N. P.; Weintraub, P. M. Pyrrolidine and hexahydro-1H-azepine mimics of the ‘flap up’ mannosyl cation. *Tetrahedron Lett.* **1994**, *50*, 1033–1044.

CI0000272



### 3.12 Classificació de camins de reacció mitjançant mesures de semblança

En col·laboració amb els professors D. L. Cooper i N. L. Allan de les universitats de Liverpool i Bristol respectivament, s'han estudiat algunes aplicacions de la semblança molecular quàntica. L'objectiu ha estat comparar les mesures de semblança basades en  $DF$  espacials de posició i de moment en diferents àmbits. En l'article [49] s'apliquen els estudis de semblança a la reactivitat, en concret s'analitzen els canvis de les mesures de moment i els índexs de semblança al llarg dels camins de reacció. A partir d'un estudi previ,<sup>50</sup> s'examina la variació d'ambdós mesures, valors esperats en l'espai  $p$  i  $MQSM$  en l'espai  $r$ , al llarg del camí de reacció de quatre reaccions simples de reordenació intramolecular:  $F_2S_2/FSSF$ ,  $HNC/NCH$ ,  $H_2SO/HSOH$  i  $H_2SCH_2/HSCH_3$ , que es mostren en la figura 3.1. Aquí només s'inclou la part del treball que fa referència a les  $MQSM$ , fent especial èmfasi a la comparació dels resultats obtinguts emprant densitats electròniques  $PASA$  i HF. Una altra de les finalitats de l'estudi és mostrar un exemple d'aplicació de les  $MQSM$  no relacionat amb  $QSAR$ .



**Figura 3.1** Les quatre reaccions de reordenació etiquetades de (A)–(D)

Una atenció particular se centra en la natura de l'estat de transició d'aquestes reaccions exotèrmiques. El postulat de Hammond,<sup>51</sup> molt utilitzat en química orgànica, diu que si l'estat de transició (*TS*) és proper en energia a un complex estable adjacent, llavors la seva estructura també és similar a la del complex. Per consegüent, una reacció exotèrmica s'ha d'esperar que el reactiu (*R*) sigui semblant al *TS*, mentre que un procés endotèrmic hauria d'estar caracteritzat per un producte (*P*) semblant al *TS*. Les reaccions de reordenació normalment involucren un nombre limitat de nuclis, i normalment s'espera que segueixin el postulat de Hammond. En les quatre reaccions discutides en l'article [49],  $F_2S_2/FSSF$  i  $HNC/NCH$  tenen un *TS* molt prematur i presenten un comportament tipus Hammond. Al contrari de la reacció  $H_2SO/HOH$ , que té un *TS* tardívol, i demostra les característiques anti-Hammond.<sup>50</sup> En aquest cas, el *TS* és estructuralment més similar al compost  $H_2SO$  que a  $HOH$ . Finalment, el *TS* per la reacció  $H_2SCH_2/HSCH_3$  presenta característiques intermèdies entre els dos tipus de comportament.

Les funcions d'ona han estat generades per *R*, *P* i *TS* de totes les reaccions presentades en la figura 3.1 emprant la teoria *SCF* HF amb el programa GAUSSIAN.<sup>43</sup> S'ha utilitzat el conjunt de funcions de base 6-31++G\*\*. Les energies HF dels *TS* i *P* referides als *R* es mostren en la taula 3.10. A més es llisten els valors del factor d'exotermicitat,  $\gamma$ ,<sup>52</sup> definit per Cioslowski com

$$g = (E_P - E_R) / (2E_{TS} - E_R - E_P) . \quad (3.86)$$

El factor  $\gamma$  pren valors entre +1 i -1, i és el quocient de la diferència i la suma de les energies d'activació de les reaccions directe i inversa. La reacció  $H_2SCH_2/HSCH_3$  és la més exotèrmica de les quatre i té el valor més negatiu de  $\gamma$ . El factor  $\gamma$  creix en magnitud al llarg de les series de reaccions (A)–(D), i no dóna cap indicació del comportament tipus Hammond o anti-Hammond.

	F <sub>2</sub> S <sub>2</sub> /FSSF	HNC/NCH	H <sub>2</sub> SO/HSOH	H <sub>2</sub> SCH <sub>2</sub> /HSCH <sub>3</sub>
$E_{TS}-E_R$	64.5	40.0	50.3	17.9
$E_P-E_R$	-9.6	-9.5	-32.9	-80.2
$\gamma$	-0.069	-0.106	-0.246	-0.691

**Taula 3.10.** Energies HF relatives (kcal/mol) i valors de  $\gamma$  per les quatre reaccions de reordenació estudiades.

De la mateixa manera que s'havia fet en l'estudi previ [50], s'utilitza la *MQSM* de tipus Coulomb entre dues molècules *A* i *B* definida en l'equació (2.14). Els valors de  $Z_{AB}$  depenen de la posició relativa en l'espai de les molècules *A* i *B*. La seva orientació mútua s'ha optimitzat de manera que es maximitza la *MQSM*. En el capítol 4 es donen detalls sobre el mètode de superposició emprat. El procés d'optimització es fa mitjançant funcions *PASA*, i posteriorment es fa un càlcul puntual *ab initio* sobre la superposició òptima. Les *PASA DF* s'han construït a partir de les funcions ASA atòmiques ajustades en la base 6-311G. S'han emprat 3 funcions per descriure l'H, 5 funcions per als àtoms C, O, F i N, i 6 funcions per al S.

Una vegada avaluades les *MQSM* de Coulomb  $Z_{R TS}$ ,  $Z_{R P}$  i  $Z_{P TS}$ , més els valors de les *QS-SM* per a cadascuna de les espècies:  $Z_{R R}$ ,  $Z_{P P}$  i  $Z_{TS TS}$ ; es calcula l'índex de Carbó, mitjançant equació (2.19):  $C_{R TS}$ ,  $C_{R P}$  i  $C_{P TS}$ . Els valors de  $C_{AB}$  són propers a la unitat com evidencien les tres primeres columnes de la taula 3.11. A més s'han calculat les distàncies euclidianes, a partir de l'equació (2.20):  $D_{R TS}$ ,  $D_{R P}$  i  $D_{P TS}$ . També s'examina la variació de dos paràmetres més,  $\alpha$  i  $\beta$ , definits per primera vegada per Cioslowski.<sup>52</sup> El paràmetre d'isosincronicitat,  $\alpha$ , ve donat per l'expressió

$$\alpha = (D_{R TS} + D_{P TS}) / D_{R P} , \quad (3.87)$$

i mai és inferior a 1. Pren un valor pròxim a 1 quan la distància des de *R* o *P* al *TS* és petita, o quan l'estat de transició perd semblança amb els reactius de la mateixa manera que guanya semblança amb els productes.



La proximitat estructural del *TS* amb els *R* ve donada per  $\beta$ , definit com

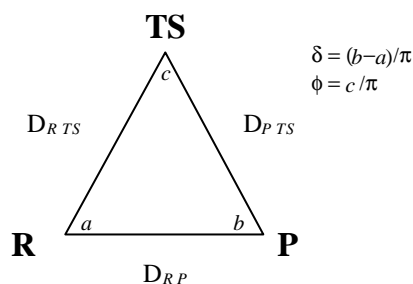
$$\beta = (D_{R\,TS} - D_{P\,TS})/D_{R\,P} \quad , \quad (3.88)$$

que pren els valors entre  $-1$  i  $1$ . Si *R* i *TS* són més propers que *P* i *TS*, llavors el *TS* és prematur i  $\beta < 0$ . Valors positius de  $\beta$  corresponen a un *TS* endarrerit.

Mesures de semblança HF										
Reacció	$C_{RP}$	$C_{R\,TS}$	$C_{P\,TS}$	$D_{RP}$	$D_{R\,TS}$	$D_{P\,TS}$	$\alpha$	$\beta$	$\delta$	$\phi$
F <sub>2</sub> S <sub>2</sub> /FSSF	0.925	0.939	0.916	14.36	12.94	15.17	1.96	-0.155	-0.086	0.338
HNC/NCH	0.997	0.998	0.997	0.849	0.815	0.894	2.01	-0.092	-0.051	0.330
H <sub>2</sub> SO/H <sub>2</sub> SOH	0.982	0.980	0.999	4.398	4.709	1.051	1.31	0.832	0.490	0.370
H <sub>2</sub> SCH <sub>2</sub> /HSCH <sub>3</sub>	0.990	0.962	0.967	3.165	6.206	5.743	3.78	0.146	0.091	0.169
Mesures de semblança PASA										
Reacció	$C_{RP}$	$C_{R\,TS}$	$C_{P\,TS}$	$D_{RP}$	$D_{R\,TS}$	$D_{P\,TS}$	$\alpha$	$\beta$	$\delta$	$\phi$
F <sub>2</sub> S <sub>2</sub> /FSSF	0.925	0.940	0.917	14.39	12.86	15.09	1.94	-0.155	-0.170	0.341
HNC/NCH	0.996	0.997	0.997	1.011	0.902	0.961	1.84	-0.059	-0.063	0.365
H <sub>2</sub> SO/H <sub>2</sub> SOH	0.982	0.980	0.999	4.407	4.654	1.079	1.30	0.811	0.924	0.390
H <sub>2</sub> SCH <sub>2</sub> /HSCH <sub>3</sub>	0.990	0.963	0.969	3.178	6.080	5.603	3.67	0.150	0.185	0.174

**Taula 3.11** Índexs de semblança calculats a partir de les MQSM de Coulomb, i valors dels paràmetres  $\alpha$ ,  $\beta$ ,  $\delta$  i  $\phi$

S'han definit dos nous paràmetres,<sup>49</sup>  $\delta$  i  $\phi$ , calculats a partir de les distàncies euclidianes  $D_{R\,TS}$ ,  $D_{P\,TS}$  i  $D_{R\,P}$ . Com mostra la figura 3.2, qualsevol reacció es pot representar mitjançant un triangle on els vèrtexs representen *R*, *P* i *TS*, i els costats són igual en allargada als valors de l'índex D. Triangles torçats cap a l'esquerra o a la dreta indiquen estats de transició matiners o endarrerits, respectivament. Una base relativament llarga suggereix grans canvis en la estructura al llarg de la reacció, i un triangle alt suggereix significants diferències estructurals entre *TS* i *R*, i *TS* i *P*. Es defineix  $\delta$  i  $\phi$  en termes dels angles en radians supeditats a *R*, *P* i *TS*, com mostra la figura 3.2.



**Figura 3.2** Definició dels paràmetres  $\delta$  i  $\phi$ .

El paràmetre  $\delta$ , que pren el valor en l'interval  $[-1,1]$ , és una mesura d'on se situa l'estat de transició, aviat o tard. Una reacció amb un *TS* aviat tindrà un valor negatiu de  $\delta$ . Per un valor donat de  $\delta$ , el valor de  $\phi$  ( $0 \leq \phi \leq 1$ ) indica quan lluny està l'estat de transició dels reactius/productes. La taula 3.11 conté els valors de  $\alpha$ ,  $\beta$ ,  $\delta$  i  $\phi$  relatius a les quatre reaccions. El paràmetre  $\delta$ , tant per les mesures HF com *PASA*, segueix l'ordre



*TS* aviat

*TS* tard

Els valors de  $\phi$  de les reaccions isoelectròniques  $H_2SCH_2/HSCH_3$  i  $H_2SO/HSOH$  reflecteixen quan diferent són els seus respectius *TS*, i la distància relativa dels *TS* a partir de *R* i *P*. En la reordenació  $H_2SCH_2/HSCH_3$ , la distància de l'enllaç S–C augmenta considerablement ( $>0.5 \text{ \AA}$ ) des del reactiu fins a l'estat de transició abans de contraure's marcadament fins a un valor final molt més proper a l'inicial.  $H_2SCH_2/HSCH_3$  és l'única de les quatre reaccions on la semblança entre *R* i *P* és més gran que la semblança entre *R* i *TS* o *P* i *TS*, i això és, en últim terme, l'origen del valor petit de  $\phi$ . Una variació semblant en la distància de l'enllaç rígid S–O no s'observa en la reacció  $H_2SO/HSOH$ .

La conclusió més important que s'extreu dels resultats presentats sobre les quatre reaccions (A)–(D) és que s'obté la mateixa tendència emprant mesures de semblança HF que *PASA*. En la taula 3.11 no s'aprecien diferències importants entre els valors *ab initio* i *PASA* dels índexs de Carbó i de les distàncies euclidianes, així com dels paràmetres  $\alpha$ ,  $\beta$  i  $\phi$ .

### 3.13 Ús de les funcions *PASA* per reduir el nombre de cicles *SCF*

Per acabar aquest capítol dedicat a les funcions *ASA*, s'ha inclòs un exemple d'aplicació de les densitats promoleculares en càlculs d'energia electrònica. Amb aquest treball és vol demostrar que les funcions *ASA* no tenen únicament aplicabilitat en la semblança quàntica, sinó que poden ser útils en altres camps de la química computacional. En concret s'ha desenvolupat una tècnica que permet generar matrius densitat inicials per al procés *SCF*. En l'article 3.4 es descriu el mètode proposat i es mostren els resultats sobre un ampli ventall d'estructures moleculars de diferents classes, incloent sistemes orgànics i complexes metàl·lics. El nombre de cicles necessaris per assolir els criteris de convergència en els càlculs d'energia electrònica són comparables o inferiors als obtinguts mitjançant altres metodologies.

La manera més usual de resoldre les equacions de Hartree-Fock-Roothaan definides en l'expressió (3.23) és mitjançant la notació matricial:

$$\mathbf{FC} = \mathbf{SCE}, \quad (3.89)$$

on els elements de  $\mathbf{F}$  es determinen mitjançant l'expressió (3.22),  $\mathbf{C}$  és la matriu formada per les incògnites, la matriu  $\mathbf{S}$  té els elements definits en la igualtat (3.18) i  $\mathbf{E}$  és una matriu diagonal amb els valors de les energies dels orbitals. No és un sistema lineal doncs els propis elements  $F_{mm}$  contenen les incògnites  $C_m$ . Consegüentment es comença postulant unes solucions raonables  $\mathbf{C}_0$  amb les quals es construeix una primera aproximació de la matriu de Fock,  $\mathbf{F}_0$ , que s'utilitza per generar uns orbitals millorats amb coeficients  $\mathbf{C}_1$ . Repetint successivament aquest procés fins a obtenir uns coeficients que no difereixen significativament del cicle anterior,  $\mathbf{C}_{n+1} \approx \mathbf{C}_n$ , s'aconsegueixen els anomenats orbitals autoconsistents de Hartree-Fock-Roothaan. El procediment descrit no sempre és convergent, només si s'utilitza una aproximació inicial convenient.

S'ha desenvolupat una tècnica que genera matrius densitat inicials aplicables en el càlcul iteratiu *SCF*, amb l'objectiu de reduir el nombre de cicles necessaris per assolir els criteris de convergència. S'ha proposat com a Hamiltonià inicial en el procés *SCF* la suma de la contribució monoelectrònica calculada a nivell *ab initio* més una estimació

de la repulsió bielectrònica resultant de la combinació de la *PASA DF* amb les funcions de base *ab initio*  $\{ \mathbf{c}_m \}$ . La matriu de Fock es pot definir com la suma

$$\mathbf{F} = \mathbf{H} + \mathbf{R}, \quad (3.90)$$

on els elements de la matriu  $\mathbf{H}$ , establerts en l'equació (3.15), són les contribucions monoelectròniques de l'energia, i els elements de  $\mathbf{R}$  es poden expressar a través dels elements de la matriu densitat (3.26) segons:

$$R_{mm} = \sum_I \sum_s D_{Is} \left[ (\mathbf{m}|\mathbf{l}s) - \frac{1}{2}(\mathbf{m}s|\mathbf{l}n) \right]. \quad (3.91)$$

Es proposa construir la matriu de Fock inicial d'acord amb l'equació

$$\mathbf{F}_0 = \mathbf{H} + \mathbf{R}_0, \quad (3.92)$$

on els elements de  $\mathbf{R}_0$  es defineixen per mitjà de les funcions *PASA* segons l'expressió

$$R_{0,mm} = \sum_a P_a \sum_{i \in a} w_i \left[ (\mathbf{m}|\mathbf{i}i) - \frac{1}{2}(\mathbf{m}|\mathbf{i}n) \right]. \quad (3.93)$$

La resolució de les integrals  $(\mathbf{m}|\mathbf{i}i)$  i  $(\mathbf{m}|\mathbf{i}n)$  involucra com a màxim tres centres.

En l'article 3.4 es presenten alguns exemples de la utilització dels coeficients  $\mathbf{C}_0$  obtinguts de la diagonalització de la matriu  $\mathbf{F}_0$  en qualitat de primera aproximació dels coeficients dels orbitals moleculars en el procés *SCF*. Els resultats indiquen que el nombre de cicles totals *SCF* és igual o inferior als valors que s'obtenen emprant l'aproximació que té per defecte el programa GAUSSIAN.<sup>43</sup> Les diferències més grans s'observen en compostos que contenen metalls de transició. A la pràctica s'ha procedit de la següent manera. Primer s'optimitzen les geometries de totes les molècules amb el mètode de HF i el conjunt de funcions de base 6-311G implementats en el programa GAUSSIAN. Dels fitxers de sortida del GAUSSIAN s'obté el conjunt de funcions de base  $\{ \mathbf{c}_m \}$ , i les matrius  $\mathbf{S}$  i  $\mathbf{H}$ . Llavors s'ha codificat un programa informàtic que calcula les integrals (3.93) amb les fórmules donades en la referència [53], i determina els elements de la matriu  $\mathbf{F}_0$ . Seguidament es calcula la descomposició de Cholesky de la matriu  $\mathbf{S}$ . Al ser una matriu definida positiva,  $\mathbf{S} > 0$ , existeix una matriu triangular

superior,  $\mathbf{T}$ , tal que  $\mathbf{S} = \mathbf{T}^T \mathbf{T}$ . El càlcul de la inversa de la matriu triangular,  $\mathbf{T}^{-1}$ , és fàcil i permet determinar la inversa de la mètrica conforme a

$$\mathbf{S}^{-1} = \mathbf{T}^{-1} \mathbf{T}^{-T} \quad (3.94)$$

Llavors es fa la transformació dels coeficients dels orbitals moleculars

$$\mathbf{C}' = \mathbf{T} \mathbf{C} \quad \wedge \quad \mathbf{C} = \mathbf{T}^{-1} \mathbf{C}', \quad (3.95)$$

la qual cosa permet definir l'equació de HF en notació matricial com

$$\mathbf{F}_0 \mathbf{T}^{-1} \mathbf{C}' = \mathbf{S} \mathbf{T}^{-1} \mathbf{C}' \mathbf{E}. \quad (3.96)$$

Multiplicant per  $\mathbf{T}^T$  l'equació (3.96),

$$(\mathbf{T}^T \mathbf{F}_0 \mathbf{T}^{-1}) \mathbf{C}' = (\mathbf{T}^T \mathbf{S} \mathbf{T}^{-1}) \mathbf{C}' \mathbf{E}, \quad (3.97)$$

definint  $\mathbf{F}'_0 = \mathbf{T}^T \mathbf{F}_0 \mathbf{T}^{-1}$  i tenint en compte que  $\mathbf{T}^T \mathbf{S} \mathbf{T}^{-1} = \mathbf{T}^T \mathbf{T}^T \mathbf{T} \mathbf{T}^{-1} = \mathbf{I}$ , s'obté

$$\mathbf{F}'_0 \mathbf{C}' = \mathbf{C}' \mathbf{E}, \quad (3.98)$$

que és una equació de valors i vectors propis. Tot seguit es diagonalitza  $\mathbf{F}'_0$  i es desfà el canvi de base (3.95) per obtenir els coeficients  $\mathbf{C}$ . La sortida del programa és un fitxer d'entrada del programa GAUSSIAN on s'especifica la matriu  $\mathbf{C}$  que s'utilitzarà com a primera aproximació dels coeficients dels orbitals moleculars en el procés SCF.

#### Article 3.4

---

**Autors:** Lluís Amat, Ramon Carbó-Dorca.

**Títol:** *Use of promolecular ASA density functions as a general algorithm to obtain starting MO in SCF calculations*

**Revista:** *International Journal of Quantum Chemistry*

**Volum:** 87      **Pàgines, inicial:** 59    **final:** 67    **Any:** 2002

---

---

# Use of Promolecular ASA Density Functions as a General Algorithm to Obtain Starting MO in SCF Calculations

---

LLUÍS AMAT, RAMON CARBÓ-DORCA

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

*Received 29 June 2001; accepted 21 September 2001*

---

**ABSTRACT:** Atomic shell approximation (ASA) constitutes a way to fit first-order density functions to a linear combination of spherical functions. The ASA fitting method makes use of positive definite expansion coefficients to ensure appropriate probability distribution features. The ASA electron density is sufficiently accurate for the practical implementation of quantum similarity measures, as was proved in previous published work. Here, a new application of the ASA density formalism is analyzed, and employed to obtain an initial guess of the density matrix for SCF procedures. The number of cycles needed to assess the convergence criterion in electronic energy calculations appears comparable to or less than those obtained by other means. Several molecular structures of different classes, including organic systems and metal complexes, were chosen as representative test cases. In addition, an ASA basis set for atoms Sc-Kr fitted to an ab initio 6-311G basis set is also presented. © 2002 John Wiley & Sons, Inc. *Int J Quantum Chem* 87: 59–67, 2002

**Key words:** density function; ASA; PASA; LCAO MO; SCF

---

## Introduction

The systematic study of large molecular systems, which can deal with the realistic simulation of biological compounds, has motivated the development of approximations sufficiently accurate

*Correspondence to:* R. Carbó-Dorca; e-mail: director@iqc.udg.es.

Contract grant sponsor: CICYT project.

Contract grant number: SAF2000-223.

Contract grant sponsor: Fundacio Maria Francisca de Roviralta.

as to be able to replace ab initio computations. This is the case of molecular quantum similarity measures, where in order to compute precise measure values, it has been used as a simple, yet accurate approximation to the molecular density function, based on a sum of spherically symmetric atomic densities. Such a theoretical model of electron density fitting has been named atomic shell approximation (ASA) [1–7]. Furthermore, when dealing with molecules, a promolecular description of the charge density distribution was employed, based on the sum of atomic ASA densities, which were previously fitted to an ab initio atomic basis set.

Although the schemes which provide a partition of a given molecule into its atomic components are not possible to derive from the quantum mechanical postulates, the concept of a promolecule has been used in many theoretical electron density distribution analyses. For instance, the definition of density deformation maps was based on subtracting from the electron density of the molecule the electron density of the promolecule [8–10]. These theoretical schemes have been useful in obtaining chemical information belonging to bonding interactions between atoms. The promolecular ASA (PASA) density function represents the atoms in a molecule as neutral entities of spherical shape, with a radial dependence equal to the isolated atoms. This approximate quantum mechanical development avoids costly molecular *ab initio* calculations and provides a sufficient precise three-dimensional electron distribution for several purposes, among them, the molecular quantum similarity evaluation of QSAR models, involving large molecular systems and needing the optimization of the relative position of the corresponding molecular pairs [7, 11–18].

In this work, a new application of PASA density functions is presented, related to a general and homogeneous generation of an initial guess of the density matrix in SCF calculations. This useful tool constitutes a quantum mechanical development providing the connection between ASA density function structure and the SCF-LCAO-MO approach. The present methodology can be applied to the study of large molecular systems, such as proteins or can be used in the simulation of liquid solutions, where the effects of the surrounding space would be approximated by ASA densities, and the active molecule would be processed at the *ab initio* level. Here, in order to assess the reliability of the procedure, PASA densities are used to compute initial two-electron repulsion integrals, which are appropriately added to the core Hamiltonian matrix to obtain a first guess of the Fock operator, which is subsequently diagonalized to construct an initial MO set for the SCF calculation. The number of SCF cycles needed to achieve the requested convergence in electronic energy calculations are compared with two alternative approaches: the use of the core Hamiltonian matrix only, and the use of CNDO [19], INDO [20], or Extended Hückel (EHT) [21] Hamiltonians.

This article takes account of the main concepts for the generation of starting MO in SCF calculations, and includes the application of the methodology

to several compounds involving atoms up to the fourth period. First, the procedure was tested over a set of small organic molecules, obtaining very similar results with respect to existing methods, except for certain compounds containing transition metals. In view of this fact, the novel algorithm to obtain an initial MO guess has been applied to a set of experimental transition metal complexes extracted from the Cambridge Structural Database (CSD) [22], improving considerably the results obtained from other techniques. The article also contains an Appendix which covers recent advances in the ASA fitting of atomic basis sets. The main steps of the developed algorithm, based on elementary Jacobi rotations techniques [23], are described, and ends with the results for the ASA fitting of atoms Sc-Kr with a 6-311G *ab initio* basis set.

---

## Theoretical Framework

### DENSITY FUNCTION

In the LCAO-MO approximation, the electronic density function of a closed-shell system is constructed as:

$$\rho(\mathbf{r}) = \sum_{\mu,\nu} D_{\mu\nu} \chi_{\mu}^*(\mathbf{r}) \chi_{\nu}(\mathbf{r}) \quad (1)$$

where  $\{D_{\mu\nu}\}$  are the elements of the charge-bond order matrix, defined from the MO coefficients as  $D_{\mu\nu} = 2 \sum_k^{\text{occ}} C_{\mu k}^* C_{\nu k}$ , and  $\{\chi_{\mu}\}$  the AO basis set.

Within the ASA approach [1–7], electronic densities are expressed as a linear combination of spherical *1s* functions:

$$\rho^{\text{ASA}}(\mathbf{r}) = \sum_i w_i |s_i(\mathbf{r})|^2, \quad (2)$$

where the coefficients  $\{w_i\}$  are restricted to be positive definite in order to assure the physical meaning of the fitted density. A simple but very interesting particular case of the previous development consists in using a promolecular approximation. In the PASA model, the molecular electronic density is simply a sum of independent atomic contributions,

$$\rho_A^{\text{PASA}}(\mathbf{r}) = \sum_{a \in A} P_a \rho_a^{\text{ASA}}(\mathbf{r}) = \sum_{a \in A} P_a \sum_{i \in a} w_i |s_i(\mathbf{r})|^2, \quad (3)$$

where the coefficients  $P_a$  are the total density on atom *a*, usually approximated by the atomic number  $Z_a$ , and  $\rho_a^{\text{ASA}}(\mathbf{r})$  the atomic densities built up using expression (2), where the coefficients are also normalized to one, that is,  $\sum_{i \in a} w_i = 1$ . The set of coefficients and exponents of the ASA expansion are

calculated by minimizing a quadratic error integral function between the exact and ASA atomic density functions [4, 5]. With the parameterized atomic densities  $\rho_a^{\text{ASA}}(\mathbf{r})$ , and the nuclear coordinates of any molecular system, PASA density is easily generated by using Eq. (3), producing a reasonable approximation to the molecular density, avoiding the ab initio theoretical calculation of electron density, and permitting the quantum similarity study of large molecular systems such as enzymes [3].

### INITIAL GUESS FOR THE HARTREE-FOCK WAVEFUNCTION

The SCF procedure for a closed-shell one determinant system is based on the solution of the generalized secular equation,

$$\mathbf{FC} = \mathbf{SCE}, \quad (4)$$

which is expressed in matrix form, and where  $\mathbf{F}$  is the Fock matrix,  $\mathbf{C}$  contains the molecular expansion coefficients,  $\mathbf{S}$  is the overlap matrix and  $\mathbf{E}$  is a diagonal matrix of the MO energies. The Fock matrix is defined as

$$\mathbf{F} = \mathbf{K} + \mathbf{V} + \mathbf{R}(\mathbf{C}), \quad (5)$$

where  $\mathbf{K}$  is the kinetic energy matrix,  $\mathbf{V}$  is the nuclear attraction potential matrix and  $\mathbf{R}(\mathbf{C})$  corresponds to the electronic repulsion and depends on the expansion coefficients. The elements of the two-electron repulsion matrix in this case are expressed using the charge and bond order matrix elements as

$$R_{\mu\nu} = \sum_{\lambda\sigma} D_{\lambda\sigma} \left[ (\mu\nu|\lambda\sigma) - \frac{1}{2}(\mu\lambda|\sigma\nu) \right] \quad (6)$$

It is well-known [24], that the Fock matrix depends on the expansion coefficients  $\mathbf{C}$ , and thus, Eq. (4) is nonlinear and has to be solved in an iterative manner. Consequently, an initial guess at the eigenvector matrix  $\mathbf{C}$  has to be obtained to start the SCF procedure. An immediate possible general procedure consists of using only the one-electron, core-Hamiltonian matrix,

$$\mathbf{F}_0 = \mathbf{H} = \mathbf{K} + \mathbf{V}. \quad (7)$$

Some approximations were developed to improve this initial step of the SCF procedure, and decrease as well the number of iterations. Among them, we shall mention the default option of GAUSSIAN program [25], where an INDO guess is used for first-row systems [19], CNDO for second-row [20], and EHT for the rest [21].

Here, it is proposed to use PASA density in combination with ab initio MO to compute an initial Fock matrix,

$$\mathbf{F}_0 = \mathbf{H} + \mathbf{R}_0, \quad (8)$$

where the elements of the approximated electronic repulsion matrix will be defined as

$$R_{0,\mu\nu} = \sum_a P_a \sum_{i \in a} w_i \left[ (\mu\nu|ii) - \frac{1}{2}(\mu i|iv) \right] \quad (9)$$

where  $P_a$  and  $w_i$  are defined in turn as in Eq. (3), and the repulsion integrals  $(\mu\nu|ii)$  and  $(\mu i|iv)$  are computed using a blend of AO set and the ASA set of functions. That is,

$$\begin{aligned} (\mu\nu|ii) &= \iint \chi_\mu^*(\mathbf{r}_1) \chi_\nu(\mathbf{r}_1) \mathbf{r}_{12}^{-1} |s_i(\mathbf{r}_2)|^2 d\mathbf{r}_1 d\mathbf{r}_2, \\ \text{and } (\mu i|iv) &= \iint \chi_\mu^*(\mathbf{r}_1) s_i(\mathbf{r}_1) \mathbf{r}_{12}^{-1} s_i(\mathbf{r}_2) \\ &\quad \times \chi_\nu(\mathbf{r}_2) d\mathbf{r}_1 d\mathbf{r}_2, \end{aligned} \quad (10)$$

$\{\chi_\mu\}$  being the AO basis set functions. Because any ASA density function is built as a linear combination of 1s Gaussian functions, the mixed up integrals in Eq. (9) are not really time-consuming. The initial integrals do not need to be stored, and they involve three centers as a maximal order. The present approach can be easily generalized to deal with initial density matrix guesses in simple open-shell RHF calculations. The initial MO treatment may be similar as described above. For instance, in multiplet states, MOs are described as consisting of two sets of functions: a closed-shell MO set occupied by two electrons, and an open-shell set containing all the singly occupied MOs [26]. In this case, both open- and closed-shell MO sets can be constructed from an initial Fock matrix obtained as in Eqs. (8) and (9).

### RESULTS

Before presenting the results of our calculations, it seems appropriate to specify the followed computational procedure. All molecules have been fully optimized at the HF level using the 6-31G basis set implemented in the GAUSSIAN program [25]. Then, three SCF calculations with different initial Fock matrices are run over the optimized molecular geometry of each molecule to determine the number of cycles: (i) the core Hamiltonian, which corresponds to the keyword *core* in the GAUSSIAN program; (ii) the default method in the GAUSSIAN program, which requests INDO, CNDO or EHT



Hamiltonians depending on the involved atoms; and (iii) force as initial MO the vectors obtained from the diagonalization of a Fock matrix as defined in Eq. (8). In all the computations, the SCF convergence criterion on root mean square density matrix differences is fixed to  $10^{-8}$  in order to obtain the same final electronic energy. In the last case, the PASA density functions are constructed centering three functions on H atoms, five functions on B, C, N, O, and F atoms, six functions on Na, S, and Cl atoms, and seven functions on Sc-Kr atoms. ASA coefficients and exponents used here can be downloaded from a WWW site [27]. See Appendix for more details concerning the fitting of Sc-Kr atoms.

To test the new methodology, an extensive set of molecules involving various structural features has been analyzed. In all cases the most stable molecular geometries have been considered. The structural diversity involves a series of diatomic molecules and simple organic compounds, as well as representatives of more complex molecules containing first-row transition metals. The computed number of SCF cycles using the three aforementioned approximations to generate an initial density matrix are summarized in Table I. As can be seen from these results, the use of the core Hamiltonian gives predominantly the highest number of SCF cycles, whereas the default option in the GAUSSIAN program and the PASA-MO formalism yield similar results. The main differences between these last two processes occur in the first-row transition metal complexes. For instance, the computed number of SCF cycles using PASA densities in  $\text{TiF}_4$ ,  $\text{VF}_5$ ,  $\text{CrO}_2\text{F}_2$ ,  $\text{CrO}_2\text{Cl}_2$ ,  $\text{CrO}_4^{2-}$ ,  $\text{Cr}(\text{CO})_6$ ,  $\text{Cr}(\text{CO})_5\text{N}_2$ ,  $\text{MnO}_4^-$ ,  $\text{Fe}(\text{CO})_5$ ,  $\text{Fe}(\text{CO})_4\text{H}_2$ ,  $\text{Ni}(\text{CO})_4$ , and  $\text{CuF}$  compounds becomes less than using the default option in the GAUSSIAN program. It must be mentioned that for charged molecules, corresponding in the present application example to the  $\text{BrO}^-$ ,  $\text{MnO}_4^-$ , and  $\text{CrO}_4^{2-}$  compounds, negative charges are shared out in an equitable manner between all the oxygen atoms of the corresponding molecule. That is, defining coefficient  $P_0$  in Eq. (3) as the atomic number of O atom plus the molecular charge divided by the total number of O belonging to the molecule studied. The simulation of negative charges in PASA formalism does not have a unique definition, although the adopted methodology ensures that the integration of PASA density function over all the space gives the total number of electrons of charged molecules.

In addition, Table I lists the quadratic error integral function between PASA and exact molecu-

lar densities,  $\varepsilon^{(2)}$  value, as well as the number of spherical functions employed to generate the PASA densities for each compound studied.  $\varepsilon^{(2)}$  values are used to evaluate the accuracy of PASA densities, and they are divided by the square number of electrons. It is possible to see from these results that there is indeed very satisfactory agreement between PASA and exact densities, especially if some of those results showing higher deviations can be rationalized, for example, the  $\text{CrO}_4^{2-}$  and  $\text{BrO}^-$  molecules, which correspond to charged compounds. Atomic fittings were performed over neutral atoms, so that some deficiencies can be expected in the description of electronic densities of charged molecules.

Finally, the proposed use of ASA density functions to construct initial MO in SCF calculations is illustrated in molecular electronic structure calculations of experimental transition metal complexes extracted from the Cambridge Structural Database (CSD) [22]. Chemical formulas and CSD reference codes for a variety of compounds, which present notably different structures, are listed in Figure 1. X-ray crystallographic structures were selected to avoid the multiple conformer problem and to permit the reproducibility of the present results. It is important to emphasize that molecules are considered in the experimental configuration and no molecular geometry optimization is performed. For this molecular series, all selected systems possess an even number of electrons, and only closed-shell (singlet) electronic states have been considered. The calculated number of SCF cycles for EHT and PASA-MO methodologies are presented in Table II, along with quadratic error results between PASA and exact density and the total number of  $1s$  fitted functions for each compound. A comparison of the results for both approaches indicates that the number of SCF cycles using PASA-MO formalism is, at least, comparable to the number of cycles for the EHT approach, and in several examples greatly lower. It is concluded that the use of the PASA-MO approach offers attractive possibilities to achieve faster MO-LCAO-SCF calculations for compounds including heteroatoms up to the fourth period.

---

## Discussion

The utility of PASA densities to obtain starting MO in SCF calculations has been demonstrated by the present results. The main improvement with respect to existing GAUSSIAN methods is accomplished when transition metal complexes are given

**TABLE I**  
**SCF cycles and quadratic error results between PASA and exact density function for different molecular systems.<sup>a</sup>**

Molecule	<i>H</i>	INDO CNDO		PASA-MO	<i>n</i> <sup>b</sup>	$\epsilon^{(2)}$
		EHT				
BH <sub>3</sub>	10	9		8	14	$5.60 \times 10^{-4}$
B <sub>2</sub> H <sub>6</sub>	11	9		9	28	$2.02 \times 10^{-4}$
CH <sub>4</sub>	10	9		8	17	$2.69 \times 10^{-4}$
C <sub>2</sub> H <sub>2</sub>	11	9		9	16	$1.76 \times 10^{-4}$
C <sub>2</sub> H <sub>4</sub>	11	10		9	22	$1.73 \times 10^{-4}$
C <sub>2</sub> H <sub>6</sub>	11	9		9	28	$1.53 \times 10^{-4}$
CH <sub>3</sub> NH <sub>2</sub>	14	12		12	25	$1.78 \times 10^{-4}$
N <sub>2</sub>	11	9		9	10	$2.81 \times 10^{-4}$
NH <sub>3</sub>	12	10		11	14	$2.92 \times 10^{-4}$
N <sub>2</sub> H <sub>2</sub>	12	10		10	16	$3.02 \times 10^{-4}$
N <sub>2</sub> H <sub>4</sub>	14	11		11	22	$2.27 \times 10^{-4}$
HNC	14	11		11	13	$1.84 \times 10^{-4}$
HCN	13	11		11	13	$2.16 \times 10^{-4}$
HNO	16	12		12	13	$4.71 \times 10^{-4}$
NH <sub>2</sub> OH	15	13		12	19	$3.38 \times 10^{-4}$
CO	15	10		10	10	$2.75 \times 10^{-4}$
O <sub>2</sub>	25	9		9	10	$7.37 \times 10^{-4}$
H <sub>2</sub> O	13	11		10	11	$5.22 \times 10^{-4}$
H <sub>2</sub> O <sub>2</sub>	13	11		10	16	$4.84 \times 10^{-4}$
CH <sub>3</sub> OH	15	11		11	22	$2.63 \times 10^{-4}$
H <sub>2</sub> CO	15	12		11	16	$3.03 \times 10^{-4}$
CH <sub>3</sub> COOH	18	13		13	32	$1.27 \times 10^{-4}$
CH <sub>3</sub> CH <sub>2</sub> COOH	21	15		15	43	$9.41 \times 10^{-5}$
F <sub>2</sub>	9	8		8	10	$7.89 \times 10^{-4}$
HF	11	9		9	8	$5.29 \times 10^{-4}$
HOF	15	13		13	13	$5.91 \times 10^{-4}$
CH <sub>3</sub> F	14	10		10	19	$2.82 \times 10^{-4}$
CH <sub>3</sub> Cl	14	11		11	20	$1.34 \times 10^{-4}$
CH <sub>2</sub> F <sub>2</sub>	15	11		11	21	$2.16 \times 10^{-4}$
CH <sub>2</sub> Cl <sub>2</sub>	14	12		11	23	$8.44 \times 10^{-5}$
CHF <sub>3</sub>	14	11		11	23	$1.71 \times 10^{-4}$
CHCl <sub>3</sub>	15	13		13	26	$6.11 \times 10^{-5}$
CF <sub>4</sub>	12	11		10	25	$1.44 \times 10^{-4}$
CCl <sub>4</sub>	14	12		12	29	$4.77 \times 10^{-5}$
H <sub>2</sub> SO	19	14		14	17	$2.68 \times 10^{-4}$
F <sub>2</sub> S <sub>2</sub>	23	16		18	22	$1.13 \times 10^{-4}$
NaBr	27	12		12	13	$1.17 \times 10^{-4}$
ScF	23	24		21	12	$1.16 \times 10^{-4}$
ScF <sub>3</sub>	71	14		13	22	$7.04 \times 10^{-5}$
TiF <sub>4</sub>	497	31		15	27	$6.59 \times 10^{-5}$
VF <sub>5</sub>	<i>nc</i> <sup>c</sup>	35		17	32	$6.60 \times 10^{-5}$
CrO <sub>4</sub> <sup>2-</sup>	333	34		27	27	$4.90 \times 10^{-4}$
CrO <sub>2</sub> F <sub>2</sub>	59	33		28	27	$8.53 \times 10^{-5}$
CrO <sub>2</sub> Cl <sub>2</sub>	<i>nc</i> <sup>c</sup>	39		31	29	$5.85 \times 10^{-5}$
Cr(CO) <sub>6</sub>	<i>nc</i> <sup>c</sup>	100		20	67	$4.47 \times 10^{-5}$
Cr(CO) <sub>5</sub> N <sub>2</sub>	<i>nc</i> <sup>c</sup>	61		31	67	$4.61 \times 10^{-4}$
MnO <sub>4</sub> <sup>-</sup>	77	33		20	27	$2.05 \times 10^{-4}$
Fe(CO) <sub>5</sub>	<i>nc</i> <sup>c</sup>	29		22	57	$6.61 \times 10^{-5}$

(Continued)

**TABLE I**  
Continued

Molecule	$H$	INDO CNDO			$n^b$	$\varepsilon^{(2)}$
		EHT	PASA-MO			
Fe(CO) <sub>4</sub> H <sub>2</sub>	$nc^c$	34	28	53	$1.12 \times 10^{-4}$	
Co(CO) <sub>4</sub> H	$nc^c$	38	36	50	$9.61 \times 10^{-5}$	
Ni(CO) <sub>4</sub>	312	31	21	47	$4.76 \times 10^{-5}$	
CuH	23	21	18	10	$1.83 \times 10^{-4}$	
CuF	30	25	19	12	$1.43 \times 10^{-4}$	
Zn(CH <sub>3</sub> ) <sub>2</sub>	32	15	17	35	$6.52 \times 10^{-5}$	
GaCl	19	12	12	13	$9.37 \times 10^{-5}$	
GeH <sub>4</sub>	13	10	10	19	$1.49 \times 10^{-4}$	
GeO	18	12	12	12	$1.31 \times 10^{-4}$	
SeH <sub>2</sub>	13	11	11	13	$1.63 \times 10^{-4}$	
AsH <sub>3</sub>	13	11	11	16	$1.57 \times 10^{-4}$	
HBr	12	10	10	10	$1.77 \times 10^{-4}$	
BrO <sup>-</sup>	16	12	14	12	$3.53 \times 10^{-4}$	
BrF	17	11	11	12	$1.63 \times 10^{-4}$	
BrCl	17	11	11	13	$1.13 \times 10^{-4}$	
KrF <sub>2</sub>	18	12	11	17	$1.28 \times 10^{-4}$	

<sup>a</sup> These computations have been performed on a SGI Origin 2000/8 R10000 processors using 1 node.

<sup>b</sup> Total number of 1s ASA functions.

<sup>c</sup> No converged.

as examples. Furthermore, the introduction of the PASA concept is quite promising, since it constitutes the starting point for a new computational method able to expand quantum chemistry to very large systems. In this initial phase of the methodological development, only a simple application in the reduction of SCF cycles has been tested, and in this way the PASA-MO approach constitutes a general and homogeneous systematic algorithm to produce the initial MO guess in SCF procedures.

## Appendix

### ASA FITTING ALGORITHM

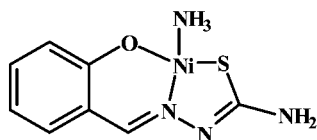
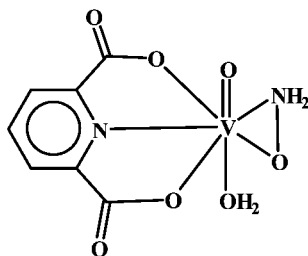
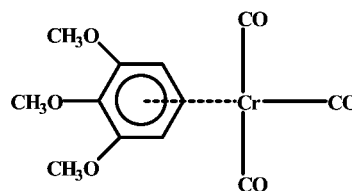
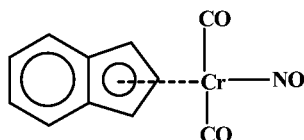
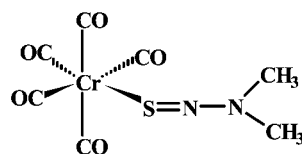
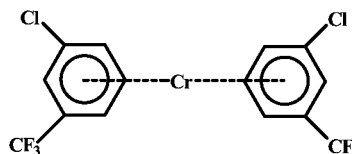
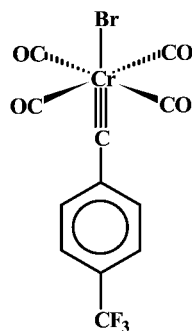
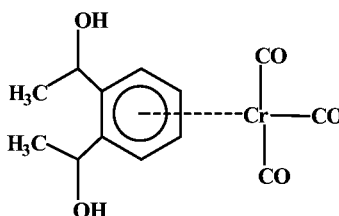
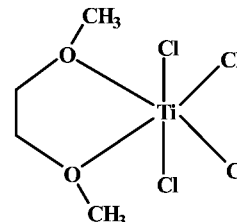
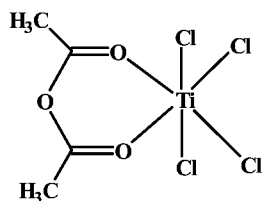
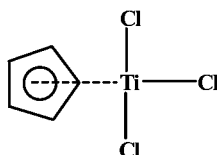
Since the detailed and strict mathematical development of the ASA fitting can be found in previous papers [4–7], only the basic ideas of the algorithm will be briefly explained. Fundamentally, spherically symmetric atomic densities,  $\rho_a^{\text{ASA}}(\mathbf{r})$ , are fitted to any ab initio density function  $\rho_a(\mathbf{r})$  by minimizing the functional:

$$\varepsilon^{(2)} = \int |\rho_a(\mathbf{r}) - \rho_a^{\text{ASA}}(\mathbf{r})|^2 d\mathbf{r}, \quad (11)$$

which is integrated over all space. Ab initio RHF electronic energies and density matrix for the in-

cluded atoms in this study have been calculated using the ATOMIC program [28].

Among known fitting techniques, a least squares method in combination with a Lagrange multiplier to preserve density normalization is the most usual [29–36]. However, this algorithm does not guarantee that the expansion coefficients  $\{w_i\}$  would be in any case positive definite. To add this supplementary constraint, a robust and fast algorithm has been developed in our laboratory [4–7]. The first step in the procedure consists of defining a new set of coefficients:  $\{x_i\}$ , where  $\forall i: w_i = |x_i|^2$ . Then, starting from a set of coefficients which present the normalization condition of the density function, an elementary Jacobi rotations (EJR) technique is applied to minimize the  $\varepsilon^{(2)}$  function, Eq. (11). Within orthogonal transformations, both constraints required for ASA coefficients are fulfilled along the whole process. In addition, the efficiency of the EJR optimization algorithm was improved by expressing sine and cosine functions by means of a Taylor series [5]. This approximation is effective when the optimization procedure is near the minimum, and produces a dramatic reduction in the computation time. Recently, this algorithm has been successfully implemented to obtain accurate ASA fitting of molecular electron densities [6].

**AMSCNI11**

**AQHPRV**

**ANLCRB10**

**BABFOZ**

**BASTOE10**

**BAVGIO**

**BCPCBC**

**BENGLY**

**BOMFAK**

**BUDHIR**

**CEHPIO**

**FIGURE 1.** Cambridge Structural Data base structures and reference codes of selected metal complexes.

With respect to ASA exponents, formerly they were generated from an even tempered sequence, and no optimization was performed [1]. To improve the fitting results without the necessity of increasing the number of fitted functions, a Newton method was incorporated into the code in order to enable exponents' optimization. The deduced equations for the analytical gradient and hessian of the  $\varepsilon^{(2)}$  func-

tion were detailed in the first paper of this series [4]. The Newton method has an intrinsic sensitivity, but possesses a rather limited applicability in a delimited zone, even more when the function to be optimized has numerous local optima. However, with a fine grid carried out under even-tempered parameters, initial sets of ASA exponents are generated, which provide reasonable starting points for

**TABLE II**  
**SCF cycles and quadratic error results between PASA and exact density function for CSD compounds containing first-row transition metals.<sup>a</sup>**

Molecule	EHT	PASA-MO	$n^b$	$\varepsilon^{(2)}$
AMSCNI11	40	36	108	$4.84 \times 10^{-5}$
AQHPRV	39	40	108	$2.97 \times 10^{-5}$
ANLCRB10	118	54	133	$3.40 \times 10^{-5}$
BABFOZ	76	65	103	$4.01 \times 10^{-5}$
BASTOE10	55	50	101	$3.39 \times 10^{-5}$
BAVGIO	60	60	143	$2.44 \times 10^{-5}$
BCPCBC	94	80	121	$2.56 \times 10^{-5}$
BENGLY	60	47	139	$2.82 \times 10^{-5}$
BOMFAK	34	17	91	$2.91 \times 10^{-5}$
BUDHIR	38	17	84	$2.68 \times 10^{-5}$
CEHPIO	31	18	65	$3.44 \times 10^{-5}$

<sup>a</sup> GAUSSIAN computations have been carried out in a PC ADM K7 1000 MHz, running under Linux.

<sup>b</sup> Total number of 1s ASA functions.

the Newton method. This methodology gives accurate ASA density functions, but on the other hand, it has been limited to atomic fittings of a relative small number of functions.

In summary, the proposed algorithm has two essential features that distinguish it from other algorithms: (i) it preserves the physical meaning of the fitted density function by forcing the linear coefficients to be positive definite and (ii) it combines the optimization of exponents and coefficients in order to obtain the optimal fitting with the minimal number of functions.

#### ATOMIC FITTING EXAMPLE

As an application example, an ASA basis set for atoms Sc-Kr fitted to an ab initio 6-311G basis set has been calculated. In a way, this part of the present article represents the continuation of previous work [6, 7], where the fitting for atoms H-Ar within the same ab initio basis set level has been given. Results are presented in Table III using seven fitted functions per atom. The values of  $\varepsilon^{(2)}$  are divided by the square number of electrons in order to give comparable and normalized results. Present results are consistent with previous calculations carried out for atoms H-Ar [6, 7]. In addition, Table III lists the relative errors between ASA and exact atomic quantum self-similarity measures. An overlap quantum self-similarity measure is defined by means of the integral:  $Z_{aa} = \int |\rho_a(r)|^2 dr$ . The result-

**TABLE III**  
**Fitting results for the 6-311G basis set of atoms Sc to Kr using seven fitted atomic functions.**

	$\varepsilon^{(2)a}$	% $Z_{aa}^b$
Sc ( <sup>2</sup> D)	$1.41 \times 10^{-4}$	-0.067
Ti ( <sup>3</sup> F)	$1.40 \times 10^{-4}$	-0.067
V ( <sup>4</sup> F)	$1.43 \times 10^{-4}$	-0.068
Cr ( <sup>5</sup> D)	$1.45 \times 10^{-4}$	-0.068
Mn ( <sup>6</sup> S)	$1.46 \times 10^{-4}$	-0.069
Fe ( <sup>5</sup> D)	$1.48 \times 10^{-4}$	-0.070
Co ( <sup>4</sup> F)	$1.49 \times 10^{-4}$	-0.071
Ni ( <sup>3</sup> F)	$1.51 \times 10^{-4}$	-0.072
Cu ( <sup>2</sup> D)	$1.60 \times 10^{-4}$	-0.073
Zn ( <sup>1</sup> S)	$1.15 \times 10^{-4}$	-0.005
Ga ( <sup>2</sup> P)	$1.70 \times 10^{-4}$	-0.071
Ge ( <sup>3</sup> P)	$1.72 \times 10^{-4}$	-0.088
As ( <sup>4</sup> S)	$1.75 \times 10^{-4}$	-0.083
Se ( <sup>3</sup> P)	$1.75 \times 10^{-4}$	-0.082
Br ( <sup>2</sup> P)	$1.85 \times 10^{-4}$	-0.082
Kr ( <sup>1</sup> S)	$1.87 \times 10^{-4}$	-0.082

<sup>a</sup> Quadratic error integral.

<sup>b</sup> % error in atomic quantum self-similarity measure  $Z_{aa}$ .

ing ASA  $Z_{aa}$  measures are in good agreement with exact values, giving relative errors less than 0.09%. Exponents and coefficients for the ASA atomic basis set employed in this work are accessible from a WWW address [27].

#### References

- Constans, P.; Carbó, R. *J Chem Inf Comput Sci* 1995, 35, 1046-1053.
- Constans, P.; Amat, L.; Fradera, X.; Carbó-Dorca, R. In *Advances in Molecular Similarity*; Carbó-Dorca, R.; Mezey, P. G., Eds.; JAI Press: Greenwich, CT, 1996; Vol. 1, pp. 187-211.
- Gironés, X.; Amat, L.; Carbó-Dorca, R. *J Mol Graph Model* 1998, 16, 190-196.
- Amat, L.; Carbó-Dorca, R. *J Comput Chem* 1997, 18, 2023-2039.
- Amat, L.; Carbó-Dorca, R. *J Comput Chem* 1999, 20, 911-920.
- Amat, L.; Carbó-Dorca, R. *J Chem Inf Comput Sci* 2000, 40, 1188-1198.
- Carbó-Dorca, R.; Amat, L.; Besalú, E.; Gironés, X.; Robert, D. In *The Fundamentals of Molecular Similarity*; Carbó-Dorca, R., Ed.; Kluwer: Dordrecht. To appear.
- Ritchie, J. P. *J Am Chem Soc* 1985, 107, 1829-1837.
- Schwarz, W. H. E.; Ruedenberg, K.; Mensching, L. *J Am Chem Soc* 1989, 111, 6926-6933.
- Ruedenberg, K.; Schwarz, W. H. E. *J Chem Phys* 1990, 92, 4956-4969.

11. Robert, D.; Gironés, X.; Carbó-Dorca, R. *J Chem Inf Comput Sci* 2000, 40, 839–846.
12. Lobato, M.; Amat, L.; Besalú, E.; Carbó-Dorca, R. *Quant Struct-Act Relat* 1997, 16, 465–472.
13. Amat, L.; Robert, D.; Besalú, E.; Carbó-Dorca, R. *J Chem Inf Comput Sci* 1998, 38, 624–631.
14. Robert, D.; Amat, L.; Carbó-Dorca, R. *J Chem Inf Comput Sci* 1999, 39, 333–344.
15. Robert, D.; Gironés, X.; Carbó-Dorca, R. *J Comput Aided Mol Design* 1999, 13, 597–610.
16. Robert, D.; Carbó-Dorca, R. *SAR QSAR Environ Res* 1999, 10, 401–422.
17. Amat, L.; Carbó-Dorca, R.; Ponec, R. *J Med Chem* 1999, 42, 5169–5180.
18. Carbó-Dorca, R.; Robert, D.; Amat, L.; Gironés, X.; Besalú, E. *Molecular Quantum Similarity in QSAR and Drug Design; Lecture Notes in Chemistry*; Springer: New York, 2000; Vol. 73.
19. Segal, G.; Pople, J. *J Chem Phys* 1966, 44, 3289–3296.
20. Pople, J. A.; Beveridge, D.; Dobosh, P. *J Chem Phys* 1967, 47, 2026–2033.
21. Hoffmann, R. *J Chem Phys* 1963, 39, 1397–1412.
22. Allen, F. H.; Bellard, S.; Brice, M. D.; Cartwright, B. A.; Doubleday, A.; Higgs, H.; Hummelink, T.; Hummelink-Peters, B. G.; Kennard, O.; Motherwell, W. D. S.; Rodgers, J. R.; Watson, D. G. *Acta Crystallogr* 1979, B35, 2331.
23. Jacobi, C. G. J. *J Reine Angew Math* 1846, 30, 51–94.
24. Roothaan, C. C. J. *Rev Mod Phys* 1951, 23, 69–89.
25. 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, A. 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.; Baboul, A. G.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. 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.7*; Gaussian, Inc.: Pittsburgh, PA, 1998.
26. Carbó, R.; Domingo, L.; Gregori, J. *Int J Quantum Chem* 1980, 17, 725–736.
27. ASA coefficients and exponents for different fitted atomic basis functions can be downloaded from the WWW site: <http://iqc.udg.es/cat/similarity/ASA/basiset.html>.
28. Carbó-Dorca, R. *ATOMIC Program 1995*, based on: Roos, B.; Salez, C.; Veillard, A.; Clementi, E. A general program for calculation of SCF orbitals by the expansion method. *IBM Research/RJ518(#10901)*, 1968.
29. Baerends, E. J.; Ellis, D. E.; Ros, P. *Chem Phys* 1973, 2, 41–51.
30. Samba, H.; Felton, R. H. *J Chem Phys* 1975, 62, 1122–1126.
31. Dunlap, B. I.; Connolly, J. W. D.; Sabin, J. R. *J Chem Phys* 1979, 71, 3396–3402.
32. Delley, B.; Ellis, D. E. *J Chem Phys* 1982, 76, 1949–1960.
33. Andzelm, J.; Wimmer, E. *J Chem Phys* 1992, 96, 1280–1303.
34. Mestres, J.; Solà, M.; Duran, M.; Carbó, R. *J Comput Chem* 1994, 15, 1113–1120.
35. Gallant, R. T.; St.-Amant, A. *Chem Phys Lett* 1996, 256, 569–574.
36. Goh, S. K.; St.-Amant, A. *Chem Phys Lett* 1997, 264, 9–16.



## Discussió

En aquest capítol s'ha demostrat que la tècnica de rotacions de Jacobi es pot aplicar a l'ajust de *DF* aproximades amb coeficients definits positius. La parametrització de les bases de funcions atòmiques, en especial la 3-21G, ha permès posteriors estudis de *MQSM* emprant una aproximació promolecular. Només és necessari conèixer les coordenades atòmiques i especificar un dels conjunts de funcions ASA parametritzades, per generar automàticament la densitat electrònica de qualsevol molècula. La qualitat de les funcions promoleculars s'ha il·lustrat amb varis exemples de *MQSM*. En el primer treball s'han comparat els valors de les *MQSM* de tipus solapament emprant densitats HF i *PASA* en un conjunt de derivats del metà. S'ha pogut observar que els errors relatius entre ambdues mesures són, en tots els exemples estudiats, inferiors al 2 %. En el segon treball s'ha utilitzat les *MQSM* per determinar el millor mètode teòric de càlcul de la geometria molecular del compost cis-platin. Aquest estudi ha estat possible gràcies a la dilucidació d'una base de funcions ASA atòmica que arriba fins al radó, ajustades a la densitat HF d'una base d'Huzinaga. En el tercer article s'analitza un suposat intermedi en la reacció d'inhibició de la glicosidasa a través de l'establiment de semblances amb compostos actius i no actius. El següent exemple, relacionat amb l'anterior, ha estat l'examen dels camins de reacció en reaccions de reordenació intramolecular. S'ha analitzat el comportament Hammond i anti-Hammond de quatre reaccions simples de transposició mitjançant quantitats derivades de les *MQSM*. En resum, tots els exemples presentats demostren que l'aproximació *PASA* és suficientment precisa pel càlcul pràctic de les mesures i índexs de semblança quàntica. A més, i tal com es veurà en els capítols 6 i 7, les *PASA DF* han permès l'extensió de les *MQSM* en l'àmbit de la racionalització i predicció de l'activitat de fàrmacs.

Un segon aspecte metodològic fa referència al protocol establert a l'hora de calcular les funcions ASA. L'algorisme proposat per l'ajust d'àtoms té dues parts essencials que el distingeixen d'altres algorismes. En primer lloc preserva el significat físic de la *DF* aproximada forçant els coeficients a ser definits positius, i d'altra banda combina l'optimització d'exponents i coeficients amb l'objectiu d'assolir un ajust òptim amb el menor nombre de funcions. Quant al mètode d'ajust molecular, mencionar que la *PASA DF* proporciona uns excel·lents coeficients que serveixen de punt de partida a l'algorisme *EJR*. A més, la millora en el càlcul de l'angle de rotació *EJR* amb el



desenvolupament en sèries de Taylor del sinus i cosinus ha funcionat especialment bé en l'ajust de molècules.

El principal inconvenient de la parametrització de conjunts de funcions *ASA* atòmiques és el temps de computació que requereix. Realitzar els càlculs en una xarxa de punts definida sobre els paràmetres *even-tempered* suficientment densa, i fer-ho per tots els àtoms considerats i per diferent nombre de funcions, suposa moltes hores de computació. Un segon inconvenient de l'ajust atòmic és la limitació en el nombre de funcions. Normalment no es poden superar les 8–10 capes per àtom degut a les restriccions imposades en els coeficients *ASA*. Si s'utilitzen més funcions l'algorisme tendeix a eliminar-ne, ja sigui fent zero el coeficient o bé igualant dos exponents de la sèrie.

L'últim treball inclòs en aquest capítol ha estat un exemple d'aplicació de les *PASA DF* en càlculs d'energia, en concret en la generació de matrius densitats inicials per al càlcul *SCF*. Les principals diferències, quan es compara amb les aproximacions emprades en el programa *GAUSSIAN*, s'observen en complexes amb metalls de transició. La reducció del nombre de cicles *SCF* ha estat una aplicació senzilla, però serveix d'introducció de l'aproximació *PASA* en l'àmbit dels càlculs d'energia electrònica. Actualment s'està en una fase preliminar de desenvolupament d'un possible nou mètode de càlcul en química quàntica de sistemes molt grans emprant *PASA DF*.

## Referències

1. C. C. J. Roothaan. New developments in molecular orbital theory. *Rev. Mod. Phys.* **1951**, *23*, 69–89.
2. E. J. Baerends, D. E. Ellis, P. Ros. Self-consistent molecular Hartree-Fock-Slater calculations. I. The computational procedure. *Chem. Phys.* **1973**, *2*, 41–51.
3. H. Sambe, R. H. Felton. A new computational approach to Slater's SCF- $X\alpha$  equation. *J. Chem. Phys.* **1975**, *62*, 1122–1126.
4. B. Delley, D. E. Ellis. Efficient and accurate expansion methods for molecules in local density models. *J. Chem. Phys.* **1982**, *76*, 1949–1960.
5. B. I. Dunlap, J. W. D. Connolly, J. R. Sabin. On some approximations in applications of  $X\alpha$  theory. *J. Chem. Phys.* **1979**, *71*, 3396–3402.
6. J. Andzelm, E. Wimmer. Density functional Gaussian-type-orbital approach to molecular geometries, vibrations, and reaction energies. *J. Chem. Phys.* **1992**, *96*, 1280–1303.
7. R. T. Gallant, A. St-Amant. Linear scaling for the charge density fitting procedure of the linear combination of Gaussian-type orbitals density functional method. *Chem. Phys. Letters* **1996**, *256*, 569–574.
8. S. K. Goh, A. St-Amant. Using a fitted electronic density to improve the efficiency of a linear combination of Gaussian-type orbitals calculation. *Chem. Phys. Letters* **1997**, *264*, 9–16.
9. J. Mestres, M. Solà, M. Duran, R. Carbó. On the calculation of *ab initio* quantum molecular similarities for large systems: fitting the electron density. *J. Comput. Chem.* **1994**, *15*, 1113–1120.
10. E. Besalú, R. Carbó, J. Mestres, M. Solà. Foundations and recent developments on molecular quantum similarity. *Topics in Current Chemistry* **1995**, *173*, 31–62.
11. P. Constans, R. Carbó. Atomic shell approximation: electron density fitting algorithm restricting coefficients to positive values. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1046–1053.
12. P. Constans, Ll. Amat, X. Fradera, R. Carbó-Dorca. Quantum molecular similarity measures (QMSM) and the atomic shell approximation (ASA). Publicat en el llibre: *Advances in molecular similarity*. R. Carbó-Dorca, P. G. Mezey (Eds.). JAI Press, Greenwich, CT, Vol. 1, pàgines 187–211, 1996.
13. R. Carbó-Dorca, E. Besalú, Ll. Amat, X. Fradera. Quantum molecular similarity measures: concepts, definitions, and applications to quantitative structure-property relationships. Publicat en el llibre: *Advances in molecular similarity*. R. Carbó-Dorca, P. G. Mezey (Eds.). JAI Press, Greenwich, CT, volum 1, pàgines 1–41, 1996.
14. Ll. Amat, R. Carbó-Dorca. Quantum similarity measures under atomic shell approximation: first order density fitting using elementary Jacobi rotations. *J. Comput. Chem.* **1997**, *18*, 2023–2039.
15. R. Carbó-Dorca, Ll. Amat, E. Besalú, M. Lobato. Quantum similarity. Publicat en el llibre: *Advances in molecular similarity*. R. Carbó-Dorca, P. G. Mezey (Eds.). JAI Press, London, volum 2, pàgines 1–41, 1998.
16. X. Gironés, Ll. Amat, R. Carbó-Dorca. A comparative study of isodensity surfaces using *ab initio* and ASA density functions. *J. Mol. Graph. Model.* **1998**, *16*, 190–196.

17. Ll. Amat, R. Carbó-Dorca. Fitted electronic density functions from H to Rn for use in quantum similarity measures: cis-diammine-dichloroplatinum(II) complex as an application example. *J. Comput. Chem.* **1999**, *20*, 911–920.
18. R. Carbó-Dorca, Ll. Amat, E. Besalú, X. Gironés, D. Robert. Quantum mechanical origin of QSAR: theory and applications. *J. Mol. Struct. (Theochem)* **2000**, *504*, 181–228.
19. Ll. Amat, R. Carbó-Dorca. Molecular electronic density fitting using elementary Jacobi rotations under atomic shell approximation. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1188–1198.
20. K. Ruedenberg, R. C. Raffenti, R. D. Bardón. Publicat en el llibre: Energy, structure and reactivity. Proceedings of the 1972 Boulder seminar research conference on theoretical chemistry. D. W. Smith (ed.). Wiley, New York, pàgina 164, 1973.
21. S. Huzinaga, M. Klobukowski. Well-tempered Gaussian basis set expansions of Roothaan-Hartree-Fock atomic wavefunctions for lithium through mercury. *J. Mol. Struct. (Theochem)* **1988**, *44*, 1–87.
22. K. Ruedenberg, W. H. E. Schwarz. Nonspherical atomic ground-state densities and chemical deformation densities from x-ray scattering. *J. Chem. Phys.* **1990**, *92*, 4956–4969.
23. Jacobi, C. G. J. Über ein leichtes Verfahren die in der theorie der Säcularstörungen vorkommenden gleichungen numerisch aufzulösen. *J. Reine Angew. Math.* **1846**, *30*, 51–94.
24. K. J. Miller, K. Ruedenberg. Electron correlation and separated-pair approximation. An application to Beryllium like atomic systems. *J. Chem. Phys.* **1968**, *48*, 3414–3443.
25. D. M. Silver, E. L. Mehler, K. Ruedenberg. Electron correlation and separated pair approximation in diatomic molecules. I. Theory. *J. Chem. Phys.* **1970**, *52*, 1174–1180.
26. E. L. Mehler, K. Ruedenberg, D. M. Silver. Electron correlation and separated pair approximation in diatomic molecules. II. Lithium hydride and boron hydride. *J. Chem. Phys.* **1970**, *52*, 1181–1205.
27. D. M. Silver, K. Ruedenberg, E. L. Mehler. Electron correlation and separated pair approximation in diatomic molecules. III. Imidogen. *J. Chem. Phys.* **1970**, *52*, 1206–1227.
28. R. Carbó, Ll. Domingo, J. J. Peris. Elementary unitary MO transformations and SCF theory. *Adv. in Quantum Chem.* **1982**, *15*, 215–265.
29. R. Carbó, Ll. Domingo, J. J. Peris, J. J. Novoa. Energy variation and elementary Jacobi rotations. *J. Mol. Struct. (THEOCHEM)* **1983**, *93*, 15–33.
30. R. Carbó, Ll. Domingo, J. J. Novoa. Multiconfigurational calculations using elementary Jacobi rotations. *J. Mol. Struct. (THEOCHEM)* **1985**, *120*, 357–363.
31. R. Carbó, J. Miró, Ll. Domingo, J. J. Novoa. Jacobi rotations: a general procedure for electronic energy optimization. *Adv. in Quantum Chem.* **1989**, *20*, 375–441.
32. C. Edmiston, K. Ruedenberg. Localized atomic and molecular orbitals. *Rev. Mod. Phys.* **1963**, *35*, 457–465.
33. R. C. Raffentti, K. Ruedenberg, C. L. Janssen, H. F. Schaefer. Efficient use of Jacobi rotations for orbital optimization and localization. *Theor. Chim. Acta* **1992**, *86*, 149–165.
34. D. A. Pierre. Optimization theory with applications. John Wiley & Sons, New York, 1969.
35. Programa GATOMIC. Ll. Amat, R. Carbó-Dorca. Institut de química computacional, Universitat de Girona, 1997.

36. Programa ATOMIC. R. Carbó-Dorca. Institut de química computacional, Universitat de Girona, 1995.
37. C. C. J. Roothaan, P. S. Bagus. Atomic self-consistent field calculations by the expansion method. Publicat en el llibre: *Methods in computational physics. Advances in research and applications*. B. Alder, S. Fernbach, M. Rotenberg (Eds.). Academic Press, New York, Volum 2, pàgines 47–94, 1963.
38. B. Roos, C. Salez, A. Veillard, E. Clementi. A general program for calculation of SCF orbitals by the expansion method. IBM Research/RJ518(#10901), 1968.
39. J. S. Binkey, J. A. Pople, W. J. Hehre. Self-consistent molecular orbital methods. 21. Small split-valence basis sets for first-row elements. *J. Am. Chem. Soc.* **1980**, *102*, 939–947.
40. M. S. Gordon, J. S. Binkley, J. A. Pople, W. J. Pietro, W. J. Hehre. Self-consistent molecular orbital methods. 22. Small split-valence basis sets for second-row elements. *J. Am. Chem. Soc.* **1982**, *104*, 2797–2803.
41. K. D. Dobbs, W. J. Hehre. Molecular orbital theory of the properties of inorganic and organometallic compounds. 4. Extended basis sets for third- and fourth-row, main-group elements. *J. Comput. Chem.* **1986**, *7*, 359–378.
42. <http://iqc.udg.es/cat/similarity/ASA/basisset.html>
43. GAUSSIAN 98, Revision A.6. M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, V. G. Zakrzewski, J. A. Montgomery Jr., R. E. Stratmann, J. C. Burant, S. Dapprich, J. M. Millam, A. D. Daniels, K. N. Kudin, M. C. Strain, O. Farkas, J. Tomasi, V. Barone, M. Cossi, R. Cammi, B. Mennucci, C. Pomelli, C. Adamo, S. Clifford, J. Ochterski, G. A. Petersson, P. Y. Ayala, Q. Cui, K. Morokuma, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. Cioslowski, J. V. Ortiz, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. Gomperts, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, C. Gonzalez, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, J. L. Andres, C. Gonzalez, M. Head-Gordon, E. S. Replogle, J. A. Pople. Gaussian, Inc.: Pittsburgh, PA, 1998.
44. F. H. Allen, S. Bellard, M. D. Brice, B. A. Cartwright, A. Doubleday, H. Higgs, T. Hummelink, B. G. Hummelink-Peters, O. Kennard, W. D. S. Motherwell, J. R. Rodgers, D. G. Watson. The Cambridge crystal data center: computer-based search, retrieval, analysis and display of information. *Acta Crystallogr.* **1979**, *B35*, 2331–2339.
45. S. Huzinaga (Ed.). Gaussian basis sets for molecular calculations. Physical sciences data 16. Elsevier, Amsterdam, 1984.
46. R. Krishnan, J. S. Binkley, R. Seeger, J. A. Pople. Self-consistent molecular orbital methods. XX. A basis set for correlated wave functions. *J. Chem. Phys.* **1980**, *72*, 650–654.
47. A. D. McLean, G. S. Chandler. Contracted Gaussian basis sets for molecular calculations. I. Second row atoms,  $Z=11-18$ . *J. Chem. Phys.* **1980**, *72*, 5639–5648.
48. L. A. Curtiss, M. P. McGrath, J.-P. Blaudeau, N. E. Davis, R. C. Binning, L. Radom. Extension of Gaussian-2 theory to molecules containing third-row atoms Ga–Kr. *J. Chem. Phys.* **1995**, *103*, 6104–6113.
49. Ll. Amat, R. Carbó-Dorca, D. L. Cooper, N. L. Allan. Classification of reaction pathways via momentum-space and quantum molecular similarity measures. *Chem. Phys. Lett.* **2003**, *367*, 207–213.
50. M. Solà, J. Mestres, R. Carbó, M. Duran. Use of *ab initio* quantum similarity measures as an interpretative tool for the study of chemical reactions. *J. Am. Chem. Soc.* **1994**, *116*, 5909–5915.

51. G. S. Hammond. A correlation of reaction rates. Hammond postulate. *J. Am. Chem. Soc.* **1955**, *77*, 334–338.
52. J. Cioslowski. Quantifying the Hammond postulate: intramolecular proton transfer in substituted hydrogen catecholate anions. *J. Am. Chem. Soc.* **1991**, *113*, 6756–6760.
53. V. R. Saunders. An introduction to molecular integral evaluation. Publicat en el llibre: Computational techniques in quantum chemistry and molecular physics. Diercksen et al. (eds.). D. Reidel Publishing Company, Dordrecht-Holland, pàgines 347–424, 1975.

## 4. Superposició molecular

---

Un aspecte molt important a considerar en totes aquelles tècniques comparatives que utilitzen estructures moleculars tridimensionals és l'establiment d'un algorisme eficient per alinear en l'espai els compostos considerats. La superposició d'estructures moleculars té aplicació en diferents àrees de la química, per exemple en els mètodes *3D-QSAR*,<sup>1</sup> en la comparació quantitativa de l'estereoquímica molecular i en la mesura de la distorsió molecular en cristalls,<sup>2</sup> o en el reconeixement de patrons en bases de dades moleculars,<sup>3-6</sup> on és necessari fer cerques ràpides sobre conjunts que contenen molts de compostos. Els mètodes de comparació i quantificació de la similitud estructural de molècules s'han fonamentat tradicionalment en la minimització de la suma de les distàncies entre els àtoms de les molècules que es comparen.<sup>7-11</sup> Donada a priori una correspondència entre àtoms d'una de les molècules amb els de l'altra, s'optimitzen les posicions relatives de les molècules fins a aconseguir una sobreposició òptima de les dues estructures. Correspondències entre àtoms de molècules diferents porten implícita una certa arbitriietat en la seva elecció, la qual es deixa als criteris dels usuaris dels programes. No definir-les a priori porta a un problema de combinatòria, que s'ha resolt amb algorismes aproximats de naturalesa estocàstica, com per exemple la tècnica de *Simulated Annealing*.<sup>12-14</sup>

La definició de la *MQSM* depèn de la posició relativa de les molècules comparades en l'espai. En el nostre laboratori s'han desenvolupat dos algorismes de superposició molecular per ser aplicats en estudis de *MQSM*. Inicialment es va adoptar el criteri de definir la *MQSM* en el màxim absolut,<sup>15,16</sup>

$$Z_{AB}(\Omega; \mathbf{Q}) = \underset{\Theta}{\text{Max}} \left[ \iint \mathbf{r}_A(\mathbf{r}_1) \Omega(\mathbf{r}_1, \mathbf{r}_2) \mathbf{r}_B(\mathbf{r}_2; \mathbf{Q}) d\mathbf{r}_1 d\mathbf{r}_2 \right], \quad (4.1)$$

que requereix una recerca de la superposició òptima. Més recentment s'ha dissenyat un nou algorisme que es fonamenta en la recerca de la subestructura comuna més gran entre dues molècules.<sup>17</sup> És una aproximació que únicament utilitza conceptes topològics i geomètrics, i per això s'ha anomenat algorisme de superposició topo-geomètric (*Topo-*

*Geometrical Superposition Algorithm, TGSA*). Ambdues metodologies eviten les arbitrarietats en les correspondències atòmiques i el problema combinatori d'establir-les.

En aquest capítol es descriu l'esquema general del programa MOLSIMIL<sup>18</sup> de càlcul de mesures de semblança. Donat un conjunt de  $n$  molècules, el programa calcula la superposició òptima de cadascun dels parells de compostos involucrats en l'estudi, i dóna com a resultat una matriu de semblança de dimensió  $(n \times n)$ . Posteriorment, la matriu de semblança s'utilitza en la construcció de models matemàtics que relacionen els canvis en l'estructura d'una sèrie química amb les seves activitats. En el capítol 6 es mostren alguns exemples de l'aplicació de les *MQSM* en famílies de molècules amb interès biològic. El mètode d'optimització de l'alineament molecular s'ha deduït a partir d'una solució exacta per a densitats deformades a funcions delta de Dirac.<sup>15</sup> Quan s'aplica a funcions densitat reals, l'algorisme troba *punts maximitzadors* de la funció de semblança, a partir dels quals i mitjançant un mètode de Newton s'assoleix el màxim absolut de semblança. S'entén com a *punt maximitzador* un punt inicial de la recerca amb el mètode de Newton que es troba en la conca d'un màxim de la funció analitzada. En el darrer apartat d'aquest capítol es descriu l'algorisme *TGSA*, i es mostra un exemple comparatiu entre ambdues metodologies. Precisament s'il·lustra un exemple d'alineament molecular on hi intervé un àtom pesant, que ha estat un dels motius del desenvolupament de la metodologia *TGSA*.

## 4.1 Descripció de la posició relativa de les molècules

Siguin  $A$  i  $B$  dues molècules de  $m_A$  i  $m_B$  àtoms respectivament. S'assumeix que l'estructura dels cossos és rígida, i s'elimina en conseqüència la possibilitat que els àtoms puguin patir deformacions. Les coordenades dels àtoms de les dues molècules poden expressar-se de forma matricial mitjançant les matrius  $\mathbf{A}$  i  $\mathbf{B}$ , de dimensions  $(3 \times m_A)$  i  $(3 \times m_B)$ :

$$\mathbf{A} = (\mathbf{a}_1 \ \mathbf{a}_2 \ \dots \ \mathbf{a}_{m_A}) \wedge \mathbf{a}_i = \begin{pmatrix} a_x^{(i)} \\ a_y^{(i)} \\ a_z^{(i)} \end{pmatrix}, \mathbf{B} = (\mathbf{b}_1 \ \mathbf{b}_2 \ \dots \ \mathbf{b}_{m_B}) \wedge \mathbf{b}_i = \begin{pmatrix} b_x^{(i)} \\ b_y^{(i)} \\ b_z^{(i)} \end{pmatrix} \quad (4.2)$$

En l'equació (4.1) el vector  $\mathbf{Q}$  indica la dependència de les  $MQSM$  respecte a la posició relativa de les molècules. Considerant la molècula  $A$  fixa en l'espai,  $\mathbf{Q}$  inclou les tres translacions més les tres rotacions de la molècula  $B$ . Precisament el moviment de la molècula  $B$  es pot descriure a partir d'una geometria de referència  $\mathbf{B}_0$  segons l'equació:

$$\mathbf{B} = \mathbf{E}(\phi, \theta, \chi) \mathbf{B}_0 + \mathbf{T} \mathbf{1}^T, \quad (4.3)$$

on  $\mathbf{B}_0$  correspon a l'orientació original de la molècula traslladada de manera que el centre de masses i el de coordenades universals coincideixin. Les rotacions de la molècula  $B$  es realitzen sobre les coordenades  $\mathbf{B}_0$ , i es poden representar mitjançant la matriu de transformació  $\mathbf{E}(\phi, \theta, \chi)$ ,

$$\mathbf{B} = \mathbf{E}(\phi, \theta, \chi) \mathbf{B}_0 = \begin{pmatrix} c\phi c\theta c\chi - s\phi s\chi & -c\phi c\theta s\chi - s\phi c\chi & c\phi s\theta \\ s\phi c\theta c\chi + c\phi s\chi & -s\phi c\theta s\chi + c\phi c\chi & s\phi s\theta \\ -s\theta c\chi & s\theta s\chi & c\theta \end{pmatrix} \begin{pmatrix} b_{x,0}^{(1)} & b_{x,0}^{(2)} & \dots & b_{x,0}^{(m_B)} \\ b_{y,0}^{(1)} & b_{y,0}^{(2)} & \dots & b_{y,0}^{(m_B)} \\ b_{z,0}^{(1)} & b_{z,0}^{(2)} & \dots & b_{z,0}^{(m_B)} \end{pmatrix} \quad (4.4)$$

que es defineix sobre els eixos de coordenades que descriuen els angles d'Euler  $(\phi, \theta, \chi)$ . Una vegada orientada la molècula respecte als eixos de rotació propis, se suma a les coordenades moleculars el vector de translació  $\mathbf{T} = (T_x \ T_y \ T_z)^T$  com mostra l'equació (4.3), essent  $\mathbf{1}^T$  un vector filera d'uns de dimensió  $(1 \times m_B)$ . Així la posició i orientació de la molècula  $B$  en qualsevol punt de l'espai es determina mitjançant un vector de sis components,  $\mathbf{Q} = (T_x, T_y, T_z, \phi, \theta, \chi)$ , format pel vector de posició referit al centre de masses més els valors dels angles d'Euler.

Una altra manera de descriure la posició relativa de la molècula  $B$  respecte a la molècula  $A$  és fixant la posició de tres parelles d'àtoms de les dues molècules. S'utilitza la nomenclatura  $\{a, b, a', b', a'', b''\}$ , on  $\{a, a', a''\} \in A$  i  $\{b, b', b''\} \in B$ , i s'interpreta com l'acció de sobreposar l'àtom  $b$  sobre el  $a$ , que és la translació de la molècula, orientar l'àtom  $b'$  de manera que els eixos descrits pels segments  $\overline{aa'}$  i  $\overline{bb'}$  coincideixin, i finalment rotar  $b''$  al voltant de l'eix  $\overline{bb'}$  per fer que el pla descrit pels àtoms  $\{b, b', b''\}$  sigui el mateix que el descrit per  $\{a, a', a''\}$ . Amb aquestes operacions queda fixat el moviment de la molècula  $B$ . El conjunt d'àtoms  $\{a, b, a', b', a'', b''\}$  expressa completament, igual que el vector  $\mathbf{Q}$ , l'arranjament relatiu de les dues molècules.



## 4.2 Superposició molecular referent al màxim de la MQSM

La hipersuperfície de mesures de semblança que s'ha d'explorar conté una gran quantitat de màxims locals. Els mapes de mesures de semblança quàntica,<sup>19,20</sup> han permès comprovar que quan dos àtoms diferents d'hidrogen se sobreposen donen un màxim en la hipersuperfície de mesures de semblança. Això és així perquè els màxims de semblança se situen en els punts de màxima densitat electrònica, que coincideixen amb les posicions dels nuclis atòmics. Aquestes observacions del comportament de la funció de semblança han estat útils per idear l'algorisme de superposició molecular referent al màxim de la mesura  $Z_{AB}(\Omega; \mathbf{Q})$ . A més han permès constatar la gran dependència dels àtoms pesants en el càlcul de les MQSM, la qual cosa suposa en alguns casos que la superposició de màxima semblança no sigui la que es deduiria intuïtivament observant l'estructura química de les molècules comparades.

S'ha dissenyat un algorisme específic de les MQSM, útil per localitzar *punts maximitzadors* de la funció de semblança. La tècnica es deriva de la solució global coneguda per a densitats infinitament compactades.<sup>15</sup> La seva aplicació a densitats reals, no compactades, està justificada per la seva eficàcia i pels valors màxims trobats. Dins el mateix algorisme s'han desenvolupat diferents nivells de càlcul, en els quals les superposicions moleculars poc favorables són descartades mitjançant criteris de semblança atòmica.

### 4.2.1 Deducció de l'algorisme de superposició del màxim de semblança

Per a la deducció de l'algorisme de maximització global de la funció  $Z_{AB}(\Omega; \mathbf{Q})$  es proposa la deformació de les densitats electròniques moleculars fins a fer-les col·lapsar a funcions deltes de Dirac. Aleshores, el conjunt dels valors no nuls de  $Z_{AB}(\Omega; \mathbf{Q})$  és finit i la identificació del màxim absolut factible. Introduint un factor de deformació  $t$  en la definició de la PASA DF s'obté:

$$\mathbf{r}_A^{PASA}(\mathbf{r}; t) = \sum_{a \in A} P_a \sum_{i \in a} w_i |s_i(\mathbf{r} - \mathbf{r}_a; t)|^2, \quad (4.5)$$

que afecta les funcions esfèriques,

$$|s_i(\mathbf{r}-\mathbf{r}_a;t)|^2 = \left( \frac{2tz_i}{\pi} \right)^{3/2} \exp(-2tz_i|\mathbf{r}-\mathbf{r}_a|^2) \quad (4.6)$$

de manera que quan  $t$  tendeix a infinit la funció esdevé una funció delta de Dirac, mentre que quan és la unitat correspon a la capa no deformada. Així doncs, en el cas límit de  $t$  tendint a infinit, la nova funció densitat se simplifica a

$$\tilde{\mathbf{r}}_A(\mathbf{r}) \equiv \lim_{t \rightarrow \infty} \mathbf{r}_A^{PASA}(\mathbf{r};t) = \sum_{a \in A} P_a \mathbf{d}(\mathbf{r}-\mathbf{r}_a), \quad (4.7)$$

on  $P_a$  és la càrrega total definida sobre l'àtom  $a$ .

En l'equació (4.7), la funció associada al límit de la funció densitat s'ha marcat amb una til·la. La *MQSM* de solapament entre les funcions densitat deformades queda reduïda a l'expressió

$$\tilde{Z}_{AB}(\mathbf{Q}) = \sum_a P_a \sum_b P_b \mathbf{d}(\mathbf{r}_a - \mathbf{r}_b(\mathbf{Q})), \quad (4.8)$$

que únicament és diferent de zero quan hi ha com a mínim un àtom de la molècula  $A$  que se superposa amb un àtom de la molècula  $B$ . Els valors no nuls de  $\tilde{Z}_{AB}(\mathbf{Q})$  són finits i es poden classificar en els tres conjunts següents:

- a) El conjunt de valors de  $\tilde{Z}_{AB}(\mathbf{Q})$  corresponents a les posicions de la molècula  $B$  tal que un àtom  $b$  se sobreposa a un àtom  $a$  de la molècula  $A$ , i s'escriu

$$\mathbf{Z}_{ab} \equiv \{P_a P_b \mathbf{d}(0)\} \quad (4.9)$$

De tot el conjunt de solucions el valor màxim  $\tilde{Z}_{ab}^*$  és

$$\tilde{Z}_{ab}^* \equiv \text{Max}(\mathbf{Z}_{ab}) \quad (4.10)$$

- b) El conjunt de possibles valors de  $\tilde{Z}_{AB}(\mathbf{Q})$  donats per la coincidència de les parelles d'àtoms  $(ab)$  i  $(a'b')$

$$\mathbf{Z}_{aba'b'} \equiv \{ (P_a P_b + P_{a'} P_{b'}) \mathbf{d}(0) \mid R_{aa'} = R_{bb'} \} \quad (4.11)$$

L'existència d'aquest conjunt està limitada per la presència de distàncies iguals entre les parelles d'àtoms, tal com s'indica en la definició anterior. També és possible identificar-ne el valor màxim

$$\tilde{Z}_{aba'b'}^* \equiv \text{Max}(\mathbf{Z}_{aba'b'}) \quad (4.12)$$

- c) El conjunt de possibles valors de  $\tilde{Z}_{AB}(\mathbf{Q})$  donats per la coincidència simultània de, com a mínim, les tres parelles  $(ab)$ ,  $(a'b')$  i  $(a''b'')$

$$\mathbf{Z}_{aba'b'a''b''} \equiv \left\{ \begin{aligned} & \{ (P_a P_b + P_{a'} P_{b'} + P_{a''} P_{b''}) \mathbf{d}(0) + Q_{aba'b'a''b''} \\ & | R_{aa'} = R_{bb'} \wedge R_{aa''} = R_{bb''} \wedge R_{a'a''} = R_{b'b''} \} \end{aligned} \right. \quad (4.13)$$

Com en el cas anterior, l'existència d'aquest conjunt està condicionada a la presència de les igualtats de distàncies indicades. També, a causa del fet que les molècules  $A$  i  $B$  es prenen rígides en la maximització de la semblança i que el conjunt  $\{a, b, a', b', a'', b''\}$  equival al vector  $\mathbf{Q}$  de les sis variables que indiquen l'orientació relativa d' $A$  i  $B$ , no es pot definir, en general, un conjunt  $\mathbf{Z}_{aba'b'a''b''}$  de tres i només tres superposicions simultànies. Cal, doncs, introduir-hi el terme  $Q_{aba'b'a''b''}$  que modificarà els valors de semblança en el supòsit que la superposició  $\{a, b, a', b', a'', b''\}$  impliqui també la coincidència de més àtoms:

$$Q_{aba'b'a''b''} \equiv \sum_{a'' \neq a, a', a''} \sum_{b'' \neq b, b', b''} P_{a''} P_{b''} \mathbf{d}(\mathbf{r}_a - \mathbf{r}_b(a, a', a'', b, b', b'')). \quad (4.14)$$

El valor màxim del subconjunt  $\mathbf{Z}_{aba'b'a''b''}$  és:

$$\tilde{Z}_{aba'b'a''b''}^* \equiv \text{Max}(\mathbf{Z}_{aba'b'a''b''}) \quad (4.15)$$

En els tres conjunts definits anteriorment estan inclosos tots els valors de  $\tilde{Z}_{AB}(\mathbf{Q})$  diferents de zero. El màxim absolut de la funció límit correspon al valor més gran dels valors dels màxims dels subconjunts (a), (b) i (c):

$$\tilde{Z}_{AB}^* = \text{Max}(\tilde{Z}_{ab}^*, \tilde{Z}_{aba'b'}^*, \tilde{Z}_{aba'b'a''b''}^*) \quad (4.16)$$

Un possible algorisme de recerca del màxim absolut s'exposa en la taula 4.1. Aquest algorisme, amb algunes modificacions, és el fonament de l'esquema de maximització de la mesura  $Z_{AB}(\Omega; \mathbf{Q})$  que es presentarà en la secció següent.

---

### Subrutina MÀXIM DE SEMBLANÇA

- ✓ Donades les coordenades de la molècula A, **A**
- ✓ Donades les coordenades de la molècula B, **B**
- Inicialitza  $\tilde{Z}_{AB}^* = 0$
- Per tot  $a \in A$
- Per tot  $b \in B$ 
  - Calcula  $\tilde{Z}_{AB}^* = \text{Max}(\tilde{Z}_{AB}^*, P_a P_b \mathbf{d}(0))$
  - Per tot  $a' \in A \mid a' \neq a$
  - Per tot  $b' \in B \mid b' \neq b$ 
    - Si  $R_{aa'} = R_{bb'}$  llavors
      - Calcula  $\tilde{Z}_{AB}^* = \text{Max}(\tilde{Z}_{AB}^*, P_a P_b \mathbf{d}(0) + P_{a'} P_{b'} \mathbf{d}(0))$
      - Per tot  $a'' \in A \mid a'' \neq a' \wedge a'' \neq a$
      - Per tot  $b'' \in B \mid b'' \neq b' \wedge b'' \neq b$ 
        - Si  $R_{aa''} = R_{bb''} \wedge R_{a'a''} = R_{b'b''}$  llavors
          - Traslada  $b$  sobre  $a$
          - Orienta  $\overline{bb'}$  amb  $\overline{aa'}$
          - Rota al voltant de  $\overline{bb'}$  fins  $R_{a''b''} = 0$
          - Calcula  $\tilde{Z}_{AB}^* = \text{Max}(\tilde{Z}_{AB}^*, \tilde{Z}_{AB}^*(a, b, a', b', a'', b''))$
        - Fi condicional
          - Fi per tot  $b''$
          - Fi per tot  $a''$
      - Fi condicional
        - Fi per tot  $b'$
        - Fi per tot  $a'$
    - Fi per tot  $b$
    - Fi per tot  $a$
    - ❖ Retorna màxim global  $\tilde{Z}_{AB}^*$

---

**Taula 4.1** Algorisme de maximització global de les MQSM en el límit de densitats compactades

### 4.2.2 Maximització de les MQSM

S'ha demostrat que quan s'utilitzen funcions densitat deformades a deltes de Dirac el mètode de maximització proposat proporciona el màxim global de la funció de semblança. Quan s'aplica aquest algorisme a sistemes reals com són les densitats ASA, es fa la hipòtesi que la posició de la molècula  $B$  que s'obté al final del procés correspon a un *punt maximitzador* del màxim absolut de la funció de semblança. A partir d'aquest punt, i mitjançant un mètode de Newton, es pot assolir el màxim absolut de la mesura  $Z_{AB}(\Omega; \mathbf{Q})$ .

En la taula 4.2 es mostra l'esquema general de l'algorisme de maximització, que presenta algunes modificacions amb referència al cas de les densitats límit. Les més importants es troben en les dues comandes condicionals de l'algorisme. Mentre que en el cas límit és necessari que es compleixi estrictament que les distàncies entre les parelles d'àtoms siguin iguals, en el cas real de la funció contínua  $Z_{AB}(\Omega; \mathbf{Q})$  aquests restriccions són expressades per les desigualtats

$$\begin{aligned} Z_{a'b'}(\text{Min}(R_{a'b'})) &> \varepsilon_1 \\ Z_{a''b''}(\text{Min}(R_{a''b''})) &> \varepsilon_2 \end{aligned} \quad (4.17)$$

que depenen dels paràmetres l·lindars  $\varepsilon_1$  i  $\varepsilon_2$ , la determinació dels quals és empírica i a la pràctica evita un gran nombre de sobreposicions. Les distàncies mínimes entre els àtoms  $a'b'$  i  $a''b''$  es poden avaluar simplement a partir de les distàncies interatòmiques segons les fórmules:

$$\text{Min}(R_{a'b'}) = |R_{aa'} - R_{bb'}| \quad (4.18)$$

$$\text{Min}(R_{a''b''}) = \left[ (x_{aa''} - x_{bb''})^2 + (y_{aa''} - y_{bb''})^2 \right]^{1/2} \quad (4.19)$$

on

$$x_{aa''} = (R_{aa'}^2 + R_{aa''}^2 - R_{a'a''}^2) / (2R_{aa'}) \quad \wedge \quad y_{aa''} = (R_{aa''}^2 - x_{aa''}^2)^{1/2}, \quad (4.20)$$

i expressions similars per  $x_{bb''}$  i  $y_{bb''}$ . Determinar el valor de la semblança entre dos àtoms  $a$  i  $b$ , coneguda la distància que els separa, es pot fer mitjançant la funció:

$$Z_{ab}(\Omega; R_{ab}) = P_a P_b \sum_{i \in a} w_i \sum_{j \in b} w_j Z_{ij}(\Omega; R_{ab}), \quad (4.21)$$

on les integrals  $Z_{ij}$  entre una capa  $i \in a$  i una capa  $j \in b$  s'han definit en les equacions (3.57) i (3.85) per les mesures de solapament i Coulomb respectivament.

---

### Subrutina MÀXIM DE SEMBLANÇA

- ✓ Donades les coordenades de la molècula  $A$ ,  $\mathbf{A}$ , i la seva PASA DF,  $(n_A, \mathbf{F}_A, \mathbf{w}_A)$
- ✓ Donades les coordenades de la molècula  $B$ ,  $\mathbf{B}$ , i la seva PASA DF,  $(n_B, \mathbf{F}_B, \mathbf{w}_B)$
- Inicialitza  $\tilde{Z}_{AB}^* = 0$
- Per tot  $a \in A$
- Per tot  $b \in B$ 
  - Traslada  $b$  sobre  $a$
  - Per tot  $a' \in A \mid a' \neq a$
  - Per tot  $b' \in B \mid b' \neq b$ 
    - Si  $Z_{a'b'}(\text{Min}(R_{a'b'})) > \epsilon_1$  llavors
      - Orienta  $\overline{bb'}$  amb  $\overline{aa'}$
      - Per tot  $a'' \in A \mid a'' \neq a' \wedge a'' \neq a$
      - Per tot  $b'' \in B \mid b'' \neq b' \wedge b'' \neq b$ 
        - Si  $Z_{a''b''}(\text{Min}(R_{a''b''})) > \epsilon_2$  llavors
          - Rotar al voltant de  $\overline{bb'}$  fins mínim  $R_{a''b''}$
          - Calcula  $Z_{AB}^* = \text{Max}(Z_{AB}^*, Z_{AB}(a, b, a', b', a'', b''))$
      - Fi condicional
        - Fi per tot  $b''$
        - Fi per tot  $a''$
    - Fi condicional
      - Fi per tot  $b'$
      - Fi per tot  $a'$
  - Fi per tot  $b$
  - Fi per tot  $a$
  - ❖ Retorna maximitzador absolut  $Z_{AB}^* = Z_{AB}(a^*, b^*, a'^*, b'^*, a''^*, b''^*)$

---

**Taula 4.2** Algorisme de maximització quasiglobal de les MQSM

El procés descrit a la taula 4.2 és molt costós, d'ordre  $m_A^3 m_B^3$ . Ha estat possible desacoblar els bucles anuats de la taula 4.2 i obtenir algorismes simplificats però alhora també altament robustos.<sup>15</sup> Així s'han desenvolupat dos algorismes paral·lels al general però d'ordre inferior,  $m_A^2 m_B^2$  i  $m_A m_B$ . L'eliminació dels bucles es fa introduint criteris de semblança entre àtoms per seleccionar les parelles òptimes. Aquestes simplificacions fan que el procés sigui uns quants ordres de magnitud més ràpid. En la taula 4.3 es mostra la primera aproximació. La principal diferència respecte a l'algorisme general de la taula 4.2 és la supressió del bucle més intern.

---

### Subrutina MÀXIM DE SEMBLANÇA

- ✓ Donades les coordenades de la molècula A,  $\mathbf{A}$ , i la seva PASA DF,  $(n_A, \mathbf{F}_A, \mathbf{w}_A)$
- ✓ Donades les coordenades de la molècula B,  $\mathbf{B}$ , i la seva PASA DF,  $(n_B, \mathbf{F}_B, \mathbf{w}_B)$
- ✓ Donada una matriu  $(m_A \times m_B)$  de mesures atòmiques a distància zero  $\mathbf{Z} = \{Z_{ab}(0)\}$
- Inicialitza  $\tilde{Z}_{AB}^* = 0$
- Per tot  $a \in A$
- Per tot  $b \in B$ 
  - Per tot  $a' > a$
  - Per tot  $b' > b$ 
    - Redefineix  $a, b, a', b' \mid Z_{ab}(0) > Z_{a'b'}(0)$
    - Si  $Z_{a'b'}(\text{Min}(R_{a'b'})) > \epsilon_1$  llavors
      - Defineix  $a''$  i  $b''$  que maximitzin  $Z_{a''b''}(\text{Min}(R_{a''b''})) \quad \forall a'' \wedge \forall b''$
      - Traslada  $b$  sobre  $a$
      - Orienta  $\overline{bb'}$  amb  $\overline{aa'}$
      - Rotar al voltant de  $\overline{bb'}$  fins mínim  $R_{a''b''}$
      - Calcula  $Z_{AB}^* = \text{Max}(Z_{AB}^*, Z_{AB}(a, b, a', b', a'', b''))$
    - Fi condicional
  - Fi per tot  $b'$
  - Fi per tot  $a'$
- Fi per tot  $b$
- Fi per tot  $a$
- ❖ Retorna una estimació del maximitzador absolut  $Z_{AB}^* = Z_{AB}(a^*, b^*, a^*, b^*)$

---

**Taula 4.3** Algorisme de maximització quasiglobal de les MQSM aproximació II

Enlloc d'eliminar el bucle més intern, el que es fa és traslladar el càlcul de la mesura de semblança  $Z_{AB}(\Omega; \mathbf{Q})$  en el segon bucle, i determinar els àtoms  $a''$  i  $b''$  de manera que es maximitzi la mesura  $Z_{a''b''}(\text{Min}(R_{a''b''}))$ , per tot  $a'' \in A$  i per tot  $b'' \in B$ .

Per últim s'ha dissenyat un algorisme més simplificat, d'ordre  $m_A m_B$ , on les parelles  $(a'b')$  i  $(a''b'')$  es dedueixen aplicant criteris de mesures de semblança entre àtoms. Seguint un procediment semblant a l'algorisme descrit en la taula 4.3, se seleccionen les parelles  $(a'b')$  i  $(a''b'')$  a través dels valors màxims de semblança  $Z_{a'b'}$  i  $Z_{a''b''}$  a la mínima distància que es poden trobar els respectius nuclis atòmics. Aquest algorisme simplificat correspon al descrit en la taula 4.4.

---

### Subrutina MÀXIM DE SEMBLANÇA

- ✓ Donades les coordenades de la molècula  $A$ ,  $\mathbf{A}$ , i la seva PASA DF,  $(n_A, \mathbf{F}_A, \mathbf{w}_A)$
- ✓ Donades les coordenades de la molècula  $B$ ,  $\mathbf{B}$ , i la seva PASA DF,  $(n_B, \mathbf{F}_B, \mathbf{w}_B)$
- ✓ Donada una matriu  $(m_A \times m_B)$  de mesures atòmiques a distància zero  $\mathbf{Z} = \{Z_{ab}(0)\}$
- Inicialitza  $\tilde{Z}_{AB}^* = 0$
- Per tot  $a \in A$
- Per tot  $b \in B$ 
  - Defineix  $a'$  i  $b'$  que maximitzin  $Z_{a'b'}(\text{Min}(R_{a'b'})) \quad \forall a' \wedge \forall b'$
  - Redefineix  $a, b, a', b' \mid Z_{ab}(0) > Z_{a'b'}(0)$
  - Defineix  $a''$  i  $b''$  que maximitzin  $Z_{a''b''}(\text{Min}(R_{a''b''})) \quad \forall a'' \wedge \forall b''$
  - Traslada  $b$  sobre  $a$
  - Orienta  $\overline{bb'}$  amb  $\overline{aa'}$
  - Rotar al voltant de  $\overline{bb'}$  fins mínim  $R_{a''b''}$
  - Calcula  $Z_{AB}^* = \text{Max}(Z_{AB}^*, Z_{AB}(a, b, a', b', a'', b''))$
- Fi per tot  $b$
- Fi per tot  $a$
- ❖ Retorna una estimació del maximitzador absolut  $Z_{AB}^* = Z_{AB}(a^*, b^*)$

---

**Taula 4.4** Algorisme de maximització quasiglobal de les MQSM aproximació I



### 4.3 Derivades analítiques de les MQSM definides sobre funcions ASA

En l'equació (2.11) s'han definit d'una manera general la mesura de semblança entre dues molècules  $A$  i  $B$  emprant l'aproximació ASA com una suma de les integrals  $Z_{ij}$  entre una capa  $i$  pertany a l'àtom  $a$  de la molècula  $A$ , i una capa  $j$  pertany a l'àtom  $b$  de la molècula  $B$ . Incloent el vector  $\mathbf{Q}$ , la mesura de semblança es pot expressar:

$$Z_{AB}(\Omega; \mathbf{Q}) = \sum_{i \in A} \sum_{j \in B} w_i w_j Z_{ij}(\Omega; \mathbf{Q}). \quad (4.22)$$

Si s'utilitzen funcions PASA, l'expressió (4.22) és totalment vàlida, i s'ha de considerar que els coeficients  $w_i$  i  $w_j$  estan multiplicats per les càrregues atòmiques  $P_a$  i  $P_b$  respectivament.

Seguidament es descriuran les primeres i segones derivades de les mesures  $Z_{ij}$  de solapament i de Coulomb respecte a les sis components del vector  $\mathbf{Q}$ , les tres rotacions respecte als angles d'Euler  $(\phi, \theta, \chi)$  i les tres translacions  $(T_x, T_y, T_z)$  de la molècula  $B$ . Les derivades de la mesura global  $Z_{AB}$  s'obtenen d'aplicar els sumatoris corresponents per a cada molècula sobre les derivades de  $Z_{ij}$ , i multiplicades pels corresponents coeficients. Per simplificar la notació es defineix la constant

$$\mathbf{k} = \left( \frac{2\mathbf{z}_i \mathbf{z}_j}{(\mathbf{z}_i + \mathbf{z}_j)} \right) \quad (4.23)$$

i llavors la mesura de solapament entre les capes  $i$  i  $j$  definida en l'equació (3.57) és

$$Z_{ij} = \left( \frac{\mathbf{k}}{\mathbf{p}} \right)^{3/2} \exp(-\mathbf{k}R_{ab}^2), \quad (4.24)$$

mentre que la mesura de Coulomb descrita en (3.85) es pot expressar com

$$Z_{ij}(\mathbf{r}_{12}^{-1}) = 2\sqrt{\frac{\mathbf{k}}{\mathbf{p}}} F_0(\mathbf{k}R_{ab}^2). \quad (4.25)$$

Ambdues mesures són funció de la distància interatòmica al quadrat,

$$R_{ab}^2 = (a_x - b_x)^2 + (a_y - b_y)^2 + (a_z - b_z)^2, \quad (4.26)$$

i la deriva respecte a qualsevol dels sis components del vector  $\mathbf{Q}$  afectarà només a la variable  $R_{ab}^2$  dels termes  $Z_{ij}(\Omega)$ . L'equació (4.3) descriu la posició relativa de la molècula  $B$  respecte del vector  $\mathbf{Q}$  i referida a una geometria fixa  $\mathbf{B}_0$ . Les coordenades d'un àtom  $b$ , en qualsevol punt de l'espai, es calculen a través de l'expressió:

$$\begin{pmatrix} b_x \\ b_y \\ b_z \end{pmatrix} = \mathbf{E}(\phi, \theta, \chi) \begin{pmatrix} b_{x,0} \\ b_{y,0} \\ b_{z,0} \end{pmatrix} + \begin{pmatrix} T_x \\ T_y \\ T_z \end{pmatrix} \quad (4.27)$$

Per simplificar la notació, expressem la matriu de rotació a partir de components simples,  $\mathbf{E}(\phi, \theta, \chi) = \{e_{ij}\}$ , on la definició de cada element  $e_{ij}$  es pot deduir de l'equació (4.4). Llavors les coordenades de qualsevol àtom  $b$  de la molècula  $B$  es poden escriure conforme a:

$$\begin{cases} b_x = e_{11}b_{x,0} + e_{12}b_{y,0} + e_{13}b_{z,0} + T_x = E_x + T_x \\ b_y = e_{21}b_{x,0} + e_{22}b_{y,0} + e_{23}b_{z,0} + T_y = E_y + T_y \\ b_z = e_{31}b_{x,0} + e_{32}b_{y,0} + e_{33}b_{z,0} + T_z = E_z + T_z \end{cases} \quad (4.28)$$

Substituint l'expressió (4.28) en la definició de  $R_{ab}^2$  s'obté:

$$R_{ab}^2 = (a_x - E_x - T_x)^2 + (a_y - E_y - T_y)^2 + (a_z - E_z - T_z)^2. \quad (4.29)$$

### 4.3.1 Primera derivada de les MQSM

La primera derivada de la integral  $Z_{ij}$  de solapament respecte a qualsevol dels components del vector  $\mathbf{Q}$  té la forma

$$\frac{\partial Z_{ij}}{\partial u} = -\mathbf{k} Z_{ij} \frac{\partial R_{ab}^2}{\partial u} \quad \forall u \in \mathbf{Q} = (T_x, T_y, T_z, \mathbf{f}, \mathbf{q}, \mathbf{c}). \quad (4.30)$$

Mentre que la mesura de Coulomb serà:

$$\frac{\partial Z_{ij}(\mathbf{r}_{12}^{-1})}{\partial u} = -2\mathbf{k} \sqrt{\mathbf{k}/\pi} F_1(\mathbf{k} R_{ab}^2) \frac{\partial R_{ab}^2}{\partial u}. \quad (4.31)$$

En les expressions (4.30) i (4.31) apareix la derivada respecte a la distància entre els àtoms  $a$  i  $b$  al quadrat.

**Primera derivada de  $R_{ab}^2$  respecte als components del vector de translacions.** Són les equacions més simples:

$$\frac{\partial R_{ab}^2}{\partial T_x} = -2(a_x - b_x); \quad \frac{\partial R_{ab}^2}{\partial T_y} = -2(a_y - b_y); \quad \frac{\partial R_{ab}^2}{\partial T_z} = -2(a_z - b_z). \quad (4.32)$$

**Primera derivada de  $R_{ab}^2$  respecte als angles d'Euler.** La fórmula general per a qualsevol dels tres angles ( $\phi, \theta, \chi$ ) és

$$\frac{\partial R_{ab}^2}{\partial \alpha} = -2 \left[ (a_x - b_x) \frac{\partial E_x}{\partial \alpha} + (a_y - b_y) \frac{\partial E_y}{\partial \alpha} + (a_z - b_z) \frac{\partial E_z}{\partial \alpha} \right] \quad (4.33)$$

És possible demostrar que:

$$E_x \frac{\partial E_x}{\partial \alpha} + E_y \frac{\partial E_y}{\partial \alpha} + E_z \frac{\partial E_z}{\partial \alpha} = 0 \quad \forall \alpha = \phi, \theta \text{ i } \chi. \quad (4.34)$$

Aleshores l'equació (4.33) se simplifica a l'expressió:

$$\frac{\partial R_{ab}^2}{\partial \alpha} = -2 \left[ (a_x - T_x) \frac{\partial E_x}{\partial \alpha} + (a_y - T_y) \frac{\partial E_y}{\partial \alpha} + (a_z - T_z) \frac{\partial E_z}{\partial \alpha} \right], \quad (4.35)$$

la qual serà molt útil en la deducció de les segones derivades.

Les respectives derivades per a cadascun dels angles d'Euler són:

$$\frac{\partial E_x}{\partial \phi} = -E_y, \quad \frac{\partial E_y}{\partial \phi} = E_x, \quad \frac{\partial E_z}{\partial \phi} = 0, \quad (4.36)$$

$$\begin{aligned} \frac{\partial E_x}{\partial \theta} &= -c\phi s\theta c\chi b_{x,0} + c\phi s\theta s\chi b_{y,0} + c\phi c\theta b_{z,0} \\ \frac{\partial E_y}{\partial \theta} &= -s\phi s\theta c\chi b_{x,0} + s\phi s\theta s\chi b_{y,0} + s\phi c\theta b_{z,0} \\ \frac{\partial E_z}{\partial \theta} &= -c\theta c\chi b_{x,0} + c\theta s\chi b_{y,0} - s\theta b_{z,0} \end{aligned} \quad (4.37)$$

i

$$\frac{\partial E_x}{\partial \chi} = e_{12}b_{x,0} - e_{11}b_{y,0}; \quad \frac{\partial E_y}{\partial \chi} = e_{22}b_{x,0} - e_{21}b_{y,0}; \quad \frac{\partial E_z}{\partial \chi} = e_{32}b_{x,0} - e_{31}b_{y,0} \quad (4.38)$$

### 4.3.2 Segona derivada de les MQSM

La segona derivada de la integral  $Z_{ij}$  de solapament té la forma

$$\begin{aligned} \frac{\partial^2 Z_{ij}}{\partial v \partial u} &= \mathbf{k}^2 Z_{ij} \frac{\partial R_{ab}^2}{\partial v} \frac{\partial R_{ab}^2}{\partial u} - \mathbf{k} Z_{ij} \frac{\partial^2 R_{ab}^2}{\partial v \partial u} = \mathbf{k} Z_{ij} \left( \mathbf{k} \frac{\partial R_{ab}^2}{\partial v} \frac{\partial R_{ab}^2}{\partial u} - \frac{\partial^2 R_{ab}^2}{\partial v \partial u} \right) \\ &= Z_{ij}^{-1} \frac{\partial Z_{ij}}{\partial v} \frac{\partial Z_{ij}}{\partial u} - \mathbf{k} Z_{ij} \frac{\partial^2 R_{ab}^2}{\partial v \partial u} \end{aligned} \quad (4.39)$$

essent  $u$  i  $v$  qualsevol de les components del vector  $\mathbf{Q}$ . I referit a la mesura de Coulomb:

$$\frac{\partial^2 Z_{ij}(\mathbf{r}_{12}^{-1})}{\partial v \partial u} = -2\mathbf{k} \sqrt{\mathbf{k}/\pi} \left[ -\mathbf{k} F_2(\mathbf{k} R_{ab}^2) \frac{\partial R_{ab}^2}{\partial v} \frac{\partial R_{ab}^2}{\partial u} + F_1(\mathbf{k} R_{ab}^2) \frac{\partial^2 R_{ab}^2}{\partial v \partial u} \right] \quad (4.40)$$

Ens fixem en el terme  $\frac{\partial^2 R_{ab}^2}{\partial v \partial u}$  que apareix en les dues mesures.

**Segona derivada de  $R_{ab}^2$  respecte a dues components del vector de translacions.** La segona derivada en els termes diagonals és

$$\frac{\partial^2 R_{ab}^2}{\partial T_x \partial T_x} = 2, \quad \frac{\partial^2 R_{ab}^2}{\partial T_y \partial T_y} = 2, \quad \frac{\partial^2 R_{ab}^2}{\partial T_z \partial T_z} = 2, \quad (4.41)$$

mentre que els termes creuats són zero:

$$\frac{\partial^2 R_{ab}^2}{\partial T_v \partial T_u} = 0 \quad \forall v \neq u. \quad (4.42)$$

**Segona derivada respecte a una component del vector de translacions i un angle d'Euler.** Respon a l'equació general:

$$\frac{\partial^2 R_{ab}^2}{\partial \alpha \partial T_u} = 2 \frac{\partial E_u}{\partial \alpha}, \quad (4.43)$$

essent  $\alpha$  qualsevol dels angles d'Euler,  $(\phi, \theta, \chi)$ , i  $u$  qualsevol de les coordenades  $(x, y, z)$ . Les primeres derivades que apareixen en l'expressió (4.43) s'han definit en les equacions (4.36), (4.37) i (4.38).

**Segona derivada respecte a dos angles d'Euler.** A partir de l'equació (4.35) es pot definir l'expressió general:

$$\frac{\partial^2 R_{ab}^2}{\partial \beta \partial \alpha} = -2 \left[ (a_x - T_x) \frac{\partial^2 E_x}{\partial \beta \partial \alpha} + (a_y - T_y) \frac{\partial^2 E_y}{\partial \beta \partial \alpha} + (a_z - T_z) \frac{\partial^2 E_z}{\partial \beta \partial \alpha} \right] \quad (4.44)$$

Les derivades on apareix la variable  $\phi$  són:

$$\frac{\partial^2 E_x}{\partial \phi \partial \phi} = -E_x; \quad \frac{\partial^2 E_y}{\partial \phi \partial \phi} = -E_y; \quad \frac{\partial^2 E_z}{\partial \phi \partial \phi} = 0 \quad (4.45)$$

$$\frac{\partial^2 E_x}{\partial\theta\partial\phi} = -\frac{\partial E_y}{\partial\theta}, \quad \frac{\partial^2 E_y}{\partial\theta\partial\phi} = \frac{\partial E_x}{\partial\theta}, \quad \frac{\partial^2 E_z}{\partial\theta\partial\phi} = 0, \quad (4.46)$$

$$\frac{\partial^2 E_x}{\partial\chi\partial\phi} = -\frac{\partial E_y}{\partial\chi}, \quad \frac{\partial^2 E_y}{\partial\chi\partial\phi} = \frac{\partial E_x}{\partial\chi}, \quad \frac{\partial^2 E_z}{\partial\chi\partial\phi} = 0 \quad (4.47)$$

De la resta, les derivades on es veu involucrada la variable  $\chi$  són:

$$\frac{\partial^2 E_x}{\partial\chi\partial\chi} = -e_{11}b_{x,0} - e_{12}b_{y,0}; \quad \frac{\partial^2 E_y}{\partial\chi\partial\chi} = -e_{21}b_{x,0} - e_{22}b_{y,0}; \quad \frac{\partial^2 E_z}{\partial\chi\partial\chi} = -e_{31}b_{x,0} - e_{32}b_{y,0} \quad (4.48)$$

i

$$\begin{aligned} \frac{\partial^2 E_x}{\partial\theta\partial\chi} &= c\phi s\theta s\chi b_{x,0} + c\phi s\theta c\chi b_{y,0} \\ \frac{\partial^2 E_y}{\partial\theta\partial\chi} &= s\phi s\theta s\chi b_{x,0} + s\phi s\theta c\chi b_{y,0} \\ \frac{\partial^2 E_z}{\partial\theta\partial\chi} &= c\theta s\chi b_{x,0} + c\theta c\chi b_{y,0} \end{aligned} \quad (4.49)$$

I per últim les segones derivades de  $\theta$ :

$$\begin{aligned} \frac{\partial^2 E_x}{\partial\theta\partial\theta} &= -c\phi c\theta c\chi b_{x,0} + c\phi c\theta s\chi b_{y,0} - c\phi s\theta b_{z,0} \\ \frac{\partial^2 E_y}{\partial\theta\partial\theta} &= -s\phi c\theta c\chi b_{x,0} + s\phi c\theta s\chi b_{y,0} - s\phi s\theta b_{z,0} \\ \frac{\partial^2 E_z}{\partial\theta\partial\theta} &= s\theta c\chi b_{x,0} - s\theta s\chi b_{y,0} - c\theta b_{z,0} \end{aligned} \quad (4.50)$$

## 4.4 Esquema general del programa MOLSIMIL

El programa MOLSIMIL calcula la superposició òptima de tots els possibles parells de molècules del conjunt estudiat. Essent  $n$  el nombre de compostos, el programa retorna una matriu de mesures de semblança  $\mathbf{Z}(n \times n)$ . L'entrada de dades del programa és la geometria de les molècules considerades, que pot provenir de càlculs teòrics o de determinacions experimentals. També s'ha d'especificar el tipus de mesura de semblança que es vol realitzar, solapament o Coulomb, i el nivell d'aproximació que s'utilitzarà en la subrutina MAXIM DE SEMBLANÇA. Normalment els hidrògens no es tenen en compte en el procés de superposició molecular. És per això que abans de construir les funcions PASA per a tots els compostos de la sèrie analitzada, es reordenen els àtoms de manera que els hidrògens quedin al final de la matriu de les coordenades i així és més fàcil desestimar-los en la subrutina d'alineament molecular. En la taula 4.5 es mostra l'esquema general del programa MOLSIMIL en forma de pseudo codi.

---

### Programa MOLSIMIL

- ✓ Donat un conjunt  $M$  format per  $n$  molècules
- ✓ Donada la geometria de totes les molècules que pertanyen al conjunt  $M: \{m_A, \mathbf{P}_A, \mathbf{A}\} \forall A \in M$
- Reordena els àtoms  $\forall A \in M$  de manera que els hidrògens quedin al final de cada matriu  $\mathbf{A}$ .
- Demana **Subrutina PASA DF**  $\forall A \in M$ . Necessita:  $m_A, \mathbf{P}_A$  i especificar un conjunt de funcions ASA de base. Retorna:  $\{n_A, \mathbf{w}_A, \mathbf{F}_A\}$
- Per tot  $A \in M$ 
  - Calcula  $Z_{AA}$  i recull el valor en la matriu de semblança  $\mathbf{Z}$
  - Per tot  $B > A \wedge B \in M$ 
    - Demana la **Subrutina MÀXIM DE SEMBLANÇA**. Necessita:  $\mathbf{A}, \mathbf{B}, \{n_A, \mathbf{w}_A, \mathbf{F}_A\}, \{n_B, \mathbf{w}_B, \mathbf{F}_B\}$ . Retorna:  $Z_{AB}^*$ ,  $(a^*, b^*, a'^*, b'^*) \equiv \mathbf{B}^*$
    - Demana la **Subrutina NEWTON**. Necessita:  $\mathbf{A}, \mathbf{B}^*, \{n_A, \mathbf{w}_A, \mathbf{F}_A\}, \{n_B, \mathbf{w}_B, \mathbf{F}_B\}$ . Retorna:  $Z_{AB}^{\max}$
    - Recull el valor  $Z_{AB}^{\max}$  en la matriu de semblança  $\mathbf{Z}$
  - Fi per tot  $B$
- Fi per tot  $A$
- ❖ Sortida del programa: matriu de semblança  $\mathbf{Z} = \{Z_{AB} \forall A, B \in M\}$

---

**Taula 4.5** Esquema general del Programa MOLSIMIL

## 4.5 Algorisme de superposició topo-geomètric

Recentment s'ha desenvolupat en el nostre laboratori un mètode de sobreposició molecular basat en criteris topològics i geomètrics.<sup>17</sup> L'algorisme *TGSA*, descrit en la taula 4.6, selecciona l'alineament molecular en el qual se sobreposen el màxim nombre d'àtoms del mateix tipus. Durant el procés de superposició molecular únicament es consideren les distàncies interatòmiques i el tipus d'àtoms, de manera que no es realitza cap càlcul de la integral de semblança. Això suposa un estalvi computacional molt important si es compara amb el mètode de màxima semblança. A més, la superposició molecular és única per a totes les mesures de semblança que es puguin definir, mentre que en el mètode de màxima semblança els alineaments moleculars depenen de la *MQSM* avaluada.

La subrutina *TGSA* necessita l'especificació de tres paràmetres. El primer és el llindar que determina quan dos àtoms d'una mateixa molècula estan enllaçats. S'utilitza la norma de considerar dos àtoms enllaçats si la distància que els separa és igual o inferior al 110% de la suma dels seus radis covalents. El segon paràmetre, representat per  $\epsilon_1$  en la taula 4.6, permet discernir quan dues distàncies interatòmiques de molècules diferents són comparables. I el tercer llindar,  $\epsilon_2$ , és la tolerància que indica quan dos àtoms de molècules diferents es considera que estan sobreposats.

Per escollir entre dos alineaments que tenen el mateix nombre d'àtoms sobreposats, llavors s'avalua quin dels dos té l'índex

$$T_{AB} = \left[ \frac{(d_{AA}d_{BB})^{1/2}}{d_{AB}} \right]^{1/2} \quad (4.51)$$

superior, definint-se les distàncies intermoleculares com

$$d_{AB} = \sum_i^{m_A} \sum_j^{m_B} \left[ (a_x^{(i)} - b_x^{(j)})^2 + (a_y^{(i)} - b_y^{(j)})^2 + (a_z^{(i)} - b_z^{(j)})^2 \right] \quad (4.52)$$



---

## Subrutina TGSA

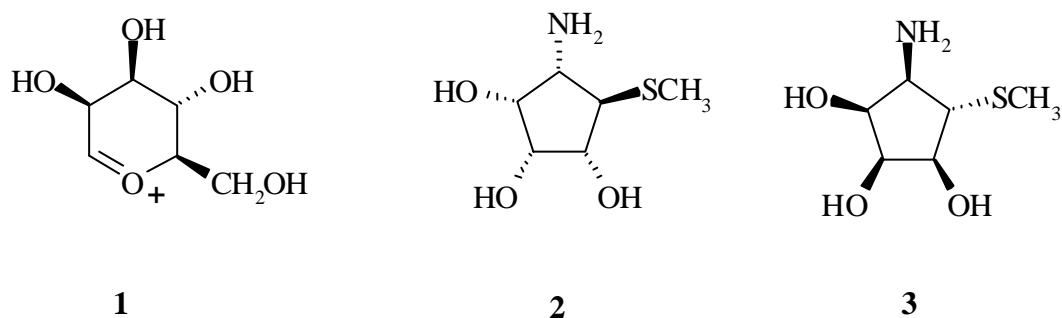
- ✓ Donades les coordenades de la molècula  $A$ ,  $\mathbf{A}$ , nombres atòmics,  $\mathbf{P}_A$ , i la matriu de connectivitats
- ✓ Donades les coordenades de la molècula  $B$ ,  $\mathbf{B}$ , nombres atòmics,  $\mathbf{P}_B$ , i la matriu de connectivitats
- Calcula  $d_{AA}$  i  $d_{BB}$
- Inicialitza  $\mathbf{u}^*=0 \wedge T^*=0$
- Per tot  $a \in A$
- Per tot  $b \in B$ 
  - Traslada  $b$  sobre  $a$ 
    - Per tot  $a' \in A \mid a' \neq a \wedge a'a$  estan enllaçats
    - Per tot  $b' \in B \mid b' \neq b \wedge b'b$  estan enllaçats
      - Si  $|R_{aa'} - R_{bb'}| < \epsilon_1$  llavors
        - Orienta  $\overline{bb'}$  amb  $\overline{aa'}$ 
          - Per tot  $a'' \in A \mid a'' \neq a' \wedge a'' \neq a \wedge (a''a \vee a''a')$  estan enllaçats
          - Per tot  $b'' \in B \mid b'' \neq b' \wedge b'' \neq b \wedge (b''b \vee b''b')$  estan enllaçats
            - Si  $|R_{aa''} - R_{bb''}| < \epsilon_1 \wedge |R_{a'a''} - R_{b'b''}| < \epsilon_1$  llavors
              - Rotar al voltant de  $\overline{bb'}$  fins mínim  $R_{a''b''}$
              - Calcula el nombre d'àtoms  $\mathbf{u}$  del mateix tipus sobreposats, és a dir, amb distància inferior a  $\epsilon_2$
              - Calcula  $T_{AB}$  a partir de l'equació (4.51)
              - Si  $\mathbf{u} > \mathbf{u}^* \vee (\mathbf{u} = \mathbf{u}^* \wedge T_{AB} > T^*)$  llavors
                - $\mathbf{u}^* = \mathbf{u}$ ,  $T^* = T_{AB}$  i guarda l'orientació  $(a^*, b^*, a'^*, b'^*, a''^*, b''^*)$
              - Fi condicional
            - Fi condicional
          - Fi per tot  $b''$
          - Fi per tot  $a''$
        - Fi condicional
      - Fi per tot  $b'$
      - Fi per tot  $a'$
    - Fi per tot  $b$
    - Fi per tot  $a$
    - ❖ Retorna la superposició topo-geomètrica òptima  $(a^*, b^*, a'^*, b'^*, a''^*, b''^*)$

---

**Taula 4.6** Algorisme de superposició topo-geomètrica.<sup>17</sup> Per defecte s'utilitza  $\epsilon_1=0.1$  i  $\epsilon_2=0.3$  a.u.

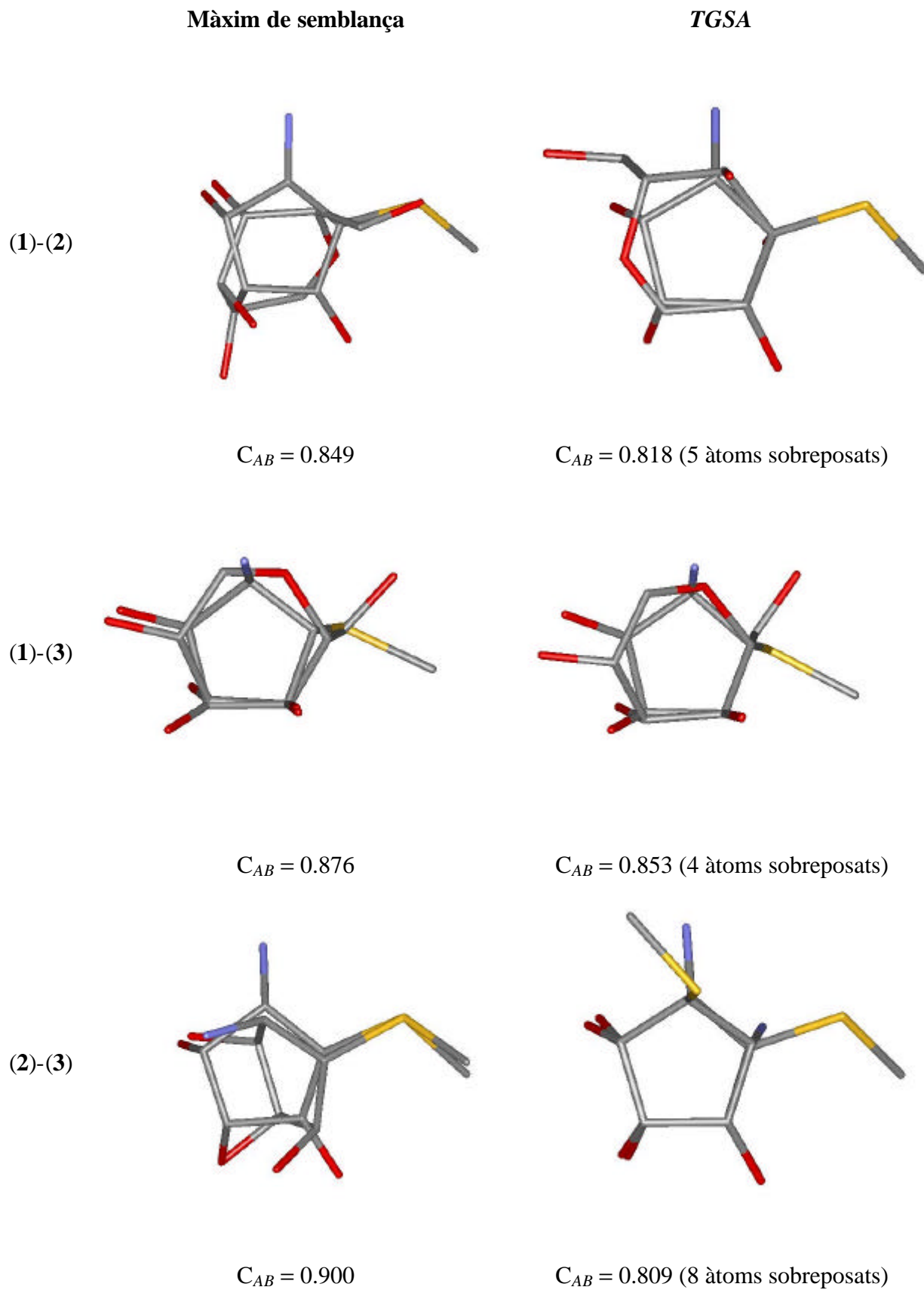
### 4.5.1 Exemple de sobreposició on hi intervenen àtoms pesants

En l'article 3.3 s'ha mostrat un exemple d'un possible intermedi de la reacció d'inhibició de la glicosidasa. Com ja s'ha comentat, en la bibliografia apareix una controvèrsia sobre la possibilitat que el catió mannopiranosil, esquema **1** de la figura 6.1, sigui un intermedi de la reacció d'inhibició de la  $\alpha$ -mannosidasa. Aquesta hipòtesi, avalada per Winkler i Holan,<sup>21,22</sup> els ha permès sintetitzar el compost (+)-mannostatin A (**2**), com un potent inhibidor de la mannosidasa. Però altres autors,<sup>23,24</sup> contradiuen aquest supòsit afirmant que el compost inactiu (-)-mannostatin A (**3**) és més semblant al catió mannosil que el seu enantiòmer **2**.



**Figura 6.1** Estructura molecular dels inhibidors de la mannosidasa.

En l'article 3.3 s'ha estudiat la semblança de les estructures mostrades en la figura 6.1 per comprovar quin dels dos enantiòmers del mannostatin A, **2** o **3**, és més semblant al catió **1**. En la present anàlisi serveix d'exemple il·lustratiu del tipus de superposició molecular que s'obté emprant els algorismes de màxima semblança i *TGSA*. L'interès de la comparació és constatar la influència de l'àtom de sofre dels compostos **2** i **3** en l'alineament molecular del màxim de semblança. La superposició de màxima semblança s'ha deduït amb el programa *MOLSIMIL* i *MQSM* de tipus Coulomb calculades sobre *FMASA DF* adaptades a la densitat molecular *ab initio* HF/6-311. D'altra banda, s'ha determinat la sobreposició topo-geomètrica òptima amb l'algorisme descrit en la taula 4.6, i sobre aquest alineament s'ha fet una mesura puntual de la integral de Coulomb emprant les *FMASA DF*. En la figura 6.2 es mostren les tres superposicions resultants d'ambdues metodologies.



**Figura 6.2** Sobreposicions moleculars dels inhibidors de la mannosidasa.

Les tres superposicions mostrades en la figura 6.2 són molt eloqüents de quins són els màxims de semblança quan es comparen molècules amb àtoms pesants. La sobreposició dels àtoms de sofre dominen l'alineament final de les molècules en el màxim de semblança. El sofre se sobreposa amb un àtom d'oxigen, un de carboni i un altre de sofre, respectivament en les tres *MQSM* estudiades. La diferència més apreciable s'observa en l'alineament dels enantiòmers **2** i **3**. Amb l'algorisme del màxim de semblança els dos àtoms de sofre coincideixen, mentre que l'algorisme *TGSA* sobreposa els dos anells de cinc carbonis i els tres grups hidroxil deixant els sofres sense coincidir amb cap carboni, ni oxigen, ni nitrogen. Segons la metodologia del màxim de semblança les molècules **2** i **3** són les més semblants de la sèrie estudiada, amb un índex de Carbó de 0.9, mentre que per l'algorisme *TGSA* són les més dissemblants, amb un valor de  $C_{AB} = 0.8$ . On coincideixen ambdues metodologies és a assenyalar l'enantiòmer no actiu **3** com el més semblant al catió mannosil, i per tant semblaria posar en dubte la consideració del catió **1** com un intermedi en la reacció d'inhibició de la mannosidasa.



## Discussió

En aquest capítol s'ha proposat un mètode de superposició molecular fonamentat en la recerca del màxim de semblança entre les densitats electròniques de les molècules comparades. L'algorisme desenvolupat explora totes les possibles sobreposicions atòmiques, amb la finalitat de trobar l'alineament molecular en el qual se sobreposa el màxim nombre d'àtoms amb nombre atòmic superior de les dues molècules. Això es deu a la naturalesa de les funcions de densitat electrònica de primer ordre, que presenten màxims en les posicions nuclears. Llavors, sobre el *punt maximitzador* localitzat es fa un refinament mitjançant un mètode de Newton. Per accelerar la recerca del *punt maximitzador* s'han dissenyat diferents nivells de càlcul, en els quals s'utilitzen criteris de semblança entre àtoms per rebutjar les superposicions moleculars menys propícies.

Precisament, un dels inconvenients del mètode de recerca del màxim de semblança és el cost computacional que suposa haver de calcular repetidament la mesura de semblança. Això ha fet inviable de moment realitzar el procés de superposició mitjançant *DF* calculades a nivell *ab initio*, a no ser que s'estudiïn sistemes moleculars de dimensions molt reduïdes. Per aquest motiu la recerca de la sobreposició molecular òptima es fa sempre emprant *DF* aproximades, encara més quan s'estudien sistemes biològics. Una altra crítica és la dependència de les mesures de semblança respecte als àtoms pesants, que en alguns casos, produeix alineaments moleculars de màxima semblança no coincidents amb el que es deduiria intuïtivament observant l'estructura comuna de les molècules comparades. Aquesta ha estat una de les raons del desenvolupament del mètode *TGSA*, on es determina la sobreposició molecular a partir només de distàncies i tipus atòmics. L'existència de les dues metodologies planteja un dilema sobre quin criteri s'ha d'elegir a l'hora de dur a terme els estudis de *MQSM*: definir la mesura de semblança en el màxim o calcular la mesura de semblança sobre la sobreposició de màxima similitud estructural entre les molècules comparades.

Des d'un punt de vista teòric, l'algorisme del màxim de semblança resol el problema del càlcul pràctic de la *MQSM*, que per la seva definició requereix el reconeixement del valor del màxim encobriment de les densitats electròniques de les molècules que s'han de comparar. Aquest tipus de mesura de semblança basada en les

funcions de densitat de primer ordre connecta el concepte de molècula amb el d'un objecte tridimensional, i permet fer la recerca de la sobreposició molecular òptima sense necessitat d'haver d'utilitzar mètodes de combinatòria. A més s'ha resolt un dels inconvenients de les tècniques de correspondències entre àtoms com és el de considerar les molècules com un conjunt de boles, i s'adapta a la distribució de probabilitats de la descripció quàntica.

Des d'un punt de vista químic, quan s'estudia l'activitat biològica de sèries de compostos derivats d'una estructura comuna, on únicament varien uns substituents, cal suposar que la interacció de tots ells amb el receptor és a través d'una conformació i orientació similar, que està supeditada al contorn de la cavitat activa. Aleshores, alineaments moleculars basats en l'estructura comuna semblen més coherents que possibles superposicions moleculars derivades del màxim de semblança, que poden estar influenciades per la presència d'àtoms pesants en els substituents, prescindint de la resta de l'esquelet comú.

Possiblement serà l'exploració sistemàtica de moltes famílies de molècules el que indicarà quin dels dos criteris és més adequat per a una determinada sèrie química. En el capítol 6 s'analitza la influència de la superposició molecular, així com altres factors, en els paràmetres estadístics dels models *QSAR* resultants.

## Referències

1. A. Atai, N. Tomioka, M. Yamada, A. Inoue, Y. Kato. Molecular superposition for rational drug design. Publicat en el llibre: 3D QSAR in drug design: theory methods and applications. H. Kubinyi (ed.). ESCOM Science Publishers B.V., Leiden, The Netherlands, pàgines 200–225, 1993.
2. S. C. Nyburg. Some uses of a best molecular fit routine. *Acta Cryst.* **1974**, *B30*, 251–253.
3. E. M. Rasmussen, G. M. Downs, P. Willett. Automatic classification of chemical structure databases using a highly parallel array processor. *J. Comput. Chem.* **1988**, *9*, 378–386.
4. Y. C. Martin. 3D database searching in drug design. *J. Med. Chem.* **1992**, *35*, 2145–2154.
5. P. Willet. Similarity-searching and clustering algorithms for processing databases of two-dimensional and three-dimensional chemical structures. Publicat en el llibre: Molecular similarity in drug design. P. M. Dean (ed.). Blackie Academic & Professional, London, pàgines 110–137, 1995.
6. J. S. Mason. Experiences with searching for molecular similarity in conformationally flexible 3D databases. Publicat en el llibre: Molecular similarity in drug design. P. M. Dean (ed.). Blackie Academic & Professional, London, pàgines 138–162, 1995.
7. A. D. McLachlan. A mathematical procedure for superimposing atomic coordinates of proteins. *Acta Cryst.* **1972**, *A28*, 656–657.
8. E. Gavuzzo, S. Pagliuca, V. Pavel, C. Quagliata. Generation and best fitting of molecular models. *Acta Cryst.* **1972**, *B28*, 1968–1969.
9. A. D. McLachlan. Rapid comparison of protein structures. *Acta Cryst.* **1982**, *A38*, 871–873.
10. P. R. Gerber, K. Muller. Superimposing several sets of atomic coordinates. *Acta Cryst.* **1987**, *A41*, 426–428.
11. P. K. Redington. Molfit: a computer program for molecular superposition. *Comput. Chem.* **1992**, *16*, 217–222.
12. S. K. Kirkpatrick, G. M. Smith. Optimization by simulated annealing. *Science* **1983**, *220*, 671–680.
13. S. Kirkpatrick. Optimization by simulated annealing – Quantitative studies. *J. Stat. Phys.* **1984**, *34*, 975–986.
14. M. T. Barakat, P. M. Dean. Molecular structure matching by simulated annealing. 1. A comparison between different cooling schedules. *J. Comput.-Aided Mol. Des.* **1990**, *4*, 295–316.
15. P. Constans, Ll. Amat, R. Carbó-Dorca. Toward a global maximization of the molecular similarity function: superposition of two molecules. *J. Comput. Chem.* **1997**, *18*, 826–846.
16. Ll. Amat, R. Carbó, P. Constans. Algorisme d'optimització global de les mesures de semblança quàntica molecular. *SCIENTIA gerundensis* **1996**, *22*, 109–121.
17. X. Gironés, D. Robert, R. Carbó-Dorca. TGSA: a molecular superposition program based on topological considerations. *J. Comput. Chem.* **2001**, *22*, 255–263.
18. Programa MOLSIMIL97. Ll. Amat, P. Constans, E. Besalú, R. Carbó-Dorca. Institut de química computacional, Universitat de Girona, 1997.
19. R. Carbó-Dorca, E. Besalú, Ll. Amat, X. Fradera. Quantum molecular similarity measures: concepts, definitions, and applications to quantitative structure-property relationships. Publicat en el llibre:



- Advances in molecular similarity. R. Carbó-Dorca, P. G. Mezey (Eds.). JAI Press, London, volum 1, pàgines 1-41, 1996.
20. Ll. Amat, X. Fradera, R. Cabó. Sobre els mapes de semblança quàntica molecular. *SCIENTIA gerundensis* **1996**, 22, 97-107.
  21. D. A. Winkler, G. Holan. Design of potential anti-HIV agents. 1. Mannosidase inhibitors. *J. Med. Chem.* **1989**, 32, 2084-2089.
  22. D. A. Winkler. Molecular modeling studies of "Flap Up" mannosyl cation mimics. *J. Med. Chem.* **1996**, 39, 4332-4334.
  23. S. Knapp, T. G. Murali Dhar. Synthesis of the mannosidase II inhibitor mannostatin A. *J. Org. Chem.* **1991**, 56, 4096-4097.
  24. S. B. King, B. Ganem. Synthetic studies on mannostatin A and its derivatives: a new family of glycoprotein processing inhibitors. *J. Am. Chem. Soc.* **1994**, 116, 562-570.

## 5. Anàlisis QSAR

---

Degut a la complexitat dels sistemes biològics, caracteritzar i interpretar els processos químics involucrats en les reaccions farmacològiques és molt difícil de tractar de manera rigorosa i realista. Si barregen no només aspectes químics relacionats amb el compost actiu, sinó també cal tenir en compte aspectes biològics tan complexes com la interacció lligand–receptor o el trasllat del fàrmac al lloc d'acció. En molts de casos la manca de coneixements i mitjans fa obviar molts factors que poden ser molt influents en l'activitat. Possiblement han estat aquests inconvenients els que han propiciat el desenvolupat de diferents aproximacions per establir relacions entre l'estructura i l'activitat de molècules amb activitat coneguda, amb l'objectiu de trobar equacions matemàtiques que relacionin descriptors moleculars amb l'activitat biològica, per posteriorment aplicar els models obtinguts a molècules amb activitat desconeguda i així fer-ne una predicció. El desenvolupament de nous procediments va molt lligat a la recerca de nous paràmetres o descriptors útils per caracteritzar l'activitat biològica de les molècules estudiades. A partir de l'àmplia bibliografia existent, es pot fer una classificació dels principals descriptors que són utilitzats per representar les estructures moleculars. L'intent d'organització dels descriptors moleculars és arbitrari, però força establert en la comunitat científica. La primera subdivisió faria referència als models QSAR que utilitzen paràmetres derivats de la química orgànica, i que se solen anomenar anàlisis QSAR clàssiques.<sup>1-5</sup> També reben el nom d'aproximació de Hansch, en referència al científic Corwin Hansch que en els anys seixanta va ser el propulsor d'un procediment general basat en descriptors físico-químics. Inicialment els paràmetres emprats es podien englobar en tres categories: hidrofòbics, electrònics i estèrics. Posteriorment s'ha inclòs altres descriptors de temàtica diversa, sobretot propietats experimentals, com la solubilitat o el punt de fusió. També s'han utilitzat models químics corresponents a forces secundàries com les induïdes per l'efecte dipol-dipol o per enllaços d'hidrogen, o les de transferència de càrrega. El segon grup el formen les propietats determinades teòricament, i que inclouria els descriptors topològics,<sup>6-10</sup> paràmetres deduïts de la química computacional,<sup>11-14</sup> o descriptors espectroscòpics.<sup>15-17</sup>

Els descriptors calculats a partir de programes de química computacional són molt diversos, però els més usuals corresponen a energies HOMO-LUMO, moments dipolars, càrregues atòmiques, duresa o polarizabilitats. Més recentment, a partir dels anys 80, han sorgit els mètodes anomenats *3D-QSAR*,<sup>18</sup> els quals composarien el tercer grup. Són tècniques que consideren les molècules en la seva estructura tridimensional, i normalment requereixen de processos de superposició molecular. Dins aquest grup s'hi inclourien els mètodes de semblança molecular,<sup>19,20</sup> i entre ells els estudis de semblança molecular quàntica que s'analitzaran en els propers capítols. Tanmateix, les dues aproximacions més comunament utilitzades són les tècniques CoMFA<sup>21</sup> i GRID,<sup>22</sup> que es fonamenten en càlculs de l'energia d'interacció avaluada en una xarxa de punts definida en l'espai que envolta les molècules estudiades. Usen potencials de Lennard-Jones i de Coulomb per representar els efectes estèrics i electrostàtics respectivament. A partir del mètode CoMFA s'han derivat altres aproximacions *3D-QSAR* que es poden trobar en la bibliografia.<sup>23-28</sup>

Independentment del mètode escollit per obtenir els descriptors moleculars, les anàlisis *QSAR* segueixen unes pautes comunes que es poden esquematitzar en tres etapes. En primer lloc s'han de tenir dades experimentals de l'activitat per a un conjunt conegut de molècules, que s'anomena conjunt d'entrenament o sèrie d'exploració. El segon procés és la determinació d'uns paràmetres empírics o teòrics que caracteritzin les molècules estudiades. I finalment, mitjançant tècniques estadístiques, se seleccionen els descriptors i propietats moleculars més representatives de l'activitat biològica analitzada. En definitiva, els models *QSAR* pretenen, mitjançant la generació d'equacions matemàtiques, descriure el comportament d'una activitat biològica amb la finalitat de fer prediccions sobre molècules no sintetitzades. És per això que un dels principals aspectes a tenir en compte dels models matemàtics generats ha de ser la seva capacitat predictiva.

## 5.1 Models de regressió multilinear

Com ja s'ha comentat, l'objectiu dels estudis *QSAR* és la generació de models matemàtics que connectin les dades experimentals, com són les activitats biològiques, amb un conjunt de paràmetres o descriptors determinats a partir de les estructures moleculars. La tècnica més utilitzada per obtenir els models matemàtics és l'anàlisi de regressió multilinear (*Multilinear Regression analysis, MLR*). Donat el vector que conté els paràmetres dependents,  $\mathbf{y}=(y_1, y_2, \dots, y_n)^T$ , i una matriu ( $n \times k$ ) de descriptors,  $\mathbf{X}=\{x_{ij}\}$ , l'equació de regressió multilinear en forma matricial és

$$\mathbf{y} = \mathbf{X} \mathbf{b} \quad , \quad (5.1)$$

i el vector amb els coeficients *MLR* es determina mitjançant l'expressió:

$$\mathbf{b} = [\mathbf{X}^T \mathbf{X}]^{-1} \mathbf{X}^T \mathbf{y} \quad . \quad (5.2)$$

Amb els valors calculats de  $\mathbf{b}$  es fa una estimació dels valors de la propietat experimental,  $\mathbf{y}'=(y'_1, y'_2, \dots, y'_n)^T$ , segons:

$$\mathbf{y}' = \mathbf{X} \mathbf{b} \quad (5.3)$$

que permeten definir l'error de la predicció o residuals:

$$\mathbf{y} - \mathbf{y}' = \mathbf{y} - \mathbf{X} \mathbf{b} = \mathbf{e} \quad (5.4)$$

La matriu que relaciona el vector de valors observats amb els valors estimats per la variable dependent,

$$\mathbf{y}' = \mathbf{H} \mathbf{y} \quad , \quad (5.5)$$

s'anomena matriu de predicció i es defineix com:

$$\mathbf{H} = \mathbf{X}[\mathbf{X}^T \mathbf{X}]^{-1} \mathbf{X}^T \quad . \quad (5.6)$$

Els elements de la matriu  $\mathbf{H}$  es poden determinar a partir de l'expressió

$$h_{ij} = \mathbf{x}_i^T [\mathbf{X}^T \mathbf{X}]^{-1} \mathbf{x}_j, \quad (5.7)$$

on  $\mathbf{x}_i^T = (x_{i1}, x_{i2}, \dots, x_{ik})$  és la filera  $i$  de la matriu  $\mathbf{X}$ .

Tot seguit es descriuen alguns dels paràmetres estadístics més usats en l'àmbit de les equacions *MLR*.

### 5.1.1 Regressió lineal

El cas més simple de *MLR* és trobar la regressió lineal entre una variable independent  $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$  i una variable dependent  $\mathbf{y} = (y_1, y_2, \dots, y_n)^T$ . L'equació

$$\mathbf{y} = a + b\mathbf{x} \quad (5.8)$$

s'anomena recta de regressió *X/Y*. Els coeficients  $\{a, b\}$  es determinen mitjançant un mètode de mínims quadrats, imposant que la suma d'errors quadrats

$$SS_{\text{Res}} = \sum_i^n (y_i - y'_i)^2 = \sum_i^n [y_i - (a + bx_i)]^2 = \sum_i^n e_i^2 \quad (5.9)$$

sigui mínima. És fàcil demostrar que els coeficients de regressió òptims són:<sup>29</sup>

$$b = \frac{n \sum_i^n x_i y_i - \left( \sum_i^n x_i \right) \left( \sum_i^n y_i \right)}{n \sum_i^n x_i^2 - \left( \sum_i^n x_i \right)^2} = \frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_i^n (x_i - \bar{x})^2} \quad (5.10)$$

i

$$a = \bar{y} - b\bar{x} \quad (5.11)$$

essent  $\bar{x} = \frac{1}{n} \sum_i^n x_i$  i  $\bar{y} = \frac{1}{n} \sum_i^n y_i$  les mitjanes aritmètiques.

### 5.1.2 Variança, desviació estàndard i covariança

La variança és la suma de les desviacions al quadrat d'un conjunt de variables i dividida pel nombre de graus de llibertat:

$$s_X^2 = \frac{\sum_i^n (x_i - \bar{x})^2}{n-1} ; s_Y^2 = \frac{\sum_i^n (y_i - \bar{y})^2}{n-1} \quad (5.12)$$

Mentre que la desviació estàndard és simplement l'arrel quadrada de la variança:

$$s_X = \sqrt{\frac{\sum_i^n (x_i - \bar{x})^2}{n-1}} ; s_Y = \sqrt{\frac{\sum_i^n (y_i - \bar{y})^2}{n-1}} \quad (5.13)$$

La covariança es defineix com la suma dels productes de les desviacions dels parells  $\{x_i, y_i\}$  respecte a les seves mitjanes i dividit pels graus de llibertat:

$$c_{XY} = \frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{n-1} \quad (5.14)$$

Emprant les definicions de variança i covariança, el coeficient  $b$  de l'equació (5.10) es pot rescriure com

$$b = \frac{c_{XY}}{s_X^2} \quad (5.15)$$

### 5.1.3 Coeficient de correlació i de determinació

El coeficient de correlació es pot definir a partir de la covariança  $c_{XY}$  i de les desviacions estàndards de les dues variables d'acord amb:

$$r = \frac{c_{XY}}{s_X s_Y} \quad (5.16)$$

Substituint les identitats en l'equació (5.16) s'obté l'expressió:

$$r = \frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{\left[ \sum_i^n (x_i - \bar{x})^2 \right]^{1/2} \left[ \sum_i^n (y_i - \bar{y})^2 \right]^{1/2}} = \frac{n \sum_i^n x_i y_i - \left( \sum_i^n x_i \right) \left( \sum_i^n y_i \right)}{\left[ n \sum_i^n x_i^2 - \left( \sum_i^n x_i \right)^2 \right]^{1/2} \left[ n \sum_i^n y_i^2 - \left( \sum_i^n y_i \right)^2 \right]^{1/2}} \quad (5.17)$$

El coeficient de correlació és una mesura del grau de relació present. Però també és usual utilitzar el coeficient de correlació al quadrat, que es pot determinar fàcilment a partir de l'equació (5.17), o bé a través dels vectors  $\mathbf{y}$  i  $\mathbf{y}\mathbf{c}$  com es mostrarà seguidament. Una vegada determinats els valors dels coeficients  $\{a,b\}$  de la recta de regressió  $X/Y$  es fa una estimació dels valors de la propietat experimental:

$$y'_i = a + b x_i \quad (5.18)$$

Substituint el valor de  $a$  per l'equació (5.11), agrupant els termes en  $X$  i en  $Y$ , elevant al quadrat i fent el sumatori per a totes les observacions, s'obté la suma d'errors quadrats de regressió:

$$SS_{\text{Reg}} = \sum_i^n (y'_i - \bar{y})^2 = b^2 \sum_i^n (x_i - \bar{x})^2 = b \sum_i^n (x_i - \bar{x})(y_i - \bar{y}). \quad (5.19)$$

D'una manera similar es pot desenvolupar la suma d'errors quadrats definida en l'equació (5.9) fins a arribar a la igualtat:<sup>29</sup>

$$\sum_i^n (y_i - y'_i)^2 = \sum_i^n (y_i - \bar{y})^2 - \frac{\left[ \sum_i^n (x_i - \bar{x})(y_i - \bar{y}) \right]^2}{\sum_i^n (x_i - \bar{x})^2} = (n-1) \left[ s_Y^2 - \frac{c_{XY}^2}{s_X^2} \right] \quad (5.20)$$

o també

$$\sum_i^n (y_i - \bar{y})^2 = \sum_i^n (y_i - y'_i)^2 + \sum_i^n (y'_i - \bar{y})^2, \quad (5.21)$$

que se sol simplificar amb la notació:

$$S_{yy} = SS_{\text{Res}} + SS_{\text{Reg}}. \quad (5.22)$$

Aleshores, el coeficient de correlació al quadrat es pot determinar a partir de l'equació (5.20) com:

$$r^2 = \frac{C_{XY}^2}{S_X^2 S_Y^2} = 1 - \frac{\sum_i^n (y_i - y'_i)^2}{\sum_i^n (y_i - \bar{y})^2} = \frac{\sum_i^n (y'_i - \bar{y})^2}{\sum_i^n (y_i - \bar{y})^2}, \quad (5.23)$$

i emprant la notació abreviada:

$$r^2 = 1 - \frac{SS_{\text{Res}}}{S_{yy}} = \frac{SS_{\text{Reg}}}{S_{yy}} \quad (5.24)$$

El paràmetre  $r^2$  és sovint anomenat coeficient de determinació i s'interpreta com la proporció de la variança que tenen en comú  $X$  i  $Y$ .

#### 5.1.4 Coeficient de múltiple determinació

En el cas general d'una anàlisi de regressió multilinear, corresponent a l'equació (5.1), la igualtat (5.22) també es compleix.<sup>29</sup> Llavors el coeficient de múltiple correlació al quadrat o coeficient de múltiple determinació es determina segons:

$$r^2 = \frac{SS_{\text{Reg}}}{S_{yy}} \quad (5.25)$$

En el cas d'un problema de *MLR*, la suma d'errors quadrats de la regressió lineal serà

$$\begin{aligned} SS_{\text{Reg}} &= b_1 \sum_i^n (x_{i1} - \bar{x}_1)(y_i - \bar{y}) + b_2 \sum_i^n (x_{i2} - \bar{x}_2)(y_i - \bar{y}) + \dots + b_k \sum_i^n (x_{ik} - \bar{x}_k)(y_i - \bar{y}) \\ &= \sum_j^k b_j \sum_i^n (x_{ij} - \bar{x}_j)(y_i - \bar{y}) \end{aligned} \quad (5.26)$$



### 5.1.5 Importància estadística dels models *MLR*

La importància estadística de dos models *MLR* amb el mateix nombre de punts  $n$  i mateix nombre de paràmetres  $k$  és fàcil d'avaluar, únicament s'han de comparar els valors dels coeficients de correlació. Però quan el nombre de punts o el nombre de paràmetres és diferent llavors és difícil decidir quin dels dos models és el millor estadísticament parlant. Una possible solució és analitzar quin model té menys probabilitat de donar uns resultats estadístics bons quan se substitueix el conjunt de dades originals per unes d'estocàstiques.<sup>30</sup> Donada una matriu  $(n \times k)$  de variables independents,  $\mathbf{X}=\{x_{ij}\}$ , i un vector amb els paràmetres dependents,  $\mathbf{y}=(y_1, y_2, \dots, y_n)^\top$ , primer es calcula el coeficient de correlació  $r$ . Després s'analitzaria la mateixa correlació emprant un conjunt de variables generades aleatòriament enlloc de les variables originals  $\mathbf{y}$  i  $\mathbf{X}$ . El coeficient de correlació  $R$  del model *MLR* generat amb les variables estocàstiques serà, amb gran probabilitat, inferior al valor inicial de  $r$ . Repetint el mateix experiment moltes vegades, hi ha una probabilitat no nul·la,  $P$ , que el coeficient de correlació per a un dels conjunts de variables generades aleatòriament sigui igual o superior a  $r$ . Aquesta probabilitat depèn del nombre de punts  $n$ , del valor del coeficient de correlació  $r$  i del nombre de paràmetres  $k$ . Quan més petit sigui el valor de  $P$ , més difícil és obtenir una correlació amb  $R > r$ . S'ha proposat un model geomètric molt simple que permet el càlcul analític de la probabilitat  $P$  donat el valors de  $n$ ,  $k$  i  $r$ .<sup>30</sup> Respon a la fórmula:

$$P = \frac{\int_0^{\arccos(r)} \cos^{k-1} \mathbf{q} \sin^{n-k-2} \mathbf{q} d\mathbf{q}}{\int_0^{\pi/2} \cos^{k-1} \mathbf{q} \sin^{n-k-2} \mathbf{q} d\mathbf{q}} \quad (5.27)$$

que es pot calcular numèricament.

## 5.2 Capacitat de predicció dels models *MLR*

Conjuntament amb la construcció de l'equació *MLR* i el càlcul dels paràmetres estadístics de l'ajust, és molt important determinar la capacitat predictiva del model. El procediment més usual per fer-ho és a través del coeficient de validació creuada (*Cross-validation, CV*).<sup>31-33</sup> Consisteix en eliminar cert nombre d'objectes del conjunt inicial, construir un nou model amb les restants observacions, i utilitzar el model reduït per fer la predicció de la variable dependent dels objectes exclosos inicialment. El procés es repeteix tantes vegades com calgui fins a obtenir un vector amb tots els valors predits de la propietat:  $\hat{\mathbf{y}} = (\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n)^T$ . La titlla sobre les variables indica que els valors s'han obtingut a partir d'una predicció, i serveix per diferenciar-los dels valors ajustats  $\mathbf{y}'$ . El més usual és extreure un sol objecte en cada càlcul, i repetir el procés pels  $n$  objectes del conjunt inicial. Aleshores el procés rep el nom de *leave-one-out (LOO)*. Normalment se sol presentar en una gràfica bidimensional els valors de la variable dependent obtinguts de la validació creuada,  $\hat{\mathbf{y}}$ , respecte als valors experimentals,  $\mathbf{y}$ .

### 5.2.1 Coeficient de la validació creuada

Per analogia amb l'expressió (5.23), una estimació del coeficient de correlació del procés de validació creuada és

$$q^2 = 1 - \frac{\sum_i^n (y_i - \hat{y}_i)^2}{\sum_i^n (y_i - \bar{y})^2} = 1 - \frac{PRESS}{S_{yy}}, \quad (5.28)$$

on els dos sumatoris s'identifiquen amb l'error de predicció de la suma de quadrats estadístic (*PRediction Error of the Sum of Squares, PRESS*)<sup>34</sup> i la suma d'errors quadrats segons la mitjana,  $S_{yy}$ . El coeficient de predicció definit en l'expressió (5.28) pot ser negatiu. Quan això succeeix significa que les prediccions del model són pitjors que si es pren la mitjana aritmètica del vector  $\mathbf{y}$  com el valor predit per a totes les dades. Per tant seria més convenient utilitzar la notació  $q^{(2)}$  en lloc de les més comunes  $Q^2$  o  $q^2$  que es troben en la bibliografia. Una altra possibilitat és calcular directament el coeficient de correlació entre els vectors  $\mathbf{y}$  i  $\hat{\mathbf{y}}$ , i que es pot representar per  $r_{cv}$ .<sup>35,36</sup> Així s'evitaria la problemàtica dels valors negatius en el coeficient  $q^2$ .

### 5.2.2 Càlcul del vector de prediccions a partir de la matriu de predicció

Tanmateix, el càlcul del coeficient de correlació derivat del procés lineal de validació creuada és costós. En el llibre [37] es demostra com el vector de les propietats predites  $\hat{\mathbf{y}} = (\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n)^\top$  per un estudi *LOO* es pot determinar a partir de la matriu de prediccions  $\mathbf{H}$ , i així s'evita el càlcul de tots els models *MLR* involucrats en la validació creuada. Tot seguit es plantegen els principals arguments matemàtics emprats en la seva deducció. Es defineix el vector columna  $\mathbf{x}_p = (x_{p1}, x_{p2}, \dots, x_{pk})^\top$  com el que agrupa els descriptors independents originals de l'objecte  $p$  que coincideix amb la filera  $p$  de la matriu  $\mathbf{X}$ . Si s'eliminen les dades de l'objecte  $p$ , això és, el valor  $y_p$  i el vector  $\mathbf{x}_p$  són igualats a zero, llavors es pot definir un nou vector de propietats  $\mathbf{y}_{(p)}$  i una nova matriu de descriptors  $\mathbf{X}_{(p)}$ . Seguint una notació similar a l'expressió (5.2), els coeficients del model lineal són

$$\mathbf{b}_{(p)} = [\mathbf{X}_{(p)}^\top \mathbf{X}_{(p)}]^{-1} \mathbf{X}_{(p)}^\top \mathbf{y}_{(p)}. \quad (5.29)$$

El vector  $\mathbf{b}_{(p)}$  permet el càlcul del valor de la validació creuada de la propietat per l'objecte  $p$ :

$$\hat{y}_p = \mathbf{x}_p^\top \mathbf{b}_{(p)} = \mathbf{x}_p^\top [\mathbf{X}_{(p)}^\top \mathbf{X}_{(p)}]^{-1} \mathbf{X}_{(p)}^\top \mathbf{y}_{(p)}. \quad (5.30)$$

És possible demostrar que la següent relació

$$[\mathbf{X}_{(p)}^\top \mathbf{X}_{(p)}]^{-1} = [\mathbf{X}^\top \mathbf{X}]^{-1} + (1 - h_{pp})^{-1} [\mathbf{X}^\top \mathbf{X}]^{-1} \mathbf{x}_p \mathbf{x}_p^\top [\mathbf{X}^\top \mathbf{X}]^{-1} \quad (5.31)$$

es compleix.<sup>37</sup> Per tant es pot escriure:

$$\hat{y}_p = \mathbf{x}_p^\top \left\{ [\mathbf{X}^\top \mathbf{X}]^{-1} + \frac{[\mathbf{X}^\top \mathbf{X}]^{-1} \mathbf{x}_p \mathbf{x}_p^\top [\mathbf{X}^\top \mathbf{X}]^{-1}}{1 - h_{pp}} \right\} \mathbf{X}_{(p)}^\top \mathbf{y}_{(p)} = \frac{\mathbf{x}_p^\top [\mathbf{X}^\top \mathbf{X}]^{-1} \mathbf{X}_{(p)}^\top \mathbf{y}_{(p)}}{1 - h_{pp}}. \quad (5.32)$$

A més,  $\mathbf{X}^\top \mathbf{y} = \mathbf{X}_{(p)}^\top \mathbf{y}_{(p)} + \mathbf{x}_p y_p$ , amb el que

$$\hat{y}_p = \frac{\mathbf{x}_p^\top [\mathbf{X}^\top \mathbf{X}]^{-1} [\mathbf{X}^\top \mathbf{y} - \mathbf{x}_p y_p]}{1 - h_{pp}} = \frac{\mathbf{x}_p^\top [\mathbf{X}^\top \mathbf{X}]^{-1} \mathbf{X}^\top \mathbf{y} - \mathbf{x}_p^\top [\mathbf{X}^\top \mathbf{X}]^{-1} \mathbf{x}_p y_p}{1 - h_{pp}} \quad (5.33)$$

i a partir de (5.2) i (5.7) s'obté fàcilment

$$\hat{y}_p = \frac{1}{1 - h_{pp}} \sum_{i \neq p}^n h_{pi} y_i = \frac{y'_p - h_{pp} y_p}{1 - h_{pp}}. \quad (5.34)$$

Una vegada determinat el vector  $\hat{\mathbf{y}}$ , es pot calcular el coeficient  $q^2$  mitjançant l'equació (5.28), o bé avaluar el coeficient  $r_{cv}$ . Recentment s'ha demostrat que també és possible determinar el vector  $\hat{\mathbf{y}}$  d'un procés general *leave-many-out* a partir de la matriu de prediccions  $\mathbf{H}$ .<sup>35</sup>

### 5.3 Mètodes d'anàlisi multivariant

Normalment les matrius dels descriptors moleculars no poden ser utilitzades directament en qualitat de variables independents en anàlisis *MLR* degut a la seva homogeneïtat o al gran nombre de descriptors involucrats, que normalment excedeixen el nombre de compostos estudiats. Per això, com a pas previ d'una anàlisi *MLR*, sol ser necessari fer un tractament de reducció de variables per obtenir un conjunt de variables significatives i poc correlades entre elles. Seguidament es descriuran les tècniques estadístiques d'anàlisi multivariant que s'utilitzen més assíduament en els estudis *QSAR* basats en *MQSM*. Corresponen a l'anàlisi de components principals (*Principal Component Analysis, PCA*),<sup>38</sup> la tècnica de mínims quadrats parcials (*Partial Least Squares, PLS*)<sup>39,40</sup> i l'anàlisi de coordenades principals.<sup>41,42</sup>

Bàsicament tots els mètodes d'anàlisi multivariant intenten explicar un conjunt extens de variables mitjançant un número reduït de variables hipotètiques anomenades factors, components o coordenades. Donat un conjunt de variables  $\mathbf{X}(n \times m)$ , es tracta de trobar unes noves variables  $\mathbf{T}(n \times k)$ , on  $k \leq m$ , i determinar la seva contribució a les variables originals per mitjà de la matriu de coeficients  $\mathbf{P}(m \times k)$ . La descomposició de la matriu de variables respondria a l'equació matricial:

$$\mathbf{X} = \mathbf{T} \mathbf{P}^T + \mathbf{E} \quad (5.35)$$

on  $\mathbf{E}$  és la matriu d'errors residuals. Determinar la matriu de coeficients  $\mathbf{P}$  és el principal problema de les anàlisis multivariants. Una vegada descomposta la matriu de variables original, es construeixen les equacions *MLR* entre les propietats experimentals i les variables transformades  $\mathbf{T}$ .

### 5.3.1 Anàlisi de components principals

L'objectiu del mètode *PCA* és extreure el màxim de la variança de les variables inicials amb cada component de la matriu transformada  $\mathbf{T}$ . Pel raonament que segueix es parteix de la matriu de les correlacions entre les variables analitzades, definida com:  $\mathbf{R} = \mathbf{X}^T \mathbf{X}$ . La matriu dels pesos dels factors, o també anomenada matriu factorial, ha de verificar la igualtat

$$\mathbf{R} = \mathbf{P} \mathbf{P}^T. \quad (5.36)$$

Però la relació (5.36) no és suficient per determinar  $\mathbf{P}$ , perquè hi ha infinites matrius que la verifiquen. L'anàlisi *PCA* escull el primer component principal de manera que expliqui la major part de la variança de les variables. Després se'l resta de les altres variables, i sobre la variabilitat restant s'escull el segon component principal amb el mateix criteri, i així successivament. El procés resultant equival al càlcul dels valors i vectors propis de  $\mathbf{R}$ , perquè qualsevol matriu simètrica, com ho és  $\mathbf{R}$ , es pot expressar com el producte:

$$\mathbf{R} = \mathbf{V} \mathbf{L} \mathbf{V}^T \quad (5.37)$$

on  $\mathbf{V}$  és una matriu ortogonal que conté els vectors propis normalitzats i  $\mathbf{L}$  és una matriu diagonal amb els valors propis. A més si  $\mathbf{R}$  és semidefinida positiva, és a dir, que els seus valors propis, tots reals, són no negatius, llavors

$$\mathbf{L}^{1/2} = \text{diag}(\sqrt{\lambda_1}, \dots, \sqrt{\lambda_m}) \quad (5.38)$$

i prenent  $\mathbf{P} = \mathbf{V} \mathbf{L}^{1/2}$  es compleix l'equació (5.36). Una vegada determinada la matriu  $\mathbf{P}$ , la matriu dels factors o components es pot calcular a partir de la relació (5.35) com  $\mathbf{T} = \mathbf{X} \mathbf{P}^{-T}$ . Normalment amb els primers components ( $k < m$ ) és possible reproduir la major part de la variança continguda en la matriu de variables inicial.

### 5.3.2 Tècnica de mínims quadrats parcials

La principal diferència entre la tècnica *PLS* i l'anàlisi *PCA* és la inclusió del vector de les variables dependents,  $\mathbf{y}$ , en el procés de reducció de variables. A l'igual que en el mètode *PCA*, *PLS* busca una matriu de components  $\mathbf{T}$ , que reculli la màxima variança continguda en les variables originals  $\mathbf{X}$ , però que alhora estiguin correlades amb el vector  $\mathbf{y}$ . Tot seguit es detalla l'algorisme *PLS* descrit en l'article [39].

Abans d'iniciar el procés és recomanable estandarditzar la matriu de dades  $\mathbf{X}$  de manera que cadascuna de les seves columnes tingui mitjana zero i desviació estàndard u

$$\mathbf{x}_i = s_{x_i}^{-1}(\mathbf{x}_i - \bar{x}_i \mathbf{1}). \quad (5.39)$$

En l'equació (5.39)  $\mathbf{x}_i = (x_{1i}, x_{2i}, \dots, x_{ni})^T$  és la columna  $i$  de la matriu  $\mathbf{X}$ ,  $s_{x_i}^{-1}$  és la seva desviació estàndard i  $\bar{x}_i$  la seva mitjana aritmètica. Aquesta estandardització no provoca canvis en el coeficient de correlació final.

La transformació *PLS* se sol realitzar mitjançant un procés iteratiu. L'algorisme [39] comença per estimar el vector de pesos  $\mathbf{w}_k$  del primer component ( $k=1$ ) com:

$$\mathbf{w}_k = \frac{\mathbf{y}^T \mathbf{X}}{\mathbf{y}^T \mathbf{y}}. \quad (5.40)$$

Llavors normalitza el vector  $\mathbf{w}_k$  d'acord amb

$$\mathbf{w}_k^T = \frac{\mathbf{w}_k^T}{\|\mathbf{w}_k^T\|}, \quad (5.41)$$

i determina el vector de *scores*  $\mathbf{t}_k$  segons

$$\mathbf{t}_k = \mathbf{X} \mathbf{w}_k. \quad (5.42)$$

El vector de *loadings*  $\mathbf{p}_k$  s'obté per regressió de  $\mathbf{X}$  amb  $\mathbf{t}_k$ :

$$\mathbf{p}_k^T = \frac{\mathbf{t}_k^T \mathbf{X}}{\mathbf{t}_k^T \mathbf{t}_k} \quad (5.43)$$

el qual també es normalitza

$$\mathbf{p}_k^\top = \frac{\mathbf{p}_k^\top}{\|\mathbf{p}_k^\top\|}. \quad (5.44)$$

Per fer les estimacions de  $\mathbf{y}$  a partir de  $\mathbf{t}_k$ , és necessari determinar el coeficient de regressió:

$$b_k = \frac{\mathbf{y}^\top \mathbf{t}_k}{\mathbf{t}_k^\top \mathbf{t}_k}. \quad (5.45)$$

Al final de la primera iteració de l'algorisme *PLS* s'extreu de la matriu inicial de dades  $\mathbf{X}$  la contribució  $\mathbf{t}\mathbf{p}^\top$ , definint-se la matriu d'errors residuals  $\mathbf{E}$ :

$$\mathbf{E}_{k+1} = \mathbf{X} - \mathbf{t}_k \mathbf{p}_k^\top, \quad (5.46)$$

i del vector de propietats  $\mathbf{y}$  la contribució  $b\mathbf{t}$ :

$$\mathbf{f}_{k+1} = \mathbf{y} - b_k \mathbf{t}_k. \quad (5.47)$$

Seguidament es repeteix tot el procés per al segon component ( $k=2$ ), substituint la matriu  $\mathbf{X}$  de les equacions (5.40), (5.42), (5.43) i (5.46) per  $\mathbf{E}_k$ , i el vector  $\mathbf{y}$  de (5.40), (5.45) i (5.47) per  $\mathbf{f}_k$ . Aquest bucle es va repetint fins que el mòdul del vector  $\mathbf{f}_k$  és més petit que una certa tolerància, o bé que la diferència entre els mòduls dels vectors  $\mathbf{f}_k$  i  $\mathbf{f}_{k-1}$  de dues iteracions consecutives és inferior a un altre líndar preestablert. Si el nombre  $k$  de components extrets iguala el nombre de descriptors continguts en  $\mathbf{X}$ , llavors la solució *PLS* és idèntica a la solució donada per *MLR*.

Els elements de la matriu de factors  $\mathbf{T}$  es defineixen en funció dels elements de les matrius  $\mathbf{W}$  i  $\mathbf{E}$  segons:

$$t_{ik} = \sum_l w_{lk} e_{il,k-1} \quad (5.48)$$

Però també es poden expressar com una combinació lineal de les dades originals,

$$t_{ik} = \sum_l w_{lk}^* x_{il}, \quad (5.49)$$

i uns nous pesos  $\mathbf{W}^*$  que es determinen d'acord amb:

$$\mathbf{W}^* = \mathbf{W}(\mathbf{P}^T \mathbf{W}) \quad (5.50)$$

Els *scores* s'han calculat de manera que tinguin la propietat de ser uns bons estimadors de la variable  $\mathbf{y}$ . És per això que es pot definir una equació *MLR* entre la matriu  $\mathbf{T}$  i el vector de propietats  $\mathbf{y}$ :

$$\mathbf{y} = \mathbf{T}\mathbf{b} + \mathbf{f} \quad (5.51)$$

Degut a la relació entre la matriu  $\mathbf{T}$  i la de les dades inicials  $\mathbf{X}$  a partir de la matriu de pesos  $\mathbf{W}^*$ , es pot rescriure l'equació (5.51) emprant la matriu  $\mathbf{X}$ :

$$\mathbf{y} = \mathbf{X}\mathbf{b}_{PLS} + \mathbf{f}, \quad (5.52)$$

on els coeficients de regressió *PLS*,  $\mathbf{b}_{PLS}$ , es calculen conforme a:

$$\mathbf{b}_{PLS} = \mathbf{W}^* \mathbf{b} \quad (5.53)$$

Els coeficients  $\mathbf{b}_{PLS}$  són útils per interpretar el model resultant a més de facilitar les prediccions de l'activitat de conjunts de dades externes al model.

### 5.3.3 Anàlisi de coordenades principals

Un mètode apropiat per a les matrius de semblança és l'anàlisi de coordenades principals, també anomenat *classical scaling*.<sup>41,42</sup> Igual que en els anteriors algorismes, consisteix en transformar la matriu de variables en una matriu simplificada amb els principals factors o components. Considera els objectes com punts en un espai euclidià multidimensional i troba les coordenades per aquests punts de manera que les distàncies entre punts ajusti el màxim possible les semblances originals. L'objectiu és obtenir una representació geomètrica de les variables en relació a una distància raonablement compatible amb la semblança.



Per desenvolupar el formulisme de *classical scaling* se sol plantejar el problema a l'invers, això és, suposant que la matriu de coordenades  $\mathbf{T}$ , de dimensions  $(n \times m)$ , és coneguda per a un conjunt de  $n$  punts en un espai euclidià. La manera de construir la matriu de distàncies al quadrat és simple partint d'aquesta matriu:

$$\mathbf{D}^{(2)} = \mathbf{c}\mathbf{1}^\top + \mathbf{1}\mathbf{c}^\top - 2\mathbf{T}\mathbf{T}^\top = \mathbf{c}\mathbf{1}^\top + \mathbf{1}\mathbf{c}^\top - 2\mathbf{B}. \quad (5.54)$$

essent  $\mathbf{B} = \mathbf{T}\mathbf{T}^\top$ ,  $\mathbf{1}$  un vector d'uns de dimensió  $n$  i  $\mathbf{c}$  un vector amb components igual als elements de la diagonal de  $\mathbf{B}$ .

Multiplicant ambdós costats de l'equació (5.54) pel factor  $-\frac{1}{2}$  i per la matriu  $\mathbf{J} = \mathbf{I} - n^{-1}\mathbf{1}\mathbf{1}^\top$ , que serveix per centrar les dades, s'obté

$$\begin{aligned} -\frac{1}{2}\mathbf{J}\mathbf{D}^{(2)}\mathbf{J} &= -\frac{1}{2}\mathbf{J}(\mathbf{c}\mathbf{1}^\top + \mathbf{1}\mathbf{c}^\top - 2\mathbf{B})\mathbf{J} \\ &= -\frac{1}{2}\mathbf{J}\mathbf{c}\mathbf{1}^\top\mathbf{J} - \frac{1}{2}\mathbf{J}\mathbf{1}\mathbf{c}^\top\mathbf{J} + \mathbf{J}\mathbf{B}\mathbf{J} \\ &= -\frac{1}{2}\mathbf{J}\mathbf{c}\mathbf{0}^\top - \frac{1}{2}\mathbf{0}\mathbf{c}^\top\mathbf{J} + \mathbf{J}\mathbf{B}\mathbf{J} = \mathbf{B} \end{aligned} \quad (5.55)$$

Els dos primers termes s'anul·len perquè centrar un vector d'uns dona un vector de zeros. I el tercer terme, centrar  $\mathbf{B}$  no té cap influència, perquè  $\mathbf{B}$  ha estat centrat.

La matriu  $\mathbf{B}$  es pot calcular definint la matriu de dades inicial  $\mathbf{X}$  com la matriu de distància  $\mathbf{D}$ , i convenientment transformada a partir de l'equació (5.55). Llavors determinar les coordenades  $\mathbf{T}$  a partir de la matriu  $\mathbf{B}$  es pot fer mitjançant la descomposició espectral

$$\mathbf{B} = \mathbf{T}\mathbf{T}^\top = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top. \quad (5.56)$$

La matriu de coordenades  $\mathbf{T}$  és simplement:  $\mathbf{T} = \mathbf{V}\mathbf{L}^{\frac{1}{2}}$ .

## 5.4 Selecció de les variables per obtenir el millor model *MLR*

En qualsevol estudi *QSAR*, l'últim estadi del procés és el càlcul d'un model *MLR* que relacioni la matriu de variables independents  $\mathbf{X}$  amb el vector de propietats  $\mathbf{y}$ . La matriu de dades  $\mathbf{X}$  pot contenir els descriptors moleculars originals o bé transformacions dels mateixos com les que s'han descrit en l'apartat anterior. Però independentment de la procedència dels elements de la matriu  $\mathbf{X}$ , generalment abans de deduir el model *MLR* s'ha de fer una selecció de les variables. Essent  $(n \times m)$  la dimensió de  $\mathbf{X}$ , interessa obtenir el millor model *MLR* amb nombre de paràmetres  $k < m$ , i quan més reduït millor. Hi ha diversos criteris per establir quin és el "millor" model, i diferents metodologies per triar les variables.

Una selecció òbvia de les variables quan s'examina una matriu de dades que ha estat transformada amb alguna de les tècniques d'anàlisi multivariant descrites en l'apartat 5.3, és escollir els primers  $k$  factors, components o coordenades. Per exemple, en els estudis *PCA* és usual elegir els primers  $k$  components que donen la màxima variança, amb la finalitat de definir un subespai de dimensió  $k$  que millor ajusti les dades originals. Sabent que entre tots els factors es reproduïx el 100 % de la variança inicial, es pot calcular el percentatge de variança que recull cada component, que és proporcional al seu valor propi. En les anàlisis de coordenades principals, igualment que en *PCA*, una bona selecció consisteix en agafar els primers  $k$  vectors propis disposats en ordre decreixent dels corresponents valors propis. Aquesta selecció dóna el millor ajust en l'espai de dimensió  $k$ , on la suma de les distàncies entre punts és màxima.

Una alternativa és el mètode de les variables més predictives (*Most Predictive Variables Method, MPVM*).<sup>43,44</sup> Aquesta tècnica selecciona les primeres  $k$  columnes de la matriu de configuració  $\mathbf{X}$  disposades en ordre decreixent de la correlació absoluta respecte a les dades estudiades:

$$\chi^2(\mathbf{y}, \mathbf{x}_i) = \frac{(\mathbf{y}^\top \mathbf{x}_i)^2}{\sum_j (y_j - \bar{y})^2 \mathbf{I}_j}, \quad (5.57)$$

on  $\lambda_j$  és el valor propi corresponent a l'eix  $j$ . El mètode determina els millors descriptors d'acord amb la seva correlació individual amb una variable externa,  $\mathbf{y}$ .

Però hi ha altres mètodes de selecció de variables. Una possibilitat és la recerca sistemàtica dels millors descriptors d'entre totes les possibles combinacions de variables del conjunt de dades. És aplicable a qualsevol matriu  $\mathbf{X}$ , independentment de la seva procedència. Totes les possibles combinacions de  $m$  descriptors amb  $k$  paràmetres es poden generar mitjançant un algorisme de sumes aniuades (*Nested Summation Symbol algorithm, NSS*).<sup>45,46</sup> Aquest procediment matemàtic permet definir bucles de profunditat variable en temps d'execució sense haver de canviar el codi font d'un programa. De tots els models *MLR* generats amb nombre de paràmetres  $k$  s'elegeix el que dona millors resultats estadístics, que generalment coincideix amb el valor del coeficient de predicció de la validació creuada superior.

Una altra qüestió és l'elecció del nombre òptim de paràmetres de la regressió multilíneal. Aquesta determinació no es pot fer a partir del coeficient de correlació  $r$  perquè sempre augmenta amb l'addició de nous paràmetres en les equacions *MLR*. En canvi el valor dels coeficients de predicció  $q^2$  o  $r_{cv}$ , no mostren un augment progressiu, sinó que solen presentar un màxim a partir del qual el valor del coeficient disminueix. Llavors s'escull el nombre de paràmetres  $k$  corresponent al màxim del coeficient de predicció. Una altra possibilitat és avaluar la probabilitat d'obtenir correlacions fortuïtes mitjançant el coeficient  $P$  definit en l'equació (5.27). Aquest coeficient també indica si el model s'ha sobreparametrizat per la inclusió de masses descriptors.

## 5.5 Validació estadística dels models *QSAR*

Qualsevol model *QSAR* necessita ser validat abans de la seva utilització en la interpretació i predicció de l'activitat biològica de compostos no sintetitzats. És necessari comprovar que el model matemàtic obtingut per a la sèrie de molècules estudiades no està sobredimensionat ni és donat a correlacions fortuïtes.<sup>47</sup> Un augment significatiu del nombre de variables incloses en el model final en comparació amb el nombre de compostos considerats fa augmentar el risc de correlacions casuals. Algunes de les metodologies més emprades per determinar la robustesa i la fiabilitat d'un model *QSAR* es descriuen tot seguit.

### 5.5.1 Validació creuada

La validació creuada és el mètode més usual per avaluar la capacitat de predicció dels models *QSAR*. Com s'ha comentat en l'apartat 5.2, en la validació creuada es calcula un vector  $\hat{\mathbf{y}} = (\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n)^T$  amb els valors predits de la propietat analitzada, normalment a partir d'un procés *LOO*. Una vegada determinat el vector  $\hat{\mathbf{y}}$  es pot calcular el coeficient  $q^2$  definit en l'equació (5.28), o bé avaluar el coeficient de correlació  $r_{cv}$  de la recta de regressió  $\hat{\mathbf{y}}/\mathbf{y}$ . Quan s'analitza un model *QSAR* amb transformacions de les matrius de dades prèvies al càlcul *MLR* s'hauria de fer la predicció de cadascuna de les propietats  $\hat{y}_i$  mitjançant models reduïts generats en les mateixes condicions que el model global.

### 5.5.2 Test d'aleatorietat sobre el vector de les activitats

Una de les validacions més usuals dels models *QSAR* és un test d'aleatorietat sobre el vector de les activitats biològiques  $\mathbf{y}$ .<sup>48</sup> Essent  $n$  el nombre de molècules amb activitat coneguda, es genera una seqüència aleatòria de la sèrie de nombres enters compresos entre 1 i  $n$ . La sèrie aleatòria de nombres enters s'utilitza per reordenar els components del vector  $\mathbf{y}$ , deixant els descriptors moleculars intactes, i es construeix un model *QSAR* en les mateixes condicions que el model original. El procés es repeteix forces vegades, guardant els valors de  $r$  i  $r_{cv}$  per a cada nou vector  $\mathbf{y}$ . Com a resultat, es representa en una gràfica els valors  $r$  en front  $r_{cv}$  del model original més els de tots els models generats eventualment. Un model *QSAR* consistent és aquell que únicament dona uns paràmetres estadístics satisfactoris per l'ordre correcte del vector  $\mathbf{y}$ .

### 5.5.3 *Dividir les dades en una sèrie d'exploració més un conjunt de test*

La validació creuada proporciona una aproximació raonable de l'habilitat del model *QSAR* ha predir l'activitat de nous compostos. Però té l'inconvenient que totes les molècules analitzades pertanyen al mateix conjunt de dades. Tanmateix, si es disposa d'una sèrie suficientment gran de molècules amb l'activitat coneguda, es pot dividir en dos grups: un s'utilitza per construir el model d'ajust i l'altre per validar-lo. Així es fan prediccions sobre membres de la mateixa família però que no han format part de la sèrie d'exploració, i dels quals es coneix la seva activitat. Això dona una idea de quina fiabilitat tindrien prediccions sobre compostos no sintetitzats. Amb les activitats predites dels compostos del conjunt de test es calcula la desviació estàndard dels errors de predicció (*Standard Deviation of Errors of Prediction, SDEP*), que es defineix com:

$$SDEP = \sqrt{n^{-1} \sum_i^n (\hat{y}_i - y_i)^2} . \quad (5.58)$$

El coeficient *SDEP* s'utilitza per estimar la qualitat de les prediccions del model.

## Referències

1. C. Hansch, T. Fujita.  $\rho$ - $\sigma$ - $\pi$  analysis. A method for the correlation of biological activity and chemical structure. *J. Am. Chem. Soc.* **1964**, *86*, 1616–1626.
2. C. Hansch, A. Leo. Substituent constants for correlation analysis in chemistry and biology. Wiley, New York, 1979.
3. R. I. Zalewski, T. M. Krygowski, J. Shorter (Eds.). Similarity models in organic chemistry, biochemistry, and related fields. Elsevier, Amsterdam, 1991.
4. H. Kubinyi. QSAR: Hansch analysis and related approaches. VCH, Weinheim, 1993.
5. C. Hansch, A. Leo. Exploring QSAR. Fundamentals and applications in chemistry and biology. ACS, Washington, DC, 1995.
6. H. Wiener. Structural determination of paraffin boiling points. *J. Chem. Phys.* **1947**, *69*, 17–20.
7. M. Randic. On characterization of molecular branching. *J. Am. Chem. Soc.* **1975**, *97*, 6609–6615.
8. L. B. Kier, L. H. Hall. Molecular connectivity in chemistry and drug research. Academic, New York, 1976.
9. L. B. Kier, L. H. Hall. Molecular connectivity in structure-activity analysis. Research Studies Press, Letchworth, 1986.
10. A. T. Balaban. Chemical graphs: looking back and glimpsing ahead. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 339–350.
11. M. R. Saunders, D. J. Livingstone. Electronic structure calculations in quantitative structure-property relationships. Publicat en el llibre: Advances in quantitative structure-property relationships. M. Charton (Ed.). JAI Press, London, volum 1, pàgines 53–79, 1996.
12. R. M. Hyde, D. J. Livingstone. Perspectives in QSAR: computer chemistry and pattern recognition. *J. Comput.-Aided Mol. Des.* **1988**, *2*, 145–155.
13. Oxford Molecular. The Medawar center, Oxford Science park, Sandford-on-Thames, Oxford OX4 4GA, U.K.
14. Molecular Simulations, Inc., 9685 Scranton Rd., San Diego, CA 92121-3752.
15. A. M. Ferguson, T. Heritage, P. Jonathon, S. E. Pack, L. Phillips, J. Rogan, P. J. Snaith. EVA: a new theoretically based molecular descriptor for use in QSAR/QSPR. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 143–152.
16. D. B. Turner, P. Willett, A. M. Ferguson, T. Heritage. Evaluation of a novel infrared range vibration-based descriptor (EVA) for QSAR studies. 1. General application. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 409–422.
17. C. M. R. Ginn, D. B. Turner, P. Willett, A. M. Ferguson, T. Heritage. Similarity searching in files of three-dimensional chemical structures: evaluation of the EVA descriptor and combination of rankings using data fusion. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 23–37.
18. H. Kubinyi (ed.). 3D QSAR in drug design: theory methods and applications. ESCOM Science Publishers B.V., Leiden, The Netherlands, 1993.

19. P. M. Dean (ed.). *Molecular similarity in drug design*. Blackie Academic & Professional, London, 1995.
20. K. Sen (Ed.). *Molecular similarity. Topics in Current Chemistry*. Springer Verlag, Berlin, volums 173 i 174, 1995.
21. R. D. Cramer III, D. E. Patterson, J. D. Bunce. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959–5967.
22. P. J. A. Goodford. Computational procedure for determining energetically favorable binding sites on biologically important molecules. *J. Med. Chem.* **1985**, *28*, 849–857.
23. G. Klebe, U. Abraham, T. Mietzner. 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.
24. A. N. Jain, K. Kolie, D. Chapman. Compass: predicting biological activities from molecular surface properties. Performance comparisons on a steroid benchmark. *J. Med. Chem.* **1994**, *37*, 2315–2327.
25. B. D. Silverman, D. E. Platt. Comparative molecular moment analysis (CoMMA): 3D-QSAR without molecular superposition. *J. Med. Chem.* **1996**, *39*, 2129–2140.
26. G. E. Kellogg, L. B. Kier, P. Gaillard, L. H. Hall. E-state fields: Applications to 3D QSAR. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 513–520.
27. G. Bravi, E. Gancia, P. Mascagni, M. Pegna, R. Todeschini, A. Zaliani. 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.* **1997**, *11*, 79–92.
28. D. D. Robinson, P. J. Winn, P. D. Lyne, W. G. Richards. Self-organizing molecular field analysis: A tool for structure-activity studies. *J. Med. Chem.* **1999**, *42*, 573–583.
29. A. L. Edwards. *An introduction to linear regression and correlation*. W. H. Freeman and Company, New York, 1984.
30. J. Pecka, R. Ponec. Simple analytical method for evaluation of statistical importance of correlations in QSAR studies. *J. Math. Chem.* **2000**, *23*, 13–22.
31. D. M. Allen. The relationship between variable selection and data augmentation and a method for prediction. *Technometrics* **1974**, *16*, 125–127.
32. S. Wold. Cross-validatory estimation of the number of components in factor and principal component models. *Technometrics* **1978**, *20*, 125–127.
33. S. Wold. Validation of QSARs. *Quant. Struct.-Act. Relat.* **1991**, *10*, 191–193.
34. D. M. Allen. The prediction sum of squares as a criterion for selecting variables. Technical report 23, Department of Statistics, University of Kentucky, 1971.
35. E. Besalú. Fast computation of cross-validated properties in full linear leave-many-out procedures. *J. Math. Chem.* **2001**, *29*, 191–204.
36. Ll. Amat, E. Besalú, R. Carbó-Dorca, R. Ponec. Identification of active molecular sites using quantum-self-similarity measures. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 978–991.
37. D. C. Montgomery, E. A. Peck. *Introduction to linear regression analysis*. Wiley, New York, 1992.
38. C. M. Cuadras. *Métodos de análisis multivariante*. EUB, Barcelona, 1996.

39. P. Geladi, B. R. Kowalski. Partial least-squares regression: a tutorial. *Analytica Chimica Acta* **1986**, *185*, 1–17.
40. S. Wold, E. Johansson, M. Cocchi. PLS-partial least-squares projections to latent structures. Publicat en el llibre: 3D QSAR in drug design: theory, methods and applications. H. Kubinyi (ed.). ESCOM Science Publishers B. V., Leiden, The Netherlands, pàgines 523–550, 1993.
41. K. V. Mardia, J. T. Kent, J. M. Bibby. *Multivariate Analysis*. Academic Press, London, 1979.
42. T. F. Cox, M. A. A. Cox. *Multidimensional Scaling*. Chapman and Hall, London, 1994.
43. C. M. Cuadras, C. Arenas. A distance based regression model for prediction with mixed data. *Commun. Statist. Theor. Methods* **1990**, *19*, 2261–2279.
44. C. M. Cuadras, C. Arenas, J. Fortiana. Some computational aspects of a distance-based model for prediction. *Commun. Statist. Simul.* **1996**, *25*, 593–609.
45. R. Carbó, E. Besalú. Definition, mathematical examples and quantum chemical applications of nested summation symbols and logical Kronecker deltas. *Comput. Chem.* **1994**, *18*, 117–126.
46. R. Carbó, E. Besalú. Definition and quantum chemical applications of nested summations symbols and logical functions: pedagogical artificial intelligence devices for formulae writing sequential programming and automatic parallel implementation. *J. Math. Chem.* **1995**, *18*, 37–72.
47. J. G. Topliss, R. J. Costello. Change correlations in structure-activity studies using multiple regression analysis. *J. Med. Chem.* **1972**, *15*, 1066–1068.
48. S. Wold, L. Erikson. Statistical validation of QSAR results. Publicat en el llibre: Chemometric methods in molecular design. H. Van de Waterbeemd (Ed.). VCH, New York, pàgines 309–318, 1995.





## 6. Matrius de semblança en anàlisis QSAR

---

Les tècniques *3D-QSAR*,<sup>1,2</sup> basades en la descripció tridimensional de les molècules, han anat guanyant pes al llarg dels anys davant les aproximacions tradicionals o clàssiques. I dins els mètodes *3D-QSAR*, les tècniques fonamentades en la semblança molecular han proliferat molt en els darrers anys.<sup>3-8</sup> La semblança molecular és una disciplina que considera les molècules en les tres dimensions i són importants, entre altres, l'anàlisi conformacional, les posicions atòmiques o la disposició espacial de les molècules comparades. Es fonamenten en el principi que molècules amb estructura semblant se suposa que també han de tenir propietats biològiques semblants i valors d'activitat comparables. El primer exemple de l'aplicació d'un índex de semblança molecular en *QSAR* va ser proposat per Hopfinger l'any 1980.<sup>9</sup> Va estudiar l'activitat d'un conjunt de triazines de Baker mitjançant mesures de volum definides com la suma dels solapaments entre tots els parells d'àtoms de les molècules comparades. Hopfinger va avaluar mesures puntuals dels compostos sobreposats en el suposat mode com s'enllacen amb el receptor. Paral·lelament a aquest treball, Carbó i col·laboradors van publicar el primer article de semblança entre densitats electròniques moleculars, també restringit a càlculs de semblança singulars, en un sol punt.<sup>10</sup> No va ser fins als treballs de Namasivayan i Dean,<sup>11</sup> i Hodgkin i Richards,<sup>12</sup> que els índexs de semblança es van considerar com unes funcions que es poden optimitzar, i es van començar a aplicar mètodes de superposició molecular.

El tractament estadístic de les matrius de semblança resultants de les mesures de semblança entre tots els parells possibles de molècules que formen el conjunt estudiat no es va produir fins a començaments dels anys noranta. Rum i Herndon van fer servir una matriu d'índexs de semblança ( $n \times n$ ), on  $n$  és el nombre de compostos analitzats, derivats de descriptors topològics, en una anàlisi de regressió que correlava algunes columnes d'aquesta matriu amb les activitats biològiques.<sup>13</sup> Good i col·laboradors, en una investigació més sistemàtica, van sobreposar molècules i van comparar la seva semblança mitjançant potencials electrostàtics i descriptors estèrics.<sup>14-16</sup> Van descriure

un protocol de l'aplicació de les matrius de semblança ( $n \times n$ ) en *QSAR* molt similar al que s'utilitza actualment en el nostre laboratori. Per primera vegada, el tractament de les matrius de semblança va incloure la reducció de dimensió i una validació estadística del procés.

En els darrers anys s'han publicat nombrosos treballs, tant teòrics com d'aplicacions pràctiques, relacionats amb la semblança molecular. Com a exemple, citar alguns dels llibres especialitzats en la matèria, [3-8], així com les referències que s'inclouen en els mateixos. Una de les principals aplicacions de la semblança molecular, encara que no la única, ha estat la definició de nous descriptors moleculars per ser utilitzats en estudis *QSAR/QSPR*. És precisament aquest àmbit de recerca el que es tractarà en el present capítol, i en particular una branca de la semblança molecular basada en descriptors mecanicoquàntics. En els estudis de semblança molecular orientats a dilucidar relacions estructura-activitat es distingeixen dos procediments, contigus en el temps, i que són característics de cada metodologia. El primer és l'obtenció de la matriu de semblança, que involucra la descripció de les molècules i la utilització d'un mètode de sobreposició, i en segon terme el tractament estadístic de la matriu de semblança amb la finalitat d'extreure informació que ajudi a explicar les propietats experimentals.

## 6.1 Evolució de les anàlisis *QSAR* basades en *MQSM*

En els propers apartats s'exposarà l'evolució soferta en els diferents processos involucrats en el càlcul de les *MQSM* en la darrera dècada fins a arribar a l'actual esquema de treball que se segueix en el nostre laboratori quan es realitzen estudis *QSAR*. En els primers treballs encaminats a determinar relacions estructura-activitat,<sup>17,18</sup> es va emprar una aproximació de tipus *CNDO* per descriure les densitats electròniques de les molècules. L'orientació relativa de les molècules en l'espai s'havia fixat a través dels moments dipolars i quadropolars dels compostos analitzats. Això va suposar la primera temptativa d'optimització de la superposició molecular que va ser implementada en el programa *MOLSIMIL88*.<sup>19</sup> Finalment, la matriu de semblança resultant es tractava geomètricament amb la finalitat d'obtenir una ordenació de diferents conjunts moleculars que permetés relacionar qualitativament les mesures de

semblança amb les propietats experimentals. Es va desenvolupar un formalisme teòric que considerava les *MQSM* com un mitjà que permet projectar conjunts de molècules en espais de dimensió finita, en els quals els compostos es poden relacionar mitjançant raonaments de caire geomètric.<sup>17</sup>

### 6.1.1 Descripció de les molècules: funció de densitat electrònica

La primera innovació en la descripció de les molècules va ser la deducció d'unes densitats electròniques moleculars de caire empíric, construïdes mitjançant una aproximació promolecular. La funció densitat es va equiparar a una combinació lineal de funcions esfèriques amb exponents ajustats de manera que la diferència entre el valor de l'autosemblança atòmica *ab initio* i la calculada amb les funcions aproximades fos mínima. Aquesta aproximació es va anomenar *EASA (Empirical Atomic Shell Approximation)*<sup>20-22</sup> i es van definir dos possibles nivells de càlcul. El més simple descrivia la densitat de cada àtom mitjançant una única funció esfèrica, mentre que en una aproximació més completa s'igualava el nombre de funcions al nombre quàntic principal de cada àtom, reproduint així l'essència de l'estructura atòmica en capes. Posteriorment s'han desenvolupat algorismes d'ajust de la densitat *ab initio* a una combinació lineal de funcions esfèriques centrades en els àtoms. En el capítol 3 s'han detallat els algorismes d'ajust de les funcions *ASA*, i s'han presentat els conjunts de funcions atòmiques que s'utilitzen actualment en la construcció de les *PASA DF*. En la majoria d'estudis realitzats amb matrius de semblança s'han emprat les funcions *ASA* ajustades a la base 3-21G i descrites en l'article 3.1.<sup>23-38</sup> Únicament en les sèries químiques on hi intervenen àtoms de nombre atòmic superior al criptó s'ha escollit les funcions *ASA* ajustades a la base de Huzinaga.<sup>27,36</sup> En tots els casos s'ha utilitzat una base mínima, equivalent a descriure els àtoms d'hidrogen amb una única funció, els àtoms del segon període amb tres funcions, els del tercer període amb quatre funcions, i així successivament.

### 6.1.2 Superposició molecular

Quant a la superposició molecular, el primer algorisme dissenyat en el nostre laboratori per cercar el màxim de semblança es basava en la generació aleatòria de *punts inicials* en la hipersuperfície de mesures de semblança, a partir dels quals es feia un refinament amb un mètode Simplex per arribar al màxim de semblança més proper.<sup>39</sup> L'eficàcia de l'acoblament d'ambdues tècniques depèn majoritàriament de dos aspectes relacionats amb els *punts inicials* generats. D'una banda quan més propers al màxim es trobin més facilitat tindrà el mètode Simplex per arribar al cim, i en segon lloc l'exploració de la superfície ha de ser molt completa per garantir que es genera un *punt inicial* del màxim absolut de la funció de semblança. El mètode d'optimització de les *MQSM* que s'utilitza actualment per cercar la sobreposició molecular referent al màxim de semblança és el descrit en el capítol 4. Malgrat la filosofia adoptada és similar a la tècnica que combina un mètode no seqüencial amb un de seqüencial, hi ha unes diferències molt importants. S'ha substituït la generació aleatòria de *punts inicials* per un algorisme que només té en compte *punts maximitzadors* de la funció de semblança, i el mètode Simplex per un mètode de Newton que és molt més eficaç. A més, s'han de realitzar molts menys càlculs de la funció de semblança perquè el refinament amb el mètode seqüencial únicament es fa sobre el *punt maximitzador* absolut, mentre que originàriament es realitzava un càlcul Simplex sobre cada *punt inicial* generat aleatòriament.

El mètode de màxima semblança implementat en el programa MOLSIMIL ha estat un dels propulsors, juntament amb la deducció de les funcions *PASA*, de l'àmplia difusió de les *MQSM* en les anàlisis *QSAR*. S'han deduït moltes correlacions entre matrius de semblança obtingudes amb el mètode de màxima semblança i l'activitat biològica de diverses sèries químiques.<sup>21-28</sup> Però en els darrers anys el mètode de màxima semblança ha anat perdent protagonisme en benefici de l'algorisme *TGSA*. Els últims estudis de relacions estructura-activitat produïts en el nostre laboratori s'han dut a terme amb l'algorisme *TGSA*.<sup>31-38</sup>

### 6.1.3 Tractament estadístic de les matrius de semblança

La matriu de mesures o índexs de semblança que s'obté del procés de superposició molecular de tots els possibles parells de molècules de la sèrie analitzada s'utilitza en qualitat de descriptors moleculars en anàlisis *QSAR*. L'objectiu és extreure el màxim d'informació que contenen les matrius de semblança relacionada amb la propietat investigada. En l'actualitat es fan estudis quantitius i s'utilitzen tècniques estadístiques per construir models matemàtics que relacionin les *MQSM* amb les propietats biològiques. El progrés experimentat en aquesta àrea ha estat possible, en part, al desenvolupament matemàtic que connecta les *MQSM* amb les propietats experimentals de les molècules.<sup>40</sup> El procediment es fonamenta essencialment en el concepte teòric de valor esperat d'un observable i de l'operador hermític que té associat. Llavors les equacions *QSAR/QSPR* queden perfectament definides sobre l'existència d'una relació lineal entre el conjunt de descriptors teòrics i les propietats moleculars.

En les primeres anàlisis *QSAR* emprant matrius de semblança quàntica,<sup>21-23</sup> es va utilitzar les tècniques *PCA* i *PLS* per reduir el nombre de variables, i es va avaluar la qualitat dels models resultants a través dels coeficients de múltiple determinació  $r^2$  definit en l'equació (5.25) i de predicció  $q^2$  descrit en l'equació (5.28). Posteriorment s'han codificat nous procediments estadístics amb la finalitat de construir models matemàtics amb major capacitat predictiva. Per exemple, s'ha programat una tècnica de *classical scaling* per transformar les matrius de semblança i obtenir un nou conjunt de variables de dimensió reduïda, sobre el qual es generen els models *MLR*. Un altre progrés ha estat la selecció dels components o coordenades principals obtingudes de la descomposició de la matriu de semblança mitjançant el mètode de les variables més predictives *MPVM*. Ambdues metodologies han estat descrites en el capítol 5, i s'han aplicat a un gran nombre d'estudis pràctics de semblança molecular quàntica.<sup>24-30,32,33</sup> També s'ha utilitzat la tècnica *PCA* en combinació amb el mètode *MPVM*,<sup>34</sup> o amb l'algorisme de sumes anuades *NSS* per seleccionar el millor model *MLR* d'entre totes les possibles combinacions dels components principals.<sup>31</sup>

La cerca dels millors models *QSAR* ha portat a desenvolupar una tècnica que combina les matrius de semblança  $\mathbf{Z}(\Omega)$  ( $n \times n$ ) obtingudes per a diferents operadors.<sup>24</sup> Les matrius de semblança més usuals en els càlculs de *MQSM* són les de solapament i de Coulomb. Mitjançant combinacions lineals amb coeficients definits positius de les

matrius  $\mathbf{Z}(\Omega)$  originals, es defineix una nova matriu de semblança que integra la informació més rellevant de cada tipus de *MQSM* segons l'activitat biològica estudiada. El conjunt de coeficients que multipliquen cadascuna de les matrius  $\mathbf{Z}(\Omega)$  ha de complir les condicions de convexitat descrites en l'equació (3.53), i es determinen de manera que el coeficient de predicció  $q^2$  del model *QSAR* resultant sigui màxim. Aquesta metodologia s'ha aplicat en diverses sèries químiques.<sup>24,25,28</sup> Però la progressiva manipulació de la informació molecular continguda en les matrius de semblança comporta un increment del risc d'obtenir correlacions casuals o sobreparametritzar el model. En atenció a això s'han codificat algorismes matemàtics que permeten validar estadísticament els models *QSAR* resultants, com és el test d'aleatorietat que ajuda a descartar correlacions fortuïtes,<sup>25</sup> o fer prediccions sobre conjunts externs de molècules.

Les tècniques *QSAR* basades en les matrius de semblança han permès caracteritzar propietats moleculars i activitats biològiques per a una gran varietat de conjunts moleculars.<sup>21-38</sup> Entre elles es podrien destacar les que fan referència a la predicció de l'activitat de compostos antimalarials<sup>31,34</sup> i antitumorals,<sup>24,27</sup> la descripció de la resposta biològica d'alguns fàrmacs a enllaçar-se amb un enzim,<sup>23,25,28,30,37,38</sup> la caracterització de la toxicitat de compostos orgànics,<sup>26,32,33,35</sup> i la determinació de l'estabilitat de les proteïnes en mutacions d'un sol aminoàcid.<sup>29</sup>

## 6.2 Influència de determinats factors en les anàlisis *QSAR*

En aquest apartat es vol analitzar com afecten determinats factors en els resultats estadístics dels models *QSAR* construïts amb matrius de semblança. En primer lloc s'estudiarà la repercussió que té en els resultats finals la utilització de diferents funcions densitat per descriure les molècules considerades. És molt important contrastar el comportament de les funcions *PASA* envers les *ab initio*, i veure si s'aprecien variacions en els models *QSAR*. En segon lloc s'examinarà la influència de la superposició molecular en les relacions estructura-activitat a través de comparar els resultats originats amb el mètode de màxima semblança i l'algorisme *TGSA*. I finalment s'observaran les conseqüències que tenen en els coeficients estadístics variacions en l'estructura molecular.

Per veure l'efecte dels diferents factors en els models *QSAR* s'ha escollit un conjunt de derivats de l'anilina que presenten una alta toxicitat en aigua. La toxicitat és un cas puntual d'activitat biològica de gran interès entre la comunitat científica. En aquest cas s'estudia la toxicitat aquàtica avaluada com la concentració letal necessària que redueix en un 50% una població d'una espècie de peixos anomenada *Poecilia reticulata*.<sup>41</sup> Les dades referides als compostos aromàtics considerats i als seus corresponents valors de  $-\log EC_{50}$  es mostren en la taula 6.1.

No	Compost	LC50	No	Compost	LC50
1	Anilina	-2.91	15	2,5-Dicloranilina	-4.99
2	2-Metilanilina	-3.12	16	3,4-Dicloranilina	-4.39
3	3-Metilanilina	-3.47	17	3,5-Dicloranilina	-4.62
4	4-Metilanilina	-3.72	18	2,3,4-Tricloranilina	-5.15
5	<i>N,N</i> -Dimetilanilina	-3.33	19	2,3,6-Tricloranilina	-4.73
6	2-Etilanilina	-3.21	20	2,4,5-Tricloranilina	-4.92
7	3-Etilanilina	-3.65	21	$\alpha,\alpha,\alpha,4$ -Tetrafluoro-3-metilanilina	-3.77
8	4-Etilanilina	-3.52	22	$\alpha,\alpha,\alpha,4$ -Tetrafluoro-2-metilanilina	-3.78
9	4-Butilanilina	-4.16	23	Pentafluoroanilina	-3.69
10	2,5-Diisopropilanilina	-4.06	24	2-Nitroanilina	-4.15
11	2-Cloranilina	-4.31	25	3-Nitroanilina	-3.24
12	3-Cloranilina	-3.98	26	4-Nitroanilina	-3.23
13	4-Cloranilina	-3.67	27	2-Cloro-4-nitroanilina	-3.93
14	2,4-Dicloranilina	-4.41			

**Taula 6.1** Derivats de l'anilina i toxicitat aquàtica expressada en  $-\log EC_{50}$ .

Els 27 compostos llistats en la taula 6.1 s'han extret d'un conjunt més ampli estudiat anteriorment amb descriptors clàssics,<sup>41</sup> i per nosaltres mateixos emprant matrius de semblança.<sup>26</sup> S'ha seleccionat aquesta sèrie química perquè és un conjunt amb poques molècules, que presenten una estructura força rígida i amb un nombre de partícules que no és prohibitiu a l'hora de realitzar els càlculs *ab initio*. En el present estudi no interessa reproduir els resultats estadístics obtinguts anteriorment per aquesta sèrie, sinó que es vol fer servir de test per veure la influència de diversos factors en els models *QSAR* finals. És per això que en les anàlisis que es presentaran a continuació



s'ha utilitzat la tècnica *PCA* per reduir la dimensió de les matrius de semblança i s'han escollit els primers components que donen la màxima variança per generar les regressions multilineals amb la toxicitat aquàtica, mentre que en l'article [26] s'havia utilitzat una tècnica de *classical scaling* per descompondre la matriu de semblança i el mètode *MPVM* per seleccionar les variables. La intenció és utilitzar un mètode estadístic simple i robust, que no influeixi excessivament en els resultats finals.

### 6.2.1 Influència de la densitat electrònica

El primer factor que s'examina és l'efecte de la densitat electrònica en les correlacions finals estructura-activitat. El principal objectiu és comprovar si hi ha diferències significatives entre l'ús de densitats *ab initio* i *PASA* en la descripció de les molècules. Un segon aspecte que s'analitzarà és la influència del nombre de funcions *ASA* en la descripció dels àtoms. Per fer-ho s'han avaluat dos conjunts de funcions *PASA*. El primer és una base mínima, que ha estat àmpliament utilitzada en anàlisis *QSAR*,<sup>23-38</sup> i que s'identificarà amb les sigles *PASA(1/3/4)*. Correspon a representar els àtoms de H amb una funció esfèrica, tres funcions per descriure els àtoms C, O, N i F, i quatre funcions per al Cl. Aquest conjunt de funcions ha estat descrit en l'article 3.1. La segona base és la *PASA(3/5/6)*, que ja havia estat definida en l'article 3.3, i on la seqüència del nombre de funcions és 3 per a l'hidrogen, 5 per als àtoms del segon període i 6 per als del tercer.

Les geometries dels 27 derivats de l'anilina s'han optimitzat amb el mètode HF i el conjunt de funcions de base 6-31G\* existent en el programa GAUSSIAN.<sup>42</sup> En l'actualitat, únicament es poden realitzar càlculs de semblança exactes en mesures puntuals, com per exemple les *QS-SM*, i en sistemes amb poques partícules, com és la sèrie considerada. En total s'ha d'avaluar la integral de semblança  $Z_{AB}$  de 351 parells de molècules diferents, més les 27 mesures d'autosemblança  $Z_{AA}$ . Per tenir una mateixa superposició molecular en les diferents anàlisis que es realitzaran, s'ha escollit l'alineament molecular que proporciona l'algorisme *TGSA* descrit en la taula 4.6. Així la superposició molecular no tindrà influència en els resultats presentats en aquest apartat.

L'esquema de treball proposat per als estudis de semblança del conjunt d'anilines és el següent. La primera fase és la determinació de les geometries moleculars en el nivell HF/6-31G\*. El següent procés és la fixació de la sobreposició estructural òptima de tots els parells possibles de molècules amb l'algorisme *TGSA*. Llavors, sobre cada superposició molecular s'avaluen les mesures de solapament i de Coulomb emprant funcions *ab initio* i *PASA*. Referent a les mesures exactes s'han determinat les matrius densitat corresponents als càlculs HF/3-21G i HF/6-311G. Mentre que per a les mesures aproximades s'han utilitzat les densitats *PASA(1/3/4)* i *PASA(3/5/6)*. En total s'han calculat 351 integrals  $Z_{AB}$  més 27 integrals  $Z_{AA}$  per a quatre funcions densitat diferents i dos tipus de mesures, la de recobriment i la de Coulomb.

Abans de construir els diferents models *QSAR*, s'han comparat els valors de semblança obtinguts amb les densitats exacta i aproximada. Confrontant els resultats HF/3-21G amb els *PASA(1/3/4)* s'observa que per les 27 mesures *QS-SM* de tipus solapament, el màxim error relatiu es produeix en la molècula **1**, i és del 1.4%. La mitjana dels errors relatius produïts en les mesures  $Z_{AA}$  és del 0.8 %. Respecte als valors de fora la diagonal de la matriu de semblança de recobriment, l'error màxim relatiu és del 3.0 % i es dona en la mesura de semblança entre les molècules **20** i **21**. La mitjana aritmètica en les integrals  $Z_{AB}$  és del 1.1%, i només 11 mesures del total de 351 tenen un error relatiu superior al 2 %. Aquests resultats estan en concordança amb els errors relatius mostrats en l'article 3.1 referits als derivats del metà. Respecte a les integrals de Coulomb, la comparació entre les mesures *PASA(1/3/4)* i HF/3-21G mostra un comportament similar al de les mesures de solapament. Per exemple, en les mesures d'autosemblança l'error màxim és de 1.7 % per la molècula **23**, mentre que la mitjana aritmètica és del 0.7 %. En els valors  $Z_{AB}$  s'obtenen errors inferiors al 1.6 %, essent la mitjana aritmètica del 0.7 %.

La confrontació dels valors de les integrals emprant les densitats *PASA(3/5/6)* i HF/6-311G demostra també la bona afinitat entre les mesures aproximades i exactes. En les mesures d'autosemblança de solapament l'error relatiu màxim entre les quantitats *PASA* i exacta es dona en el compost **5** i és del 1.6 %, mentre que la mitjana és del 0.9 %. Els errors en les mesures  $Z_{AB}$  són tots inferiors al 2%, amb un error màxim de 1.7 % per la parella de compostos **20** i **23**, i la mitjana és del 1.2 %. Les mesures de Coulomb segueixen la mateixa tendència que les deduïdes amb la base més senzilla.

El següent estadi és el tractament estadístic de les vuit matrius de semblança resultants. Primer s'aplica una tècnica *PCA* amb la finalitat de reduir la dimensió de cada matriu de semblança i llavors es construeixen els models *MLR* amb els primers components. Per valorar els models *QSAR* obtinguts es determinen els coeficients de correlació,  $r$ , i de predicció deduït del procés de validació creuada *LOO*,  $r_{cv}$ . L'elecció del nombre òptim de paràmetres que ha de tenir el model final s'ha fet a través del coeficient de predicció  $r_{cv}$ .

En la taula 6.2 es mostren els resultats de les equacions *MLR* agafant entre 1 i 4 components principals (*PCs*) deduïts de les matrius de semblança de solapament i de Coulomb obtingudes amb les densitats electròniques HF/3-21G i *PASA*(1/3/4). Els valors dels coeficients estadístics llistats en la taula 6.2 demostren clarament que no hi ha cap diferència entre utilitzar densitats electròniques *PASA*(1/3/4) i HF/3-21G en la generació de matrius de semblança que posteriorment són emprades en la deducció de models *QSAR*.

Tipus de mesura	No PCs $k$	<i>PASA</i> (1/3/4)		HF/3-21G	
		$r_{cv}$	$r$	$r_{cv}$	$R$
Solapament	1	0.801	0.830	0.801	0.830
	2	0.815	0.848	0.815	0.848
	3	0.809	0.848	0.809	0.848
	4	0.801	0.852	0.801	0.853
Coulomb	1	0.732	0.770	0.736	0.773
	2	0.820	0.858	0.822	0.859
	3	0.795	0.858	0.797	0.859
	4	0.776	0.859	0.779	0.861

**Taula 6.2** Resultats estadístics emprant densitats *ab initio* i *PASA* sobre les geometries HF/6-31G\*.

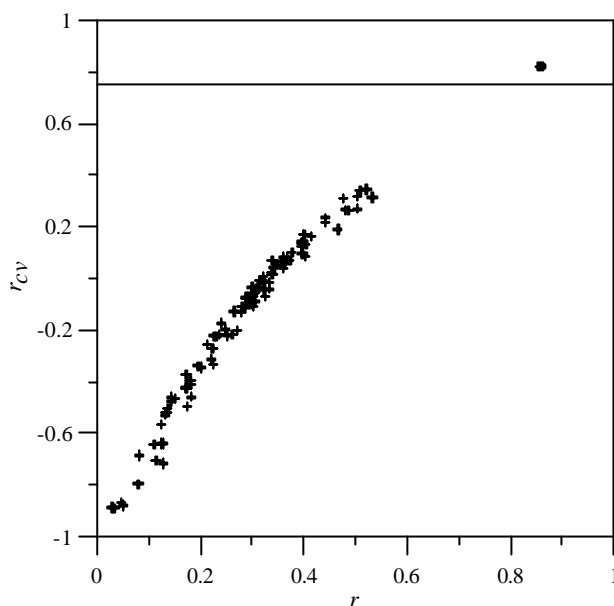
En la taula 6.3 s'exposen els coeficients estadístics resultants de les matrius de semblança deduïdes amb les densitats *PASA*(3/5/6) i HF/6-311G.

Tipus de mesura	No PCs $k$	PASA(3/5/6)		HF/6-311G	
		$r_{cv}$	$r$	$r_{cv}$	$r$
Solapament	1	0.801	0.829	0.801	0.830
	2	0.815	0.848	0.815	0.848
	3	0.809	0.848	0.809	0.848
	4	0.801	0.852	0.801	0.852
Coulomb	1	0.735	0.773	0.737	0.773
	2	0.821	0.859	0.822	0.859
	3	0.797	0.859	0.797	0.860
	4	0.779	0.860	0.779	0.861

**Taula 6.3** Resultats estadístics emprant densitats *ab initio* i PASA sobre les geometries HF/6-31G\*

De manera anàloga a les mesures derivades de la base 3-21G, els resultats de la taula 6.3 confirmen que és completament indistint utilitzar densitats electròniques PASA(3/5/6) i HF/6-311G, amb l'avantatge que tenen les primeres de consumir molt menys temps de càlcul. Una altra conclusió que s'extreu oposant les taules 6.2 i 6.3 és la gran similitud entre els coeficients estadístics PASA(1/3/4) i PASA(3/5/6). Aquests resultats validen les anàlisis QSAR que s'han produït en el nostre laboratori en els darrers anys emprant la base mínima de funcions PASA(1/3/4).<sup>23-38</sup>

Finalment, amb la millor equació MLR de les taules 6.2 i 6.3 s'ha realitzat un test d'aleatorietat per comprovar que el model no està subjecte a correlacions casuals. El valor més gran del coeficient  $r_{cv}$  s'obté amb els models emprant dos PCs deduïts de les matrius de semblança de Coulomb. El test ha consistit en permutar a l'atzar l'ordre dels elements del vector que conté les toxicitats dels 27 derivats de l'anilina, i determinar els coeficients  $r$  i  $r_{cv}$  amb el vector reordenat de les propietats però mantenint invariant la matriu dels components principals de l'anàlisi PCA. En total s'han realitzat 100 permutacions del vector de les toxicitats i s'han guardat els corresponents valors dels coeficients  $r$  i  $r_{cv}$ . En la figura 6.1 s'han representat gràficament els valors dels dos coeficients estadístics. Un model QSAR raonable es considera que ha de tenir un valor de  $q^2$  superior a 0.6,<sup>43</sup> que equivaldria aproximadament a un valor de  $r_{cv} > 0.75$ . En la figura 6.1 s'observa que l'únic model que supera aquest llindar és el real.



**Figura 6.1** Test d'aleatorietat pel model amb dos *PCs* obtingut de les matrius de semblança de Coulomb *PASA*(1/3/4) i l'algorisme *TGSA*. (●) Valor original, (+) valors permutats.

### 6.2.2 Influència de la superposició molecular

En el capítol 4 s'han descrit dos mètodes de superposició molecular. El primer es fonamenta en la definició de la mesura de semblança en el màxim, mentre que el segon té en compte només criteris topològics i geomètrics per deduir el millor alineament molecular des d'un punt de vista estructural. En aquest apartat s'examina l'efecte que té en els resultats finals dels models *QSAR* l'ús d'una o l'altra metodologia. No pretén ser un debat sobre quina de les dues tècniques és millor a l'hora d'estudiar les propietats de les sèries químiques. Ambdues han estat utilitzades amb èxit en diverses anàlisis amb matrius de semblança per establir relacions estructura-activitat.

L'exemple dels derivats de l'anilina, malgrat que són uns compostos molt homogenis, és útil per l'objectiu marcat perquè molts d'ells tenen àtoms de clor en la seva estructura. Els àtoms pesants, com s'ha constatat en el darrer apartat del capítol 4, influeixen decisivament en la deducció de l'alineament molecular. Les mesures de semblança definides per mitjà de distribucions de densitat electrònica donen una gran importància als àtoms amb nombre atòmic superior, el que provoca en alguns casos que s'ignori una part important de l'estructura molecular quan s'utilitza el mètode de màxima semblança.

En la taula 6.4 es llisten els valors dels coeficients  $r$  i  $r_{cv}$  obtinguts amb el mètode de màxima semblança i emprant les densitats  $PASA(1/3/4)$  i  $PASA(3/5/6)$ . En l'actualitat no es pot realitzar un estudi complet *ab initio* per determinar la superposició de màxima semblança perquè el procés d'optimització requereix el còmput de repetides vegades de la integral de semblança. Ha estat precisament la deducció de les funcions  $PASA$  el que ha permès superar aquest escull i poder realitzar estudis sobre molècules de grans dimensions.

Un primer examen dels valors presentats en la taula 6.4 corrobora que els resultats emprant la base de funcions mínima  $PASA(1/3/4)$  són equivalents als obtinguts amb la base més completa  $PASA(4/5/6)$ . D'altra banda, comparant els resultats de la taula 6.4 amb els deduïts amb el mètode  $TGSA$  de les taules 6.2 i 6.3, s'observen diferències significatives en els valors dels coeficients estadístics. On s'aprecien més variacions és en el coeficient de predicció  $r_{cv}$  de les mesures de Coulomb.

Tipus de mesura	No PCs $k$	$PASA(1/3/4)$		$PASA(3/5/6)$	
		$r_{cv}$	$r$	$r_{cv}$	$r$
Solapament	1	0.796	0.826	0.797	0.826
	2	0.792	0.831	0.792	0.831
	3	0.755	0.831	0.755	0.831
	4	0.694	0.845	0.690	0.845
Coulomb	1	0.761	0.794	0.764	0.796
	2	0.836	0.867	0.838	0.868
	3	0.818	0.868	0.820	0.870
	4	0.787	0.869	0.789	0.870

**Taula 6.4** Resultats estadístics emprant l'algorisme de màxima semblança i la geometria HF/6-31G\*.

Per fer més evident la diferència entre els resultats emprant l'algorisme  $TGSA$  i el de màxima semblança, en la taula 6.5 es llisten els valors de la toxicitat predits mitjançant la validació creuada  $LOO$  en els models amb 2  $PCs$ . Els resultats indiquen que majoritàriament els compostos amb àtoms de clor en la seva estructura són els que presenten les variacions més significatives entre els dos mètodes d'alineament.

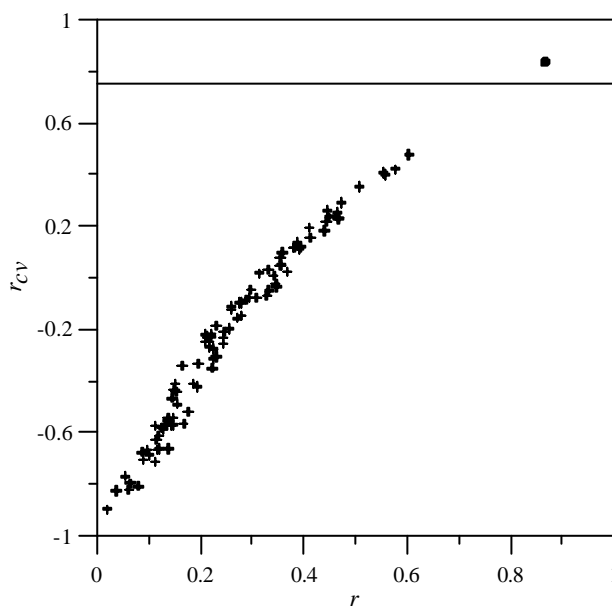
No	Mesura de solapament		Mesura de Coulomb	
	Max. Semb.	TGSA	Max. Semb.	TGSA
1	-3.59	-3.58	-3.28	-3.36
2	-3.56	-3.55	-3.44	-3.47
3	-3.53	-3.54	-3.42	-3.45
4	-3.50	-3.52	-3.35	-3.40
5	-3.53	-3.55	-3.44	-3.38
6	-3.54	-3.51	-3.50	-3.44
7	-3.51	-3.53	-3.48	-3.48
8	-3.51	-3.50	-3.45	-3.46
9	-3.47	-3.50	-3.49	-3.49
10	-3.50	-3.46	-3.84	-3.62
11	-4.21	-4.22	-3.92	-3.99
12	-4.24	-4.32	-3.97	-4.00
13	-4.24	-3.96	-3.95	-3.97
14	-4.55	-4.30	-4.60	-4.52
15	-4.49	-4.62	-4.39	-4.38
16	-4.57	-4.54	-4.51	-4.58
17	-4.50	-4.65	-4.46	-4.41
18	-4.58	-5.02	-4.75	-4.72
19	-4.94	-4.76	-5.09	-5.05
20	-4.90	-4.47	-5.01	-4.97
21	-3.57	-3.57	-3.90	-4.01
22	-3.56	-3.57	-3.90	-4.01
23	-3.59	-3.64	-3.94	-4.05
24	-3.48	-3.46	-3.54	-3.49
25	-3.54	-3.56	-3.64	-3.58
26	-3.59	-3.60	-3.61	-3.56
27	-4.24	-4.39	-4.36	-4.35

**Taula 6.5** Comparació dels valors predits de la toxicitat aquàtica emprant els dos mètodes de superposició molecular i les *PASA(1/3/4) DF*.

Un aspecte que no s'ha examinat en els apartats anteriors és quin tipus d'informació proporcionen les mesures de solapament i de Coulomb, i si és diferent. Si s'analitzen únicament els coeficients  $r$  i  $r_{cv}$  de les taules 6.2, 6.3 i 6.4, podria semblar que per l'exemple dels derivats de l'anilina les mesures de solapament i de Coulomb proporcionen models *QSAR* similars. Però en canvi els models matemàtics són molt diferents com ho demostren les dades presentades en la taula 6.5. Els valors predits de la

toxicitat varien significativament entre les mesures de solapament i de Coulomb, la qual cosa demostra que la informació que donen ambdues mesures és molt diferent. Normalment s'atribueix a les *MQSM* de solapament un fort caràcter estructural, mentre que les de Coulomb descriuen un perfil electrostàtic en les relacions estructura-activitat. En la majoria d'estudis *QSAR* orientats a predir l'activitat de fàrmacs s'ha comprovat que les mesures de Coulomb es comporten millor que les de recobriment.

Finalment, en la figura 6.2 es mostra el test d'aleatoriat resultant de la generació de 100 permutacions de l'ordre del vector de les toxicitats i posterior construcció dels corresponents models amb els dos primers *PCs* obtinguts de la descomposició de la matriu de mesures de semblança de Coulomb *PASA(1/3/4)* definides en el màxim. Els resultats són molt similars als de l'algorisme *TGSA* mostrats en la figura 6.1.



**Figura 6.2** Test d'aleatorietat pel model amb dos *PCs* obtingut de les matrius de semblança de Coulomb *PASA(1/3/4)* i el mètode de màxima semblança.

(●) Valor original, (+) valors permutats.



### 6.2.3 *Influència de la geometria molecular*

Un altre aspecte a tenir en compte en els estudis de semblança molecular és la conformació que s'escull per representar les molècules. Quan s'investiguen propietats biològiques relacionades amb l'acció de fàrmacs, el més correcte seria considerar la conformació activa de les molècules, obtinguda a través d'anàlisis de raig X o de difracció de neutrons. Però en la majoria de casos la interacció lligand–receptor no és coneguda, o bé s'ignora quina és l'estructura del receptor, o bé la cavitat activa no ha estat identificada. Llavors s'ha de recórrer a mètodes de càlcul teòric per obtenir les geometries moleculars. Els més comuns són les tècniques d'orbitals moleculars *ab initio* i semiempírics, i els mètodes de mecànica molecular o camps de força. L'elecció d'una o altra metodologia depèn d'una sèrie de factors, com el nombre de partícules que componen les molècules analitzades, el tipus d'informació requerida i el temps de càlcul disponible. Normalment no és possible realitzar un tractament *ab initio* del problema degut a la complexitat dels sistemes biològics. El més habitual és recórrer a mètodes semiempírics o de mecànica molecular per optimitzar la geometria de les molècules aïllades, sense considerar els efectes del medi que les envolta quan actuen sobre el receptor. Aquest ha estat el criteri adoptat en la majoria d'estudis de semblança molecular quàntica que s'han dut a terme en el nostre laboratori.<sup>21-29,31-38</sup> Una excepció va ser l'anàlisi d'un conjunt de derivats de la 3-amidinobenzilalanina que s'enllacen amb diferents enzims.<sup>30</sup> Precisament la deducció de l'estructura cristal·lina d'un dels compostos interaccionant amb el receptor tripsina ha permès conèixer la conformació activa d'aquestes molècules i tenir-la en compte en l'estudi de semblança.

El darrer factor que s'examina en el conjunt de 27 derivats de l'anilina és la influència de la geometria molecular en els resultats finals dels models *QSAR*. Per comprovar-ho es compararan els resultats obtinguts en els apartats anteriors, on s'han utilitzat les geometries HF/6-31G\*, amb els valors resultants de les geometries optimitzades amb el programa AMPAC<sup>44</sup> i l'Hamiltonià semiempíric AM1.<sup>45</sup> En la taula 6.6 es llisten els resultats estadístics dels models *QSAR* obtinguts emprant les geometries AM1 i les densitats *PASA*(1/3/4). Si es comparen amb els valors donats en les taules 6.2 i 6.4, s'aprecien algunes variacions en els coeficients  $r$  i  $r_{cv}$ . Els resultats estadístics per les mesures de solapament són lleugerament millors emprant les geometries AM1, mentre que en les de Coulomb la tendència és a la inversa.

Tipus de mesura	No PCs <i>k</i>	<i>TGSA</i>		Màxim de Semblança	
		<i>r<sub>cv</sub></i>	<i>r</i>	<i>r<sub>cv</sub></i>	<i>r</i>
Solapament	1	0.818	0.843	0.802	0.830
	2	0.816	0.849	0.792	0.831
	3	0.807	0.850	0.784	0.835
	4	0.789	0.851	0.726	0.840
Coulomb	1	0.740	0.777	0.766	0.797
	2	0.818	0.858	0.829	0.863
	3	0.792	0.858	0.817	0.867
	4	0.761	0.860	0.786	0.868

**Taula 6.6** Resultats estadístics emprant *PASA(1/3/4)* i les geometries AM1.

L'exemple de les anilines és molt simple, i les variacions produïdes entre les dues estructures, HF/6-31G\* i AM1, no són massa grans. Són uns compostos aromàtics que no presenten excessius problemes conformacionals. Tenen una estructura força rígida, on únicament algun dels substituents té algun grau de llibertat conformacional, però amb mínims energètics perfectament definits. Malgrat això, les dades exposades en la taula 6.6 demostren que la geometria molecular té influència en els resultats finals dels models *QSAR* quan s'utilitzen les matrius de semblança en qualitat de descriptors moleculars.



## Discussió

Les matrius de semblança molecular quàntica han demostrat que són un instrument útil en l'àmbit del modelatge molecular, on contribueixen de manera eficient en la deducció de relacions estructura-activitat. La seva aplicació en *3D-QSAR* s'ha vist beneficiada, igualment, pel desenvolupament de noves metodologies que faciliten el propi càlcul de les mesures de semblança. Així es poden destacar dos avenços importants en aquest camp. D'una banda les funcions *PASA* han millorat substancialment el temps de càlcul de les *MQSM* i en segon lloc els algorismes de superposició molecular han permès donar una correcta definició de les mesures de semblança. Una altra línia d'investigació ha anat dirigida al disseny de noves tècniques que connecten els descriptors mecanicoquàntics amb l'activitat biològica, i en definitiva, permeten generar models *QSAR* més predictius. En aquest àmbit s'ha desenvolupat un procediment que combina diferents matrius de semblança per mitjà de la teoria dels conjunts convexes.

L'estudi dels derivats de l'anilina ha mostrat alguns aspectes claus de les *MQSM* quan s'apliquen en les anàlisis *QSAR*. El primer, i tal vegada el més important, ha estat la constatació que l'elecció del tipus de densitat electrònica per descriure les molècules no té incidència en els resultats estadístics dels models *QSAR* finals. Els resultats obtinguts emprant funcions densitat *PASA* són equivalents als *ab initio*, en canvi suposen un important estalvi de temps de computació. També s'ha constatat que emprant una base mínima de funcions *ASA* atòmiques, corresponent a descriure els àtoms d'hidrogen amb una funció, els àtoms del segon període amb tres funcions, els del tercer amb quatre i així successivament, és suficient per construir la densitat de les molècules amb una aproximació promolecular. En canvi, la geometria molecular i el mètode de superposició escollit per definir l'alineament de tots els possibles parells de molècules que formen el conjunt analitzat, són dos factors que sí repercuteixen en els resultats *QSAR*. Possiblement, l'aportació de noves dades en un futur proper permetrà establir quin mètode de superposició molecular és el millor per a cada sèrie química. També, en els futurs estudis de relacions estructura-activitat, s'haurà de tenir més en compte quina conformació molecular s'escull per realitzar els càlculs de semblança. És evident que optimitzar totes les geometries moleculars en la conformació de mínima energia és només una aproximació. Darrerament s'han produït molts avenços en la

simulació i manipulació informàtica de les molècules en tres dimensions que ha proporcionat una informació essencial en l'estudi de la interacció entre lligands i receptors macromoleculars.

## Referències

1. Y. C. Martin. Quantitative drug design. A critical introduction. Marcel Dekker, New York, 1978.
2. H. Kubinyi (ed.). 3D QSAR in drug design: theory methods and applications. ESCOM Science Publishers B.V., Leiden, The Netherlands, 1993.
3. M. A. Johnson, G. Maggiora (Eds.). Concepts and applications of molecular similarity. John Wiley & Sons, Inc., New York, 1990.
4. P. M. Dean (ed.). Molecular similarity in drug design. Blackie Academic & Professional, London, 1995.
5. K. Sen (Ed.). Molecular similarity. *Topics in Current Chemistry*. Springer Verlag, Berlin, volums 173 i 174, 1995.
6. R. Carbó (Ed.). Molecular similarity and reactivity: from quantum chemical to phenomenological approaches. Understanding chemical reactivity. Kluwer Academic, volum 14, Dordrecht, 1995.
7. R. Carbó-Dorca, P. G. Mezey (Eds.). Advances in molecular similarity. JAI Press, London, volum 1, 1996. Volum 2, 1998.
8. R. Carbó-Dorca, X. Gironés, P. G. Mezey (Eds.). Fundamentals of molecular similarity. Kluwer Academic/Plenum Press, New York, 2001.
9. A. J. Hopfinger. A QSAR investigation of DHFR inhibition by Baker triazines based upon molecular shape analysis. *J. Am. Chem. Soc.* **1980**, *102*, 7196–7206.
10. R. Carbó, L. Leyda, M. Arnau. How Similar is a molecule to another? An electron density measure of similarity between two electronic structures. *Int. J. Quantum Chem.* **1980**, *17*, 1185–1189.
11. S. Namasivayam, P. M. Dean. Statistical method for surface pattern matching between dissimilar molecules: electrostatic potentials and accessible surfaces. *J. Mol. Graph.* **1986**, *4*, 46–50.
12. E. E. Hodgkin, W. G. Richards. Molecular similarity based on electrostatic potential and electric field. *Int. J. Quant. Chem., Quant. Biol. Symp.* **1987**, *14*, 105–110.
13. G. Rum, W. C. Herndon. Molecular similarity concepts. 5. Analysis of steroid-protein binding constants *J. Am. Chem. Soc.* **1991**, *113*, 9055–9060.
14. A. C. Good, S. S. So, W. G. Richards. Structure-activity relationships from molecular quantum similarity matrices. *J. Med. Chem.* **1993**, *36*, 433–438.
15. A. C. Good, S. J. Peterson, W. G. Richards. QSAR's from similarity matrices. Technique validation and application in the comparison of different similarity evaluation methods. *J. Med. Chem.* **1993**, *36*, 2929–2937.
16. A. C. Good, W. G. Richards. The extension and application of molecular similarity to drug design. *Drug Information Journal* **1996**, *30*, 371–388.
17. R. Carbó, B. Calabuig. Molecular quantum similarity measures and  $n$ -dimensional representation of quantum objects. I. Theoretical foundations. *Int. J. Quantum Chem.* **1992**, *42*, 1681–1693.
18. R. Carbó, B. Calabuig. Molecular quantum similarity measures and  $n$ -dimensional representation of quantum objects. II. Practical applications. *Int. J. Quantum Chem.* **1992**, *42*, 1695–1709.

19. R. Carbó, B. Calabuig. Molsimil-88: molecular similarity calculations using a CNDO-like approximation. *Comp. Phys. Commun.* **1989**, *55*, 117–126.
20. P. Constans, Ll. Amat, X. Fradera, R. Carbó-Dorca. Quantum molecular similarity measures (QMSM) and the atomic shell approximation (ASA). Publicat en el llibre: *Advances in molecular similarity*. R. Carbó-Dorca, P. G. Mezey (Eds.). JAI Press, London, volum 1, pàgines 187–211, 1996.
21. R. Carbó-Dorca, E. Besalú, Ll. Amat, X. Fradera. Quantum molecular similarity measures: concepts, definitions, and applications to quantitative structure-property relationships. Publicat en el llibre: *Advances in molecular similarity*. R. Carbó-Dorca, P. G. Mezey (Eds.). JAI Press, London, volum 1, pàgines 1–41, 1996.
22. X. Fradera, Ll. Amat, E. Besalú, R. Carbó-Dorca. Application of molecular quantum similarity to *QSAR*. *Quant. Struct.-Act. Relat.* **1997**, *16*, 25–32.
23. M. Lobato, Ll. Amat, E. Besalú, R. Carbó-Dorca. Structure-activity relationships of a steroid family using quantum similarity measures and topological quantum similarity indices. *Quant. Struct.-Act. Relat.* **1997**, *16*, 465–472.
24. Ll. Amat, D. Robert, E. Besalú, R. Carbó-Dorca. Molecular quantum similarity measures tuned 3D *QSAR*: an antitumoral family validation study. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 624–631.
25. D. Robert, Ll. Amat, R. Carbó-Dorca. 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.
26. D. Robert, R. Carbó-Dorca. Aromatic compounds aquatic toxicity *QSAR* using molecular quantum similarity measures. *SAR QSAR Environ. Res.* **1999**, *10*, 401–422.
27. R. Carbó-Dorca, Ll. Amat, E. Besalú, X. Gironés, D. Robert. Quantum mechanical origin of *QSAR*: theory and applications. *J. Mol. Struct. (Theochem)* **2000**, *504*, 181–228.
28. R. Carbó-Dorca, D. Robert, Ll. Amat, X. Gironés, E. Besalú, Molecular quantum similarity in *QSAR* and drug design. *Lecture Notes in Chemistry*, 73, Springer Verlag, Berlin, 2000.
29. D. Robert, X. Gironés, R. Carbó-Dorca. Quantification of the influence of single-point mutations on haloalkane dehalogenase activity: a molecular quantum similarity study. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 839–846.
30. D. Robert, Ll. Amat, R. Carbó-Dorca. Quantum similarity *QSAR*: study of inhibitors binding to Trombin, Trypsin, and Factor Xa, including a comparison with CoMFA and CoMSIA methods. *Int. J. Quantum Chem.* **2000**, *80*, 265–282.
31. X. Gironés, A. Gallegos, R. Carbó-Dorca. Modeling antimalarial activity: application of kinetic energy density quantum similarity measures as descriptors in *QSAR*. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1400–1407.
32. D. Robert, X. Gironés, R. Carbó-Dorca. Molecular quantum similarity measures as descriptors for quantum *QSAR*. *Polycyclic Aromatic Compounds* **2000**, *19*, 51–71.
33. A. Gallegos, D. Robert, X. Gironés, R. Structure-toxicity relationships of polycyclic aromatic hydrocarbons using molecular quantum similarity. *J. Comput.-Aided Mol. Design* **2001**, *15*, 67–80.
34. X. Gironés, A. Gallegos, R. Carbó-Dorca. Antimalarial activity of synthetic 1,2,4-trioxane and cyclic peroxy ketals, a quantum similarity study. *J. Comput.-Aided Mol. Design* **2001**, *15*, 1053–1063.
35. R. Carbó-Dorca, Ll. Amat, E. Besalú, X. Gironés, D. Robert. Quantum molecular similarity measures: theory and applications to the evaluation of molecular properties, biological activities and

- toxicity. Publicat en el llibre: *Fundamentals of molecular similarity*. R. Carbó-Dorca, X. Gironés, P. G. Mezey (Eds.). Kluwer Academic/Plenum Press, New York, 2001.
36. X. Gironés, R. Carbó-Dorca. Using molecular quantum similarity measures under stochastic transformation to describe physical properties of molecular systems. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 317–325.
  37. E. Besalú, X. Gironés, Ll. Amat, R. Carbó-Dorca. Molecular quantum similarity and the fundamentals of *QSAR*. *Acc. Chem. Res.* **2002**, *35*, 289–295.
  38. X. Gironés, R. Carbó-Dorca. Molecular quantum similarity-based *QSAR*'s for binding affinities of several steroids sets. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1185–1193.
  39. Ll. Amat, E. Besalú, R. Carbó, X. Fradera. Practical applications of quantum molecular similarity measures (QMSM): Programs and examples. *SCIENTIA gerundensis* **1995**, *21*, 127–143.
  40. R. Carbó, E. Besalú, Ll. Amat, X. Fradera. Quantum molecular similarity measures (QMSM) as a natural way leading towards a theoretical foundation of quantitative structure-properties relationships (*QSPR*). *J. Math. Chem.* **1995**, *18*, 237–246.
  41. E. U. Ramos, W. H. J. Vaes, H. J. M. Verhaar, J. L. M. Hermens. Quantitative structure-activity relationships for the aquatic toxicity of polar and nonpolar narcotic pollutants. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 845–852.
  42. GAUSSIAN 98, Revision A.6. M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, V. G. Zakrzewski, J. A. Montgomery Jr., R. E. Stratmann, J. C. Burant, S. Dapprich, J. M. Millam, A. D. Daniels, K. N. Kudin, M. C. Strain, O. Farkas, J. Tomasi, V. Barone, M. Cossi, R. Cammi, B. Mennucci, C. Pomelli, C. Adamo, S. Clifford, J. Ochterski, G. A. Petersson, P. Y. Ayala, Q. Cui, K. Morokuma, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. Cioslowski, J. V. Ortiz, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. Gomperts, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, C. Gonzalez, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, J. L. Andres, C. Gonzalez, M. Head-Gordon, E. S. Replogle, J. A. Pople. Gaussian, Inc.: Pittsburgh, PA, 1998.
  43. S. Clementi, S. Wold. How to choose the proper statistical method. Publicat en el llibre: *Chemometric methods in molecular design*. H. Van de Waterbeemd (Ed.). VCH, New York, pàgines 319–338, 1995.
  44. AMPAC 6.01, Semichem, Inc., 7128 Summit, Shawnee, KS 66216. D.A.
  45. M. J. S. Dewar, E. G. Zoebisch, E. F. Healy, J. J. P. Stewart. AM1: a new general purpose quantum mechanical molecular model. *J. Am. Chem. Soc.* **1985**, *107*, 3902–3909.





## 7. Aproximació *QS-SM* de fragments

---

Els treballs que es presentaran en aquest capítol tenen com a objectiu definir uns paràmetres teòrics que poden ser utilitzats en lloc dels descriptors físico-químics en els models de *QSAR* clàssica. A fi i efecte de corroborar l'eficàcia de l'alternativa proposada, primerament es comprova que determinades mesures d'autosemblança quàntica (*QS-SM*) estan estretament lligades amb alguns paràmetres empírics com el coeficient de partició octanol-aigua,  $\log P$ , i la constant  $\sigma$  de Hammett. El següent treball ha consistit en reproduir els resultats d'alguns models *QSAR* construïts mitjançant el mètode de Hansch, substituint els paràmetres físico-químics de les equacions empíriques pels seus equivalents teòrics. I finalment, en l'últim apartat d'aquest capítol, es proposa una metodologia general de recerca de les *QS-SM* òptimes que millor descriuen l'activitat biològica d'una sèrie de compostos químics. La finalitat és presentar una nova aproximació *QSAR* basada en paràmetres teòrics, amb un rang més gran d'aplicabilitat que els mètodes clàssics i al mateix temps que descriu millor les interaccions lligand-receptor. Els fragments moleculars associats a les *QS-SM* seleccionades mitjançant les tècniques estadístiques s'interpreten com les parts de la molècula amb més influència sobre l'activitat biològica.

### 7.1 *LFER* i models de *QSAR* clàssica

El primer estudi on es constata una relació entre l'estructura i l'activitat de molècules amb interès farmacològic data de l'any 1865, quan Crum-Brown i Fraser van postular que l'acció fisiològica d'una molècula,  $\Phi$ , és funció de la seva constitució física  $C$ , segons l'equació:  $\Phi=f(C)$ .<sup>1</sup> En concret van observar que una modificació química gradual de l'estructura molecular d'una sèrie de verins es traduïa en importants diferències en la seva activitat. Els subsegüents avenços en aquest camp es van produir de manera molt esglaonada. Així cap a l'any 1900 Meyer<sup>2</sup> i Overton<sup>3</sup> varen trobar relacions lineals entre la potència narcòtica de determinats compostos i el seu coeficient

de repartiment en lípids. No va ser fins a l'any 1939 quan a través d'una equació matemàtica Ferguson va relacionar algunes respostes biològiques amb paràmetres físico-químics com per exemple la solubilitat, la tensió de vapor o el punt d'ebullició.<sup>4</sup> També destacar les aportacions fetes per McGowan a començaments dels anys cinquanta,<sup>5,6</sup> en les quals es va relacionar la toxicitat d'alguns compostos amb paràmetres físico-químics com el volum molecular, el coeficient de partició, el punt d'ebullició i alguns termes d'interacció. Ja en els anys seixanta, Free i Wilson van desenvolupar un model molt simple basat en el principi d'additivitat de les contribucions de cada fragment de la molècula a l'activitat total.<sup>7</sup> En els models matemàtics s'inclouen variables lògiques que indiquen la presència o absència de determinades característiques estructurals en sèries homòlogues de compostos.

Paral·lelament als primers estudis de determinacions de relacions entre l'estructura i l'activitat, es va desenvolupar l'aproximació del model lineal relacionat amb l'energia lliure (*Lienar Free Energy Relationships, LFER*). L'any 1937, Hammett es va adonar que la reactivitat dels derivats del benzè en diferents reaccions presenta unes regularitats.<sup>8,9</sup> Concretament un estudi sobre la influència dels substituents en les velocitats d'hidròlisi dels derivats de l'àcid benzoic va permetre a Hammett determinar una relació lineal entre una constant específica per a cada substituent, representada amb la lletra grega  $\sigma$ , i el logaritme de la constant de reactivitat del producte. L'equació de Hammett és el primer referent de la simulació dels efectes dels substituents en les reactivitats dels compostos orgànics utilitzat en models matemàtics de semblança. Posteriorment a la deducció de l'equació de Hammett es va desenvolupar un model de semblança per quantificar els efectes estèrics. Va ser en els anys cinquanta quan Taft va proposar una descripció quantitativa dels efectes estèrics dels substituents per mitjà d'una anàlisi de correlació.<sup>10,11</sup>

L'origen dels mètodes *QSAR* actuals es remunta a l'any 1964, quan Hansch i Fujita varen proposar un model que relacionava l'activitat biològica amb propietats físico-químiques.<sup>12</sup> Els models de *QSAR* clàssica, o també anomenats anàlisis de Hansch, es fonamenten en l'expressió *LFER* de Hammett ampliada amb la inclusió d'altres paràmetres característics de la naturalesa físico-química de les reaccions biològiques. D'aquí va sorgir l'equació anomenada  $\rho\text{-}\sigma\text{-}\pi$ ,<sup>12</sup> en la qual es combinen

linealment dos o més descriptors que relacionen l'activitat biològica amb l'estructura molecular:

$$\log\left(\frac{1}{C}\right) = a \pi^2 + b \pi + \rho \sigma + d \quad (7.1)$$

En l'equació (7.1)  $C$  representa la concentració molar d'un producte necessària per obtenir una resposta biològica determinada,  $\pi$  és el paràmetre de partició del substituent, definit com la diferència entre els logaritmes dels coeficients octanol/aigua del producte substituït i del producte genèric de la sèrie,<sup>13</sup>  $\sigma$  és la constant de Hammett del substituent, i les constants  $\{a, b, \rho, d\}$  són la solució de l'anàlisi de regressió. En els primers models *QSAR* es va considerar que l'activitat biològica depèn exclusivament de les variables físico-químiques del fragment molecular específic que té efecte directe sobre l'activitat, i no de propietats globals de les molècules. D'altra banda, la inclusió del coeficient de partició en els models, permet, com ja havia constatat Overton i d'altres investigadors en estudis precedents, tenir en compte un factor tan important en els processos farmacològics com és la hidrofobicitat dels productes.

El model de Hansch ha anat evolucionant amb la inclusió de nous descriptors físico-químics. Una de les variacions ha estat la consideració de propietats globals de les molècules, com els coeficients de partició o refracció molar, a més de les propietats relatives als substituents. En general, el tipus de paràmetres més emprats en la construcció dels models de *QSAR* clàssica fan referència a efectes electrònics, hidrofòbics, polars i estèrics. Amb aquests descriptors es representen les interaccions més comunes que es donen en els sistemes biològics. Així, mitjançant els models de Hansch és possible deduir els requeriments bàsics d'una determinada activitat, i en conseqüència intuir el mecanisme d'acció del fàrmac. Únicament cal interpretar els descriptors moleculars seleccionats mitjançant les regressions lineals òptimes com a representatius de les interaccions més importants que es donen entre el fàrmac i el receptor. Però desenvolupaments posteriors del mètode han introduït altres paràmetres molt més heterogenis. Per exemple variables binàries com les definides en el model de Free-Wilson que indiquen la presència o absència d'un substituent. L'opció de combinar paràmetres físico-químics i variables lògiques per millorar els resultats estadístics dels models *QSAR* dificulta la dilucidació dels possibles mecanismes d'acció associats als processos farmacològics.

## 7.2 Efectes hidrofòbics

La hidrofobicitat és un factor molt important en les reaccions biològiques, i al mateix temps és un dels fenòmens menys entesos. Hi ha bàsicament dues etapes en els processos farmacològics que es regeixen per efectes hidrofòbics: una és el transport de la molècula des del punt d'administració al lloc d'acció, i l'altra fa referència a la penetració del fàrmac en les membranes cel·lulars i la posterior unió a les zones hidròfugues de l'enzim. Per quantificar la hidrofobicitat d'un compost químic se sol utilitzar una mesura de la distribució de la molècula entre dues fases immiscibles. És l'anomenat coeficient de partició, i es representa amb la lletra  $P$ . El més freqüent és utilitzar l'aigua en qualitat de fase polar i l'octanol com a fase no polar, definint-se el paràmetre molecular  $\log P_{ow}$  o logaritme del coeficient de partició en el sistema de distribució octanol-aigua. A nivell molecular, el coeficient de partició és una mesura dels canvis d'energia lliure del solut associats amb les interaccions d'atracció i repulsió en les fases aquosa i orgànica. El canvi en l'energia lliure ve donat fonamentalment per tres factors: variació en l'entalpia deguda a la interacció entre les molècules de solut i solvent, variació en l'entalpia deguda a la interacció entre les molècules de solvent, i variacions en l'entropia. Els canvis d'entropia provenen principalment de les modificacions en l'estructura de les molècules de solvent que envolten el solut.

El coeficient de partició, entès com un descriptor molecular dels efectes hidrofòbics, ha estat utilitzat en una gran varietat d'estudis *QSAR*. Prenent els treballs de Meyer<sup>2</sup> i d'Overton<sup>3</sup> de finals del segle XIX sobre anestèsics com els pioners en la utilització de models químics que van caracteritzar la hidrofobicitat en estudis estructura-activitat, és evident que els efectes hidrofòbics són anteriors al desenvolupament de descriptors electrònics, iniciats amb l'equació de Hammett que data de l'any 1937,<sup>8</sup> i estèrics, on la deducció de l'equació de Taft no es va produir fins a l'any 1952.<sup>10</sup> Treballs posteriors van confirmar la importància dels efectes hidrofòbics en els processos farmacològics. A destacar els treballs de Collander<sup>14</sup> dels anys cinquanta en els quals es relaciona el coeficient de partició entre la fase aquosa i l'orgànica amb la velocitat de penetració a través de les membranes cel·lulars de plantes. Però no va ser fins als anys seixanta, amb els treballs de Hansch i col·laboradors, que es va generalitzar la utilització del descriptor  $\log P$ , en combinació amb altres paràmetres com la  $\sigma$  de Hammett i la  $E_s$  de Taft, per construir models *QSAR*.

En la bibliografia es poden trobar molts exemples de correlacions entre l'activitat biològica i el paràmetre  $\log P$ .<sup>15</sup> Entre elles es podria destacar les que fan referència a la caracterització de la toxicitat de compostos orgànics;<sup>16</sup> la predicció de l'activitat de compostos antimalarials,<sup>17</sup> antitumorals<sup>18</sup> i anestèsics;<sup>19</sup> també ha estat utilitzat per descriure l'habilitat dels substrats a enllaçar-se a proteïnes<sup>20</sup> i en la interacció amb enzims.<sup>21</sup> Sens dubte els descriptors hidrofòbics han esdevingut un factor clau en l'elucidació de molts models *QSAR*.

Malgrat que el paràmetre  $\log P$  es pot determinar experimentalment, i es tenen els valors per a un gran nombre de molècules, vàries aproximacions teòriques han estat proposades per estimar el seu valor. Una contribució molt important de Hansch i col·laboradors va ser la constatació que el coeficient de partició de les molècules pot ser calculat a partir de les seves estructures químiques.<sup>13</sup> En els primers estudis teòrics es va demostrar la naturalesa additiva del coeficient de partició a partir de la constant  $\pi$  definida per a diferents substituents. Posteriorment, Rekker i col·laboradors van desenvolupar un mètode de càlcul teòric del paràmetre  $\log P$  a partir de fragments moleculars.<sup>22-24</sup> El seu treball es fonamenta en la determinació de les contribucions mitjanes de determinats fragments fent una anàlisi estadística sobre un ampli conjunt de valors de  $\log P$  mesurats experimentalment. Llavors el valor de  $\log P$  d'una molècula qualsevol s'avalua com la suma de les contribucions dels fragments que la formen. Però la fragmentació d'una estructura química no és única, i per tant es pot determinar més d'un valor de  $\log P$  d'una mateixa molècula. Un altre mètode teòric d'estimació del paràmetre  $\log P$  és el de Hansch i Leo, també fonamentat en el càlcul de contribucions de fragments. Són unes mesures de  $\log P$  que han evolucionat i millorat amb el temps fins a donar unes taules molt elaborades on s'exposen els valors de cada fragment.<sup>25</sup> Es defineix un petit nombre de fragments fonamentals, derivats d'unes mesures de  $\log P$  molt precises, i després s'aplica un gran nombre de factors de correcció que simulen els enllaços. A més s'utilitza una sèrie de definicions que eliminen qualsevol ambigüitat en la fragmentació de les estructures. Altres aproximacions teòriques del càlcul de  $\log P$  són les mesures basades en contribucions atòmiques. En concret, destaca el procés desenvolupat per Ghose i Crippen,<sup>26</sup> en el qual s'han classificat els àtoms més comuns que formen els compostos orgànics. També mencionar l'aproximació proposada per Klopman i Iroff,<sup>27,28</sup> que van anomenar mètode de densitat de càrrega, i on utilitzen càlculs basats en la química quàntica.

### 7.2.1 QS-SM com a alternativa a log P

S'ha proposat una alternativa als càlculs teòrics del paràmetre log  $P$  basada en mesures de semblança quàntica.<sup>29</sup> L'objectiu no és fer estimacions dels valors experimentals de log  $P$ , sinó donar un descriptor mecanicoquàntic alternatiu i aplicable a anàlisis QSAR. El procés teòric s'ha basat en la inclusió de l'afinitat hidrofòbica i lipofílica d'una molècula en la seva densitat electrònica simulant la solvatació en aigua i en octanol respectivament. Així s'ha definit una mesura d'autosemblança que involucra el producte de dues distribucions electròniques de la mateixa molècula, però cadascuna d'elles calculades sobre l'estructura molecular immerses en un solvent diferent. La utilització de QS-SM està relacionat amb el mètode basat en contribucions atòmiques de Klopman,<sup>27,28</sup> que involucra densitats de càrrega. La simulació teòrica de la diferent solvatació del solut en fase aquosa i en fase orgànica s'ha fet per mitjà d'un model dielèctric polaritzable, desenvolupat per Tomasi i col·laboradors,<sup>30,31</sup> i que es troba implementat en el programa GAUSSIAN<sup>32</sup> amb les sigles PCM (*Polarized Continuum Model*). El solvent és representat mitjançant una distribució de càrrega  $\mathbf{r}(\mathbf{r})$  en una cavitat incrustada en un medi dielèctric i polaritzable amb permitivitat  $\epsilon$ . La distribució de càrrega molecular induïx en el dielèctric una reacció potencial que actua de retruc contra la mateixa distribució de càrrega, produint alguns canvis respecte a la  $\mathbf{r}(\mathbf{r})$  de l'estat gasos. La polarització del dielèctric es crea a partir d'un sistema de càrregues virtuals sobre la superfície de la cavitat. El procediment seguit en la deducció dels descriptors teòrics ha estat, en primer lloc, l'optimització de la geometria de les molècules en fase gasosa, sobre la qual s'ha fet un càlcul puntual en els dos solvents, aigua i octanol. Tot seguit, amb les densitats de la molècula calculades en els dos medis,  $\mathbf{r}_A^w(\mathbf{r})$  i  $\mathbf{r}_A^o(\mathbf{r})$ , es defineix la mesura de semblança:

$$Z_{AA}^{wo} = \int \mathbf{r}_A^w(\mathbf{r}) \mathbf{r}_A^o(\mathbf{r}) d\mathbf{r}, \quad (7.2)$$

la qual dóna una idea de les diferències entre la configuració hidrofòbica i lipofílica de la molècula A. El mètode s'ha emprat en la determinació dels valors de la mesura  $Z_{AA}^{wo}$  d'una gran varietat de compostos químics i s'ha comprovat que existeix una correlació entre els valors de log  $P$  i la mesura de semblança:

$$\log P = b Z_{AA}^{wo} + a \quad (7.3)$$

En la taula 7.1 es resumeixen els paràmetres estadístics derivats de la recta de regressió (7.3) obtinguts per sèries d'hidrocarburs, amines, alcohols, cetones, èsters, amides, àcids carboxílics i compostos clorats.<sup>29</sup>

Sèrie	Equació lineal	<i>n</i>	<i>r</i> <sup>2</sup>
R-H	$\log P = 0.0323 \times Z_{AA}^{wo} + 0.5332$	6	0.996
R-NH <sub>2</sub>	$\log P = 0.0305 \times Z_{AA}^{wo} - 2.2295$	4	0.996
R-OH	$\log P = 0.0337 \times Z_{AA}^{wo} - 3.3314$	6	0.996
R <sub>1</sub> -CH(OH)-R <sub>2</sub>	$\log P = 0.0324 \times Z_{AA}^{wo} - 3.4537$	6	0.996
R-OH + R <sub>1</sub> -CH(OH)-R <sub>2</sub>	$\log P = 0.0311 \times Z_{AA}^{wo} - 3.1409$	12	0.972
R <sub>1</sub> COR <sub>2</sub>	$\log P = 0.0326 \times Z_{AA}^{wo} - 3.6617$	6	0.993
CH <sub>3</sub> COOR	$\log P = 0.0291 \times Z_{AA}^{wo} - 4.3443$	5	0.996
RCOOH	$\log P = 0.0307 \times Z_{AA}^{wo} - 4.5390$	5	0.998
HCONHR	$\log P = 0.0317 \times Z_{AA}^{wo} - 4.8064$	5	0.999
CH <sub>3</sub> COOR + RCOOH + HCONHR	$\log P = 0.0316 \times Z_{AA}^{wo} - 4.7706$	14	0.996
RCONH <sub>2</sub>	$\log P = 0.0307 \times Z_{AA}^{wo} - 5.0208$	5	0.999
R-Cl	$\log P = 0.0333 \times Z_{AA}^{wo} - 5.4906$	4	0.998

**Taula 7.1** Paràmetres estadístics de les relacions lineals entre la MQSM  $Z_{AA}^{wo}$  i el valor de  $\log P$  per les sèries de compostos estudiats. Resultats publicats en l'article [29].

La definició formal de la mesura  $Z_{AA}^{wo}$  implica el càlcul de la *DF ab initio* en dos solvents diferents, aigua i octanol. Posteriorment, quan s'ha aplicat la metodologia en anàlisis *QSAR*, s'ha observat que la *QS-SM* de la molècula emprant únicament *DF* en fase gas,  $Z_{AA}$ , és suficient per correlar els valors de  $\log P$ , i així s'eviten els càlculs *ab initio*. D'altra banda, també s'han reproduït algunes de les rectes de regressió de la taula 7.1 utilitzant en qualitat de descriptor molecular l'energia de repulsió bielectrònica en lloc de  $Z_{AA}^{wo}$ .<sup>33</sup> Això és possible gràcies a la interpretació de l'energia de repulsió bielectrònica com una *QS-SM*,<sup>34</sup> i per tant com un altre possible descriptor molecular útil en els estudis *QSAR/QSPR*.



### 7.3 Efectes electrònics dels substituents

L'equació de Hammett es va deduir com a resultat de l'apreciació experimental que un determinat substituent dóna efectes similars en diferents reaccions.<sup>8,9</sup> Originalment es va establir una relació empírica per descriure la influència dels substituents en les velocitats d'hidròlisi dels derivats de l'àcid benzoic:

$$\log\left(\frac{k_X}{k_H}\right) = \rho \sigma_X \quad (7.4)$$

on  $k_X$  i  $k_H$  són les constants d'equilibri o de velocitat de les reaccions dels productes substituït i sense substituir en les mateixes condicions,  $\sigma_X$  és una mesura de l'efecte electrònic de substituir H per un substituent X, i  $\rho$  és una constant que depèn del tipus i condicions de la reacció. Si l'expressió (7.4) s'escriu segons:

$$\log(k_X) = \rho \sigma_X + \log(k_H) \quad (7.5)$$

llavors l'equació de Hammett representa una relació lineal entre la constant del substituent,  $\sigma_X$ , i el logaritme de la reactivitat del producte. Donat que el logaritme de la constant d'equilibri és proporcional al canvi de l'energia lliure de Gibbs:

$$\Delta G^\circ = -RT \ln(k) \quad (7.6)$$

es diu que l'equació de Hammett i d'altres de similars estan relacionades amb l'energia lliure (*LFER*). La idea d'utilitzar un model experimental que relacioni els efectes electrònics amb els canvis estructurals es va estendre posteriorment als efectes estèrics<sup>11</sup> i hidrofòbics.<sup>13</sup>

A partir de les dades experimentals de  $pK_a$  de la reacció d'ionització dels derivats de l'àcid benzoic en aigua a 25° es van determinar els valors de la constant  $\sigma$  de Hammett. La constant de la reacció  $\rho$  es defineix igual a  $\rho$  pel procés estàndard de l'àcid benzoic i en les condicions esmentades. En la taula 7.2 es mostren alguns valors de  $\sigma_X$  referits al substituent en posició *para* respecte al grup reactiu de la sèrie aromàtica. L'hidrogen és el substituent de referència, amb el valor de la constant  $\sigma$  igual a zero com es dedueix de l'equació (7.4), mentre que els substituents amb una gran afinitat

pels electrons tenen valors de  $\sigma$  positius, i els substituents donadors d'electrons en termes relatius a l'hidrogen presenten valors negatius.

Substituent (X)	$\sigma_x$
NO <sub>2</sub>	0.81
CN	0.71
CF <sub>3</sub>	0.53
CCl <sub>3</sub>	0.46
Br	0.22
Cl	0.22
F	0.06
H	0
CH <sub>3</sub> CH <sub>2</sub>	-0.13
CH <sub>3</sub>	-0.14
CH <sub>3</sub> O	-0.28
N(CH <sub>3</sub> ) <sub>2</sub>	-0.63

**Taula 7.2** Substituents i constant  $\sigma$  de Hammett dels derivats de l'àcid benzoic *para*-substituint.<sup>35</sup>

### 7.3.1 QS-SM com a substitut de la constant ***s*** de Hammett

L'objectiu dels estudis teòrics que s'exposen seguidament ha estat demostrar que determinades QS-SM poden descriure l'efecte electrònic de la substitució sistemàtica en sèries reactives. D'una manera similar a les mesures de semblança proposades en l'apartat anterior com a alternatives al paràmetre  $\log P$ , en el present estudi es comprova la validesa del nou descriptor teòric mitjançant la deducció de relacions lineals amb la constant  $\sigma$ . Però a diferència de  $\log P$ , on s'ha proposat una QS-SM entre funcions densitat globals de la molècula, per a la constant  $\sigma$  s'utilitza una mesura definida sobre una part específica de la funció densitat. L'objectiu final és obtenir un descriptor teòric que actuï de substitut dels valors de  $\sigma$ , els quals han estat deduïts a partir dels valors experimentals de la constant d'hidròlisi dels derivats de l'àcid benzoic. Llavors, el primer estudi que s'ha suggerit és la correlació de les mesures d'autosemblança del grup COOH dels derivats de l'àcid benzoic amb la constant  $\sigma$ . Posteriorment, per demostrar

que la correlació obtinguda no és específica de la sèrie dels derivats de l'àcid benzoic, s'han analitzat altres sèries d'àcids carboxílics, en totes elles prenent com a fragment característic el grup COOH. Així, en general, les variacions de la densitat electrònica produïdes per la substitució sistemàtica de X en una sèrie reactiva R es poden quantificar a través de la *QS-SM* del fragment COOH:

$$Z_{XX,R}^{\text{COOH}}(\Omega) = \int \int \mathbf{r}_{X,R}^{\text{COOH}}(\mathbf{r}_1) \Omega(\mathbf{r}_1, \mathbf{r}_2) \mathbf{r}_{X,R}^{\text{COOH}}(\mathbf{r}_2) d\mathbf{r}_1 d\mathbf{r}_2, \quad (7.7)$$

on  $\mathbf{r}_{X,R}^{\text{COOH}}(\mathbf{r})$  és la *DF* del fragment COOH, definida en l'apartat 2.4 i corresponent al compost amb substituent X de la sèrie reactiva R. En l'article 7.1 que s'adjunta al final del present apartat, s'ha estudiat la relació lineal entre la constant  $\sigma$  i el descriptor teòric  $Z_{XX,R}^{\text{COOH}}$ ,

$$Z_{XX,R}^{\text{COOH}} = \alpha_R \sigma_X + \beta_R \quad (7.8)$$

en cinc sèries químiques, obtenint unes excel·lents correlacions. Les sèries analitzades són uns derivats de l'àcid benzoic *para*-substituït (**I**), de l'àcid 2-tiofè carboxílic substituït en la posició 5 (**II**), de l'àcid 2-furà carboxílic substituït en la posició 5 (**III**), de l'àcid cinàmic *para*-substituït (**IV**) i de l'àcid fenil acètic *para*-substituït (**V**). Les geometries moleculars s'han optimitzat emprant l'Hamiltonià semiempíric AM1<sup>36</sup> implementat en el programa MOPAC.<sup>37</sup> Referent al càlcul de les integrals  $Z_{XX,R}^{\text{COOH}}$ , s'han utilitzat mesures de tipus solapament entre funcions *PASA* construïdes a partir de la base de funcions atòmiques descrita en l'article 3.1, i definides únicament sobre electrons de valència. A més, els coeficients de les funcions *PASA* s'han modificat amb les càrregues atòmiques AM1 per així apreciar l'efecte del substituent X en el fragment molecular COOH. Els valors del coeficient de correlació són superiors a 0.96 en totes les sèries. A més dels coeficients  $r$ , en l'article 7.1 es representen gràficament les cinc rectes de regressió que permeten veure quins substituents s'han utilitzat en cadascuna de les sèries.

**Estimació de la constant  $r$ .** Una altra correlació interessant s'obté a partir dels pendents  $\alpha_R$  de l'equació lineal (7.8). En l'equació empírica de Hammett (7.4) apareix la constant  $\rho$  que depèn de la sèrie reactiva, del tipus de solvent i de la temperatura. Per caracteritzar d'una manera teòrica la sensibilitat de cada reacció amb referència a l'efecte electrònic dels substituents X i poder-la comparar amb el valor experimental  $\rho$ , s'ha definit el quocient  $\alpha_R/\alpha_0$ , essent  $\alpha_0$  el valor de la reacció dels derivats de l'àcid benzoïc (**I**). En l'article 7.1 es mostren els valors calculats de  $\alpha_R/\alpha_0$  i els experimentals de la constant  $\rho$  de les cinc sèries d'àcids analitzades. S'observa una gran correlació entre els valors experimentals i teòrics, essent el valor del coeficient  $r$  de 0.95.

**Recta de regressió  $Z_{XX,R}^{\text{COOH}}/\text{p}K_a$ .** La mesura de semblança  $Z_{XX,R}^{\text{COOH}}$  és més versàtil que la constant  $\sigma_X$ . Això queda palès en el fet que la *QS-SM* d'un substituent X varia en canviar de sèrie reactiva R, mentre que la constant de Hammett és independent de la naturalesa de la reacció. És per aquest motiu que es va suggerir la possibilitat de correlar les mesures  $Z_{XX,R}^{\text{COOH}}$  amb els valors de la constant d'ionització ( $\text{p}K_a$ ) corresponents a un mateix substituent X en diferents sèries R. En l'article 7.1 es mostra la regressió lineal obtinguda amb els productes del substituent X=H de les cinc sèries considerades. El mateix estudi no es pot fer amb  $\sigma_X$  perquè és constant en totes les sèries R, ni tampoc té cap altre equivalent *LFER* conegut. Això posa de manifest la major versatilitat dels descriptors teòrics, que ha de ser un avantatge quan s'utilitzen les mesures *QS-SM* en lloc dels valors  $\sigma$  en models *QSAR*. La constant de Hammett s'ha aplicat sistemàticament en molts tipus diferents de correlacions,<sup>15</sup> quan en principi ha estat deduïda per descriure l'efecte electrònic de la variació dels substituents en els derivats de l'àcid benzoïc. Els fàrmacs solen ser molècules molt complexes, amb més d'un grup funcional en la seva geometria, o amb combinacions de zones aromàtiques i alifàtiques. Conseqüentment és molt arriscat pretendre caracteritzar tots els efectes electrònics únicament a través de la constant  $\sigma$ . En canvi, l'estudi mitjançant *QS-SM* és específic del conjunt molecular estudiat, podent-se analitzar els efectes electrònics dels substituents en qualsevol fragment de la sèrie reactiva.

**Recta de regressió  $Z_{XX}^{NCS} / \sigma$ .** Un altre estudi presentat en l'article 7.1 fa referència als efectes electrònics dels substituents en una sèrie de compostos aromàtics derivats de l'isotiocianat, amb la fórmula genèrica:  $p\text{-X-C}_6\text{H}_4\text{-NCS}$ . És un exemple particularment interessant perquè les molècules estudiades presenten activitat antibacteriana i antifúngica sobre l'*Escherichia coli*. En un treball força antic,<sup>38</sup> s'havia demostrat que mitjançant l'equació de Hammett es podien obtenir relacions lineals entre l'activitat biològica i algunes propietats físico-químiques dels compostos analitzats. En la sèrie  $p\text{-X-C}_6\text{H}_4\text{-NCS}$ , el fragment involucrat en el procés químic és el grup NCS. En conseqüència, s'ha proposat la QS-SM del fragment NCS,  $Z_{XX}^{NCS}$ , en qualitat de descriptor teòric per mesurar l'efecte dels substituents. En l'article 7.1 es presenta la regressió lineal entre  $Z_{XX}^{NCS}$  i  $\sigma$ , que és equivalent a dir que la QS-SM està correlada amb la inhibició del creixement de l'*Escherichia coli*, perquè en l'article original [38] s'havia demostrat que existeix una relació entre  $\sigma$  i l'activitat biològica.

### Article 7.1

---

**Autors:** Robert Ponec, Lluís Amat, Ramon Carbó-Dorca.

**Títol:** *Molecular basis of quantitative structure-properties relationships (QSPR): A quantum similarity approach*

**Revista:** *Journal of Computer-Aided molecular Design*

**Volum:** 13      **Pàgines, inicial:** 259    **final:** 270    **Any:** 1999

---



## Molecular basis of quantitative structure-properties relationships (QSPR): A quantum similarity approach

Robert Ponec<sup>a</sup>, Lluís Amat<sup>b</sup> & Ramon Carbó-Dorca<sup>b,\*</sup>

<sup>a</sup>Institute of Chemical Process Fundamentals, Czech Academy of Sciences, Prague 6, Suchbátka 2, 165 02, Czech Republic; <sup>b</sup>Institute of Computational Chemistry, University of Girona, 17071 Girona, Catalonia, Spain

Received 2 July 1998; Accepted 3 September 1998

**Key words:** log P, molecular quantum similarity measures, quantitative structure-properties relationships (QSPR), substituent effect

### Summary

Since the dawn of quantitative structure-properties relationships (QSPR), empirical parameters related to structural, electronic and hydrophobic molecular properties have been used as molecular descriptors to determine such relationships. Among all these parameters, Hammett  $\sigma$  constants and the logarithm of the octanol-water partition coefficient, log P, have been massively employed in QSPR studies. In the present paper, a new molecular descriptor, based on quantum similarity measures (QSM), is proposed as a general substitute of these empirical parameters. This work continues previous analyses related to the use of QSM to QSPR, introducing molecular quantum self-similarity measures (MQS-SM) as a single working parameter in some cases. The use of MQS-SM as a molecular descriptor is first confirmed from the correlation with the aforementioned empirical parameters. The Hammett equation has been examined using MQS-SM for a series of substituted carboxylic acids. Then, for a series of aliphatic alcohols and acetic acid esters, log P values have been correlated with the self-similarity measure between density functions in water and octanol of a given molecule. And finally, some examples and applications of MQS-SM to determine QSAR are presented. In all studied cases MQS-SM appeared to be excellent molecular descriptors usable in general QSPR applications of chemical interest.

### Introduction

One of the main goals of natural sciences is the formulation of simple models and concepts in terms of which the observed phenomena can be classified, understood and, finally, described. The sophistication of these models depends on the complexity of the phenomena studied and, also, on the degree of theoretical development of a given science. In chemistry, one of the most complex fields resisting rigorous mathematical description is constituted by the relationships between the structure and (re)activity or properties of molecular structures, the so-called QSAR or QSPR. Due to the difficulties associated to QSPR analysis, an important place in the problem formulation still belongs to various empirical relationships. The well-known Hammett

or Taft equation [1–3] can serve as an example of this situation. Another important field, where empirical QSPR still play the dominant role, is the design of new materials or biologically active compounds [4–6]. Because of the immense practical importance of QSPR, considerable attention has been devoted to the elucidation of the factors responsible for the existence and validity of the empirical equations. In spite of the undeniable progress which these studies have brought to the solution of some particular problems (for instance, the theoretical rationalization of the Hammett  $\rho$  constants [7–9]), the understanding of the factors responsible for the existence of QSPR is still rather fragmentary. The main problem with these QSPR is that they cannot generally be derived from theoretically well-founded thermodynamic models, and this is why they are also called extra-thermodynamic relationships. Instead, they rely only on the intuitive

\*To whom correspondence should be addressed.

empirical idea that a similar structural change, induced in a given series of structurally related compounds by a systematic variation of substitution, will also have a similar effect on the properties or reactivity of the corresponding molecules.

Because of the great impact which this intuitively understood idea of similarity has had on chemical reasoning, it is not surprising that in recent years a lot of effort has been devoted to formulating the qualitative concept of similarity within a well defined theoretical basis. Especially promising in this respect seems to be the introduction of the quantitative similarity measures and indices based on quantum mechanical ideas [10–22]. In terms of this approach a number of qualitative chemical concepts could indeed be justified. The aim of the present work is to demonstrate that in addition to previously reported applications [23–25], quantum similarity measures (QSM) also provide an elegant and universal framework for the formulation of QSPR, whether in linear-free energy relationships (LFER) or in a more general QSAR form.

In pursuing this goal, a brief review of the basic ideas of QSM and their connections to QSPR will be presented. After introducing this general methodology, various examples to show some practical applications of the approach will be given. The provided examples involve the use of QSM as theoretical descriptors, replacing empirical parameters like the Hammett  $\sigma$  constants or octanol-water partition coefficients ( $\log P$ ), in QSPR of both chemical and pharmacological interest.

## Theoretical framework

### *Molecular quantum similarity measures (MQSM)*

The philosophy underlying the introduction of MQSM as molecular descriptors arises from the simple idea that properties of molecules are determined by their electronic structure. As a consequence, any similarity in molecular properties has to be reflected in the similarity of electronic structure. This suggests that if one wants to look for theoretical descriptors of molecular similarity it is natural to look for similarity in electron distribution, which, from a quantum mechanical point of view, contains all the information associated to a given microscopic system [26, 27]. MQSM represent an attempt to characterize quantitatively the similarity in electronic structure. On the other hand, the simplest descriptor characterizing the molecular electron structure is the first order electron density  $\rho(\mathbf{r})$ , so it is quite

natural to introduce the most easily computed MQSM just on this quantity. From the quantum mechanical point of view an MQSM involving two molecules  $A$  and  $B$  with associated first order densities  $\rho_A(\mathbf{r})$  and  $\rho_B(\mathbf{r})$ , may be defined by means of the following integral:

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

where  $\Omega(\mathbf{r}_1, \mathbf{r}_2)$  is a positive definite operator depending on the coordinates of two electrons. When both compared molecules are the same, the corresponding MQSM,  $Z_{AA}(\Omega)$  is denoted as a molecular quantum self-similarity measure (MQS-SM). In this work only the so-called overlap-like MQSM is used, which corresponds to employing the Dirac delta function  $\Omega(\mathbf{r}_1, \mathbf{r}_2) = \delta(\mathbf{r}_1 - \mathbf{r}_2)$  as a weighting operator. Then, in this case, Equation (1) simplifies to the form:

$$Z_{AB} = \int \rho_A(\mathbf{r}) \rho_B(\mathbf{r}) d\mathbf{r}. \quad (2)$$

Calculation of integral (2) can be very time consuming, since a large number of four-center integrals has to be computed within the common LCAO approximation. Moreover, the value defined by any MQSM is positionally dependent, so that optimization of mutual position of both molecules to yield the maximum measure is also required [28]. In order to reduce the above computational problems, a simplification was proposed some time ago [29, 30], known as the ASA approximation, which fits the electronic first order density functions to a linear expression composed of spherical functions. Using ASA, the MQSM computational costs are indeed substantially reduced, while still preserving a reasonable accuracy. Further reduction of computational costs can be achieved using the so-called *promolecular* approximation [30]. This approximation is based on a discrete representation of the molecular electron density function, defined as a convex superposition of the atomic electron densities in a given molecule  $A$ :

$$\rho_A^{ASA}(\mathbf{r}) = \sum_{a \in A} P_a \rho_a^{ASA}(\mathbf{r}). \quad (3)$$

In all the examples studied within this work, the coefficients  $P_a$  have been defined as the number of *valence* electrons minus the effective charge located on each atom  $a$ . In the *promolecular* form of Equation (3), every ASA atomic density function,  $\rho_a^{ASA}(\mathbf{r})$ , is

constructed using a linear combination of normalized 1S-type Gaussian functions centered on the  $a$ -th atom:

$$\rho_a^{ASA}(\mathbf{r}) = \sum_{i \in a} w_i |S_i(\mathbf{r} - \mathbf{R}_a; \zeta_i)|^2 \quad (4)$$

where the coefficients  $w_i$  are forced to fulfill the convex constraints:

$$\{w_i > 0, \forall i\} \wedge \left\{ \sum_i w_i = 1 \right\}, \quad (5)$$

to preserve the statistical distribution probability meaning of the approximate density function [31]. In order to keep Equation (5) conditions, the convex coefficient set  $w_i$  has been optimized using an elementary Jacobi rotation technique [30]. The 1S-GTO exponents  $\{\zeta_i\}$  have also been optimized simultaneously using a Newton method. Both sets of parameters have been fitted to a density function obtained using a 3–21 G basis sets for atoms H through Kr. Optimal coefficients and exponents can be downloaded from a WWW site [32]. Furthermore, the *promolecular* densities used in this work have been constructed using the following rule: one function for H, three functions for C, N, O, and F, four functions for S and Cl and five functions for Br.

Using the ASA formalism described above, the overlap-like MQSM between two molecules, as in Equation (2), is reduced to the final simple quadratic form:

$$Z_{AB} = \sum_{a \in A} \sum_{b \in B} P_a P_b Z_{ab}, \quad (6)$$

where  $P_a$  and  $P_b$  correspond to the same corrected atomic charges defined in the same way as in Equation (3), and the atomic QSM contributions  $Z_{ab}$  are calculated as the integrals:

$$Z_{AB} = \sum_{i \in a} \sum_{j \in b} w_i w_j \int |S_i(\mathbf{r} - \mathbf{R}_a)|^2 |S_j(\mathbf{r} - \mathbf{R}_b)|^2 d\mathbf{r}, \quad (7)$$

which correspond to a simple and well-known overlap between two 1S-GTO functions [33].

### MQSM and QSPR

Once the theoretical concept of MQSM and their computational implementations have been set, one can proceed to the application of the above abstract

concepts to practical chemical problems. The most appealing is the recently proposed formulation of the general theory, where the MQSMs open the way for a plausible theoretical foundation of QSPR. Since the detailed and strict mathematical development of the procedure can be found in a previous study [23], only a brief resume will be given.

Suppose known a molecular set  $\mathbf{M}=\{m_I\}$  formed by  $n$  molecules, and an attached set of density functions  $\mathbf{D}=\{\rho_I\}$  in one-to-one correspondence with  $\mathbf{M}$ . Now, having chosen a basis set of density functions, suppose computed an  $(n \times n)$  similarity matrix  $\mathbf{Z}=\{Z_{IJ}\}$ , whose elements correspond to some MQSM involving all pair of functions in  $\mathbf{D}$ . This matrix can also be viewed as a row vector whose elements are their columns:  $\mathbf{Z}=\{\mathbf{z}_I\}$ . Every column  $\mathbf{z}_I$  is in one-to-one correspondence with the functions of  $\mathbf{D}$  and hence of  $\mathbf{M}$ . The set  $\mathbf{Z}=\{\mathbf{z}_I\}$  can be interpreted as an  $n$ -dimensional discrete representation of the functions of  $\mathbf{D}$ . Since, according to quantum theory, the density functions can be regarded as the source of all the molecular information [26, 27], then any observable molecular property  $\pi_I$ , associated to some hermitian operator  $\Omega(\mathbf{r})$  and attached to molecule  $m_I$  may be calculated using:

$$\pi_I = \int \Omega(\mathbf{r}) \rho_I(\mathbf{r}) d\mathbf{r}, \quad (8)$$

which may be interpreted as a scalar product, and can be formally written as:

$$\pi_I = \langle \Omega | \rho_I \rangle. \quad (9)$$

If it is taken into account that an  $n$ -dimensional discrete representation of the density function,  $\rho_I$ , is known in terms of the columns of the similarity matrix  $\mathbf{Z}=\{\mathbf{z}_I\}$ , the above continuous representation (8) of  $\pi_I$  can be rewritten as:

$$\pi_I \approx \mathbf{a}^\top \mathbf{z}_I \quad (10)$$

where  $\mathbf{a}$  is an  $n$ -dimensional vector, representing the operator  $\Omega(\mathbf{r})$  in the  $n$ -dimensional discrete space of MQSM vectors  $\mathbf{z}_I$ . The operator-vector  $\mathbf{a}$  is generally unknown, but if the pairs  $\{\pi_I, \mathbf{z}_I\}$  are well defined, the elements of  $\mathbf{a}$  can be obtained by any least-squares technique. This computational procedure constitutes in this way a theoretical basis of QSPR. While Equation (10) represents a general theoretical form of QSPR, in some cases it is also useful to convert it to another, simpler form, which, could be closely related to the well-known LFER. In order to introduce this



simplification, Equation (10) can be rewritten in the form below, in which the self-similarity part, corresponding to the diagonal of the similarity matrix, is separated from the rest:

$$\pi_I = a_I Z_{II} + \sum_{J \neq I} a_J Z_{JI}. \quad (11)$$

Now, if the last term of the right hand side of Equation (11) is denoted as  $b_I$  and if it is further possible to admit that in a given series of molecules the terms  $\{a_I, b_I\}$  could be roughly constant, Equation (11) transforms into:

$$\pi_I \approx a Z_{II} + b, \quad (12)$$

which expresses a simple linear correlation between the property  $\pi_I$  of molecule  $m_I$  and the MQS-SM:  $Z_{II}$ , of the same molecule. Such a form of relationship is frequently found empirically. For example, within the Hammett equation, where the molecular properties (like pK or log  $k$ ) are correlated with the substituent constants  $\sigma$  in a series of substituted compounds. The following section will present several examples of the use of self-similarity measures as descriptors of substituent effects.

## Results and discussion

### *Quantum self-similarity measures as descriptors of the substituent effect*

In order to show the usefulness of MQSM as molecular descriptors, the applications of these measures to the description of substituent effect is reported as a first example. This field is traditionally represented by the broad class of the so-called LFER as, for example the Hammett and Taft equation. The intention of the present study is to show that appropriately selected MQS-SM do indeed describe the variation of substituent effect in a given reaction series, so that they can be regarded as equivalent to the usual substituent constants. In order to shed light on this equivalence, a series of dissociation equilibrium of several substituted carboxylic acids have been analyzed. As will be shown below, good correlations of MQS-SMs with Hammett  $\sigma$  constants, which can be regarded as a theoretical counterpart of empirical pK vs  $\sigma$  correlations, have been observed. Even if in principle the MQS-SM corresponding to whole molecules can be used in the correlations with  $\sigma$ , it has often been found useful

not to characterize the whole molecule by the corresponding  $Z_{II}$  values, but only with certain fragments, which can be identified, in turn, with the molecular part actively participating in a given process: the reaction center. The philosophy underlying this replacement is based on the well-known fact that the reaction center is usually the part of the molecule which is the most strongly involved in the process. In this sense, neglecting any contaminating interactions with the remaining part of the molecule may result in the increased sensitivity of the corresponding descriptor to external effects. Such a specific limitation to a particular molecular fragment is, of course, possible only when the reaction center can unambiguously be determined. In the present case, where the substituent effect on the dissociation equilibrium of carboxylic acids has been analyzed, the active part of the molecule can clearly be identified with the COOH group. As a consequence, MQS-SM  $Z_{II}$ , calculated for the active COOH fragment, should be the appropriate descriptor in the present framework. In this case the theoretical counterpart of the usual Hammett pK vs  $\sigma$  correlation should be written as:

$$\text{pK}_X^R = a_R Z_{XX,R}^{\text{COOH}} + b_R, \quad (13)$$

or, using the equivalent form:

$$Z_{XX,R}^{\text{COOH}} = \alpha_R \sigma_X + \beta_R. \quad (14)$$

In these equations the index X runs over the set of substituents and R denotes the particular reaction series. In the present case, the analyzed reaction is constituted by the classical series of p-substituted benzoic acids (**I**), 5-substituted thiophen (**II**) and furan 2-carboxylic acids (**III**), p-substituted trans-cinnamic (**IV**) and phenylacetic (**V**) acids (see Scheme 1) with substituents involving both electron donor and electron acceptor groups. The calculations for all molecules have been performed using the semiempirical AM1 [34] method included within the MOPAC [35] package. In all cases, the structures of substituted acids have been completely optimized and the resulting geometries and charge distributions have been used as the input data for the subsequent calculations of MQS-SM. The observed correlations between the calculated MQS-SM for the COOH group and the Hammett substituent constants (Equation (14)) are depicted in Figures 1–5 and the resulting statistical parameters are summarized in Table 1.

The correlations are in all cases quite satisfactory, so that the assumption that MQS-SMs represent good

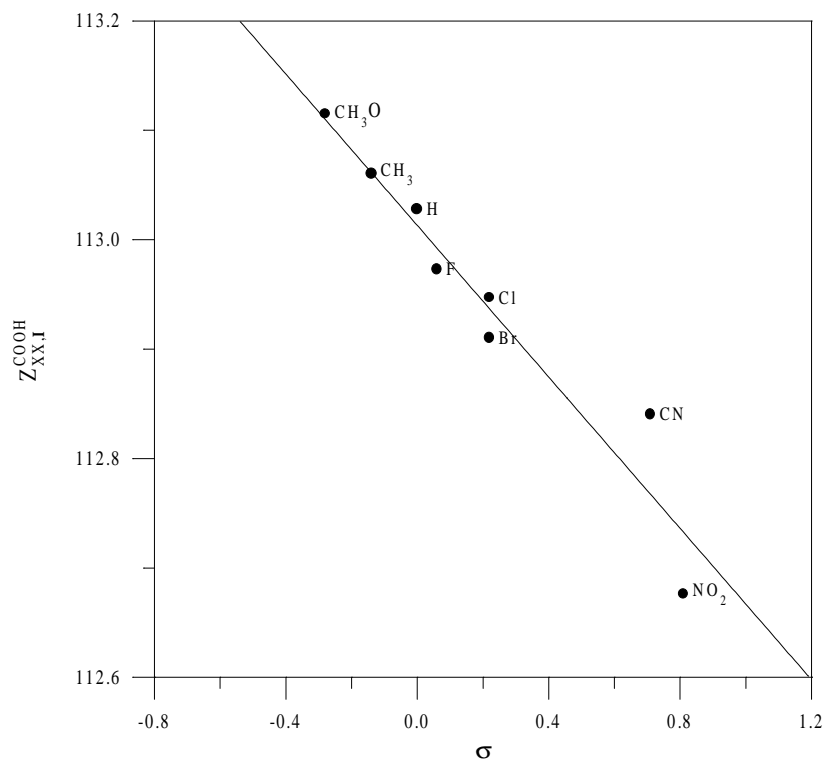


Figure 1. Dependence of calculated MQS-SM  $Z_{XX,I}^{COOH}$  for a series of substituted benzoic acids on the Hammett  $\sigma$  constants.

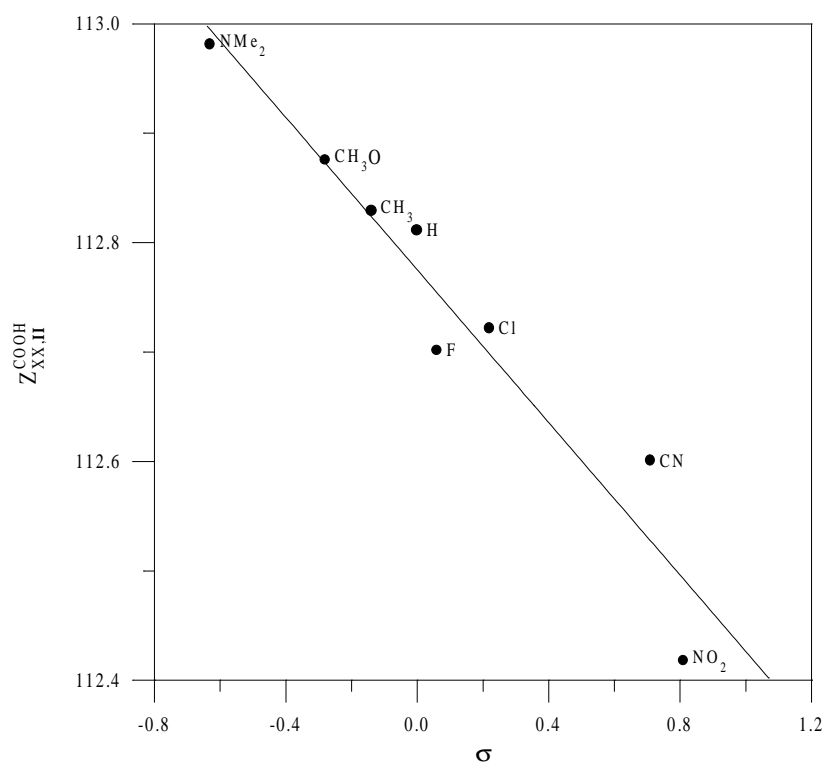


Figure 2. Dependence of calculated MQS-SM  $Z_{XX,II}^{COOH}$  for a series of substituted thiophen 2-carboxylic acids on the Hammett  $\sigma$  constants.

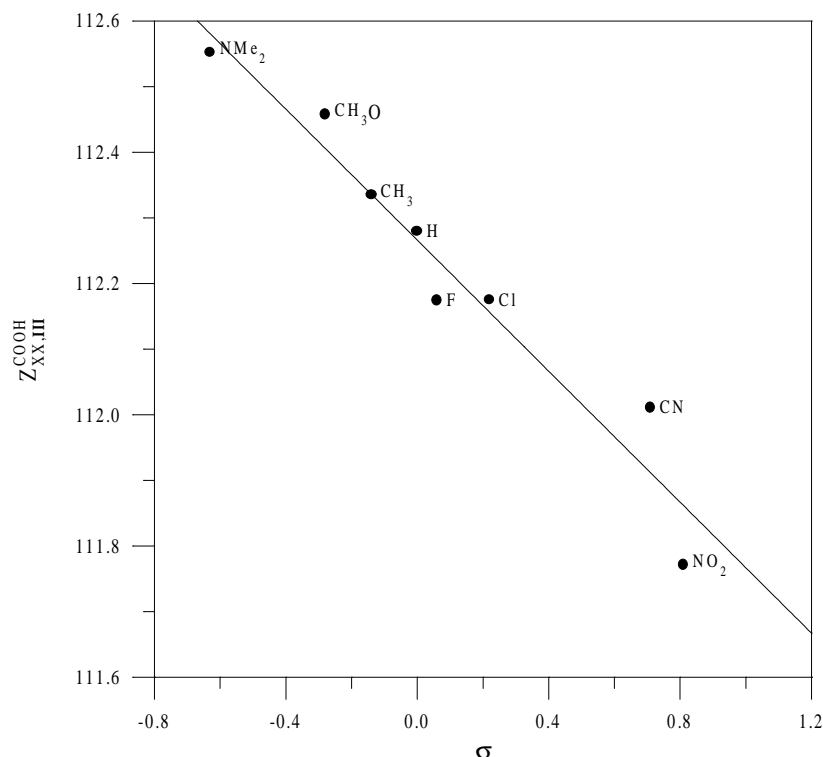


Figure 3. Dependence of calculated MQS-SM  $Z_{XX,III}^{COOH}$  for a series of substituted furan 2-carboxylic acids on the Hammett  $\sigma$  constants.

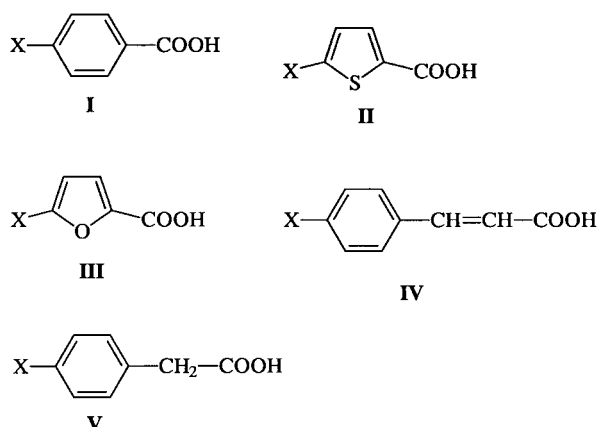
Table 1. Calculated statistical parameters of the correlations of MQS-SMs against Hammett substituent constants

Reaction series	$\alpha_R$	$\beta_R$	$n^a$	$r^b$
<b>I</b>	-0.34	113.01	8	0.969
<b>II</b>	-0.37	112.77	8	0.966
<b>III</b>	-0.50	112.27	8	0.974
<b>IV</b>	-0.22	113.15	7	0.967
<b>V</b>	-0.14	113.17	8	0.961

<sup>a</sup> Number of molecules.

<sup>b</sup> Regression coefficient.

theoretical descriptors of substituent effects is indeed justified. In addition to this primary result there are also some other conclusions which can be deduced from the observed correlations. The most interesting of them all, corresponds to the possibility of using the slopes of the reported theoretical relationships for the estimation of the relative sensitivity of a given skeleton to the transmission of the substituent effect, measured by the Hammett  $\rho$  constant. Thus, if the slope  $\alpha_o$  of the correlation (14) for the substituted benzoic acids is taken as an arbitrary unit (corresponding to  $\rho=1$



Scheme 1. Molecular structures for benzoic acid (**I**), thiophen 2-carboxylic acid (**II**), furan 2-carboxylic acid (**III**), trans-cinnamic acid (**IV**) and phenylacetic acid (**V**).

for this process), then the ratios  $\alpha_R/\alpha_o$  of the corresponding slopes for the remaining reaction series can be expected, if everything is correct, to characterize the relative sensitivity of other skeletons to the substituent effect. That is: it will be connected with the experimental  $\rho$  constants. As can be seen in Table 2,

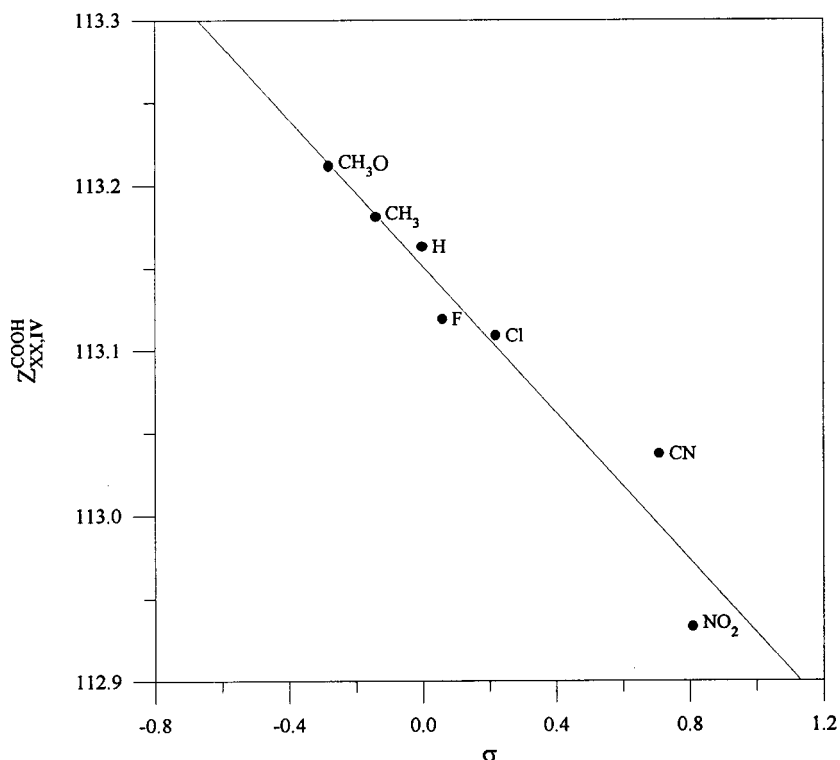


Figure 4. Dependence of calculated MQS-SM  $Z_{XX,IV}^{COOH}$  for a series of substituted trans-cinnamic acids on the Hammett  $\sigma$  constants.

where these ratios are summarized, the agreement between the experimental and 'theoretical'  $\rho$  constants is good. This suggests that the above reported linear relationships can be used advantageously as a means of calculating the experimental  $\rho$  constants. Here it is fair to say that similar estimations had been already reported on the basis of perturbation theory some time ago [7–9], but such calculations have been restricted only to aromatic conjugated systems. The above reported approach is, however, free of any limitation and can be generally used.

Another interesting and nontrivial conclusion, resulting from the above similarity approach, is represented by the observed correlations of  $pK_a$  of nonsubstituted acids (I–V) with the corresponding MQS-SM  $Z_{HH,R}^{COOH}$ . An example which surpasses the range of previously reported correlations is depicted in Figure 6. This is an especially interesting case, since it represents the example of the general relation (12) for which there is no known LFER counterpart.

In connection with the above reported relationships it is, of course, necessary to mention another important aspect. This aspect concerns the fact that while the experimental  $\rho$  constants have been obtained in wa-

Table 2. Comparison of experimental and calculated  $\rho$  constants for the dissociation of carboxylic acids I–V in water

Reaction series	$\rho_{exp}$	$\alpha_R/\alpha_o$
<b>I</b>	1	1.00
<b>II</b>	1.13 <sup>a</sup> –1.20 <sup>b</sup>	1.09
<b>III</b>	1.40	1.47
<b>IV</b>	0.46	0.64
<b>V</b>	0.56	0.42

<sup>a</sup> From ref. 36.

<sup>b</sup> From ref. 37.

ter solution, the quantum chemical descriptors used correspond to the gas phase, so that in correlations like Equation (14), the data from two different phases are in fact compared. This, of course, is not theoretically perfect, but since the experimental data have in all cases been obtained under the same conditions (water solutions at 25 °C) and, moreover, the studied systems form a series of closely structurally related molecules, it is possible to expect that the comparison of gas phase and solution data will only be affected by a systematic shift. The relative comparisons between

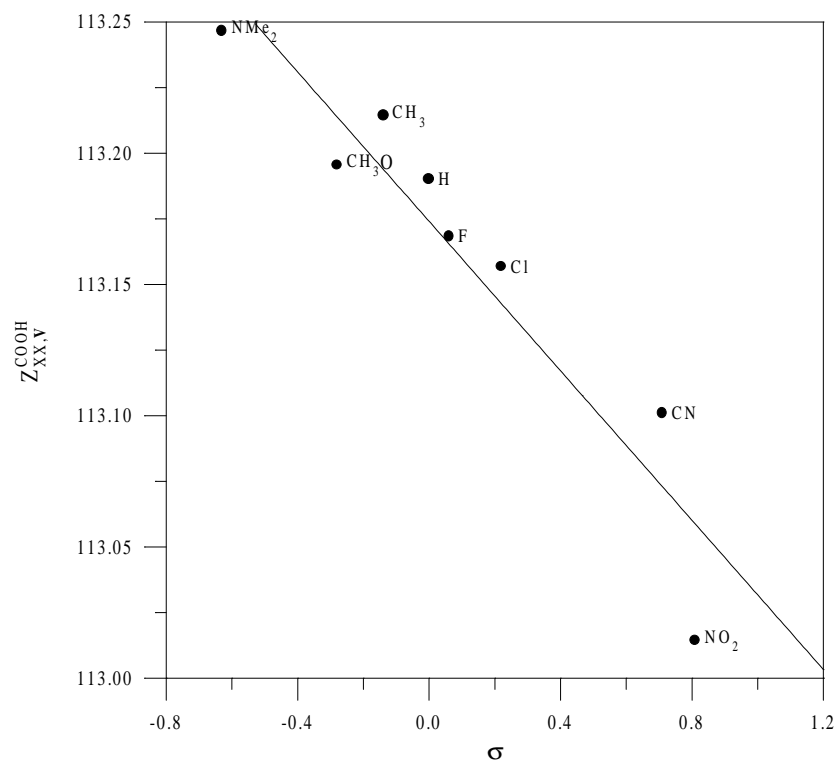


Figure 5. Dependence of calculated MQS-SM  $Z_{XX,V}^{COOH}$  for a series of substituted phenylacetic acids on the Hammett  $\sigma$  constants.

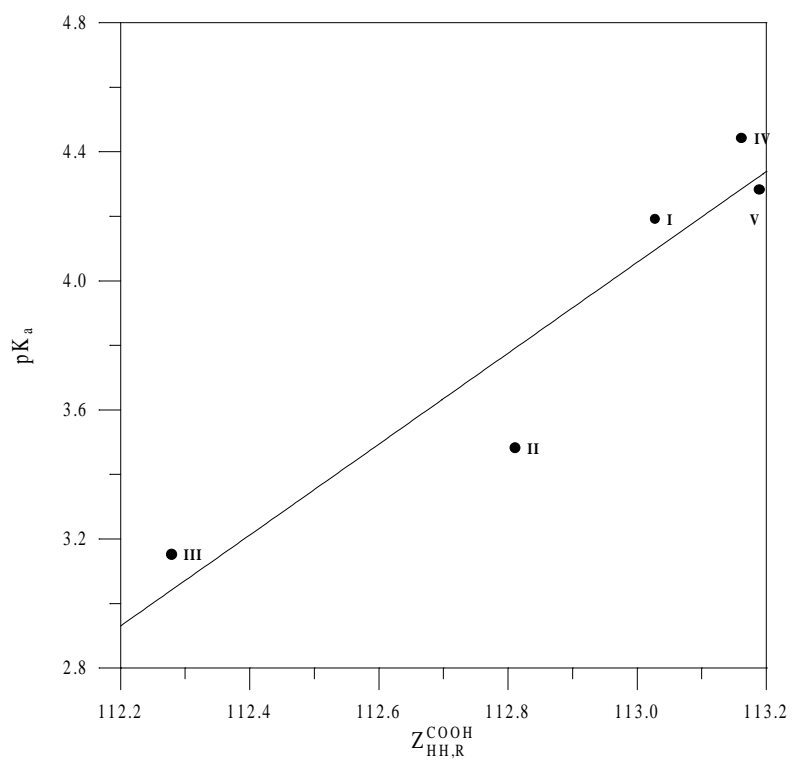


Figure 6. Dependence of calculated MQS-SM  $Z_{HH,R}^{COOH}$  for the nonsubstituted acids (I-V) on the  $pK_a$  constants.

the different series in this context may still remain meaningful. There is, however, no problem to extract the necessary theoretical descriptors from the quantum chemical calculations including the solvent effect [39–41], so that empirical correlations like Equation (14) can be made more realistic. An example of correlations where the direct inclusion of the solvent effect is of crucial importance is represented here by the theoretical prediction of the partition coefficient between water and octanol, which is frequently used as a descriptor in QSAR calculations. In the following part an example of the procedure allowing the theoretical calculation of  $\log P$  in a series of structurally related skeletons will be reported.

#### *Self-similarity measures as descriptors of $\log P$*

Partition coefficients between water and octanol are empirical parameters which have long been used as a molecular descriptor of molecular hydrophobicity. Because of this descriptive importance, various empirical additivity schemes allowing its calculation have been proposed [6]. The present work pretends to show that a new theoretical scheme, producing an excellent prediction of  $\log P$  in a series of structurally related molecules, can be formulated using MQS-SMs. This procedure is based on the direct inclusion of solvents (water and octanol in this case) into the quantum chemical calculations of electronic distribution in a molecule. A series of primary and secondary aliphatic alcohols and acetic acid esters have been used in order to test this approach. The calculations have been performed at ab initio HF level of theory using a 3-21 G\* basis set within the Gaussian 94 program [38]. After determining the optimized gas phase structures for each molecule, the solvent effect was introduced using the polarized continuum model (PCM) [40, 41] incorporated in the mentioned program code. Within this approach, the molecules are placed in a cavity surrounded by a medium where the dielectric constant  $\epsilon$ , the mutual polarization of this medium and the solute molecule are taken into account. Here, the values of dielectric constants  $\epsilon=80.4$  and  $\epsilon=10.3$  have been used for water and octanol respectively. Based on the calculations with included solvents, two *promolecular* ASA densities characterizing the modification of electron distributions in water  $\rho_{A^w}^{ASA}(\mathbf{r})$  and octanol  $\rho_{A^o}^{ASA}(\mathbf{r})$  have been determined. The theoretical descriptors have been constructed as an overlap-based self-similarity measure between ASA densities in water and octanol for each individual molecule. The

Table 3. Comparison of calculated MQS-SM with experimental  $\log P$  values for a series of selected molecules

Molecule	$\log P^a$	$Z_{A^w A^o}^b$
Methanol	-0.77	73.69
Ethanol	-0.31	91.23
1-Propanol	0.25	108.49
1-Butanol	0.88	125.50
1-Pentanol	1.56	142.30
1-Hexanol	2.03	159.68
2-Propanol	0.05	108.30
2-Butanol	0.61	125.60
2-Pentanol	1.19	142.79
3-Pentanol	1.21	142.51
2-Hexanol	1.76	159.48
3-Hexanol	1.65	159.49
Acetic acid (AcH)	-0.17	141.36
AcH methyl ester	0.18	157.71
AcH ethyl ester	0.73	175.18
AcH propyl ester	1.24	192.41
AcH butyl ester	1.78	209.31

<sup>a</sup> From Hansch et al. [42].

<sup>b</sup> See Results and discussion.

comparison of calculated MQS-SMs and experimental  $\log P$  values is summarized in Table 3 and presented graphically in Figure 7. As can be observed, the correlation splits into two separate lines for alcohols and esters, and a small systematic shift is in fact also observed between primary and secondary alcohols. As a result, within each class of molecules the correlation is indeed excellent. This situation is very interesting since it opens the possibility of using the above similarity approach as a new theoretical scheme for the calculation of  $\log P$  in classes of structurally related molecules. As a consequence, the calculated MQS-SMs can directly be used in QSAR instead of  $\log P$  themselves. An example of such a use for the correlation with biological data is presented below.

#### *MQS-SM and correlations with biological data*

The correlation of biological data with various molecular descriptors certainly constitutes an important and widely used field of the QSAR application. Because of the practical importance of such correlations for the rational drug design and recent work related to MQSM applied to QSAR [24, 25], it seems interesting to check whether the MQS-SMs could be of any help in this effort. Generally it is possible to expect that because of greater complexity of factors responsible for the biological activity, the finding of the

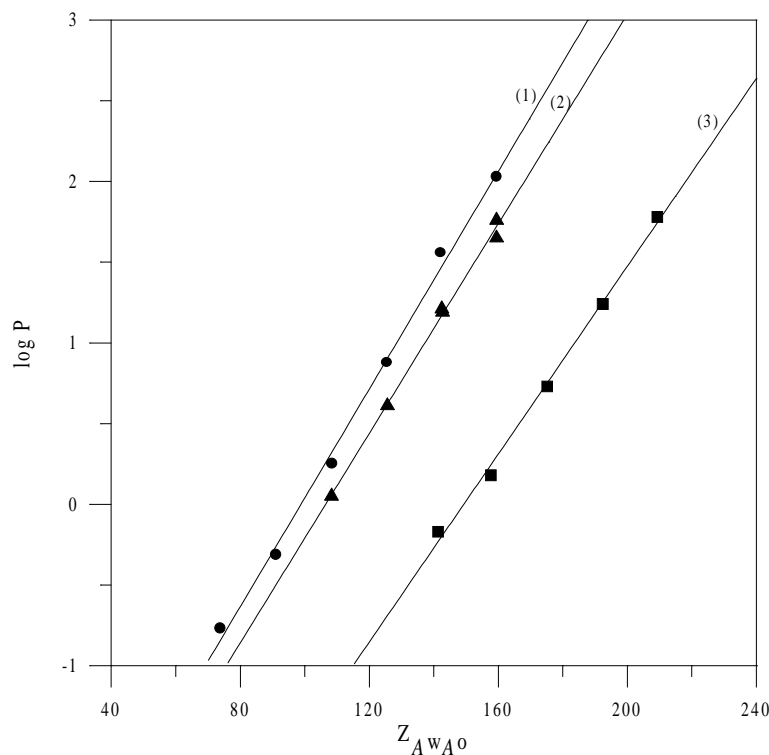


Figure 7. Dependence of calculated MQS-SM  $Z_{A^w A^o}$  on  $\log P$  for a series of primary (1) and secondary (2) aliphatic alcohols, and acetic acid esters (3).

appropriate molecular descriptor will be much more difficult and multiparameter correlations are often observed to be necessary for QSAR calculations. Among the descriptors which are used in such multiparameter correlations, the Hammett  $\sigma$  constants and  $\log P$  are usually found. The previous demonstration that MQS-SM can successfully be used as descriptors of both the substituent effect and  $\log P$ , has led to attempting the application of these new theoretical descriptors in QSAR as well. As a first example, the reported study of antibacterial and antifungal activity of substituted phenyl-isothiocyanates on *Escherichia coli* has been chosen. For this system the correlation of  $\log ED_{50}$  with Hammett  $\sigma$  constants was early reported [43]. This has been interpreted as indicating that the toxicity of these molecules is apparently due to the reaction of the active compound with the target in the cells. A plausible candidate for such a process is the nucleophilic addition of the reactive groups in the protein to the isothiocyanate group [43]. The quantum molecular similarity analysis of this process can start from the fact that the carrier of the biological activity is in this case the isothiocyanate (NCS) group. In view

of what has been previously discussed, this suggests that the appropriate theoretical descriptor could be the self-similarity measure  $Z_{XX}^{NCS}$  calculated in a series of substituted phenyl-isothiocyanates just for the active NCS group. The quantum chemical calculations, serving to generate the necessary density matrices, have been again performed by the semiempirical AM1 method and the resulting correlation of the MQS-SM with the Hammett substituent constants is depicted in Figure 8. As can be seen, the correlation is again very satisfactory so that the ability of the MQS-SMs to act as descriptors of the substituent effect is also confirmed in this case.

Having demonstrated the ability of MQS-SMs to act as descriptors of the substituent effect, in order to complete this study, one can try to find an example of the MQS-SMs application as a descriptor of  $\log P$  in biological QSAR. The next example is related to the reported data on the narcosis of tadpoles [44], which has been shown to be determined primarily by the  $\log P$  [6]. The class of the studied molecules included the series of aliphatic alcohols and acetic acid

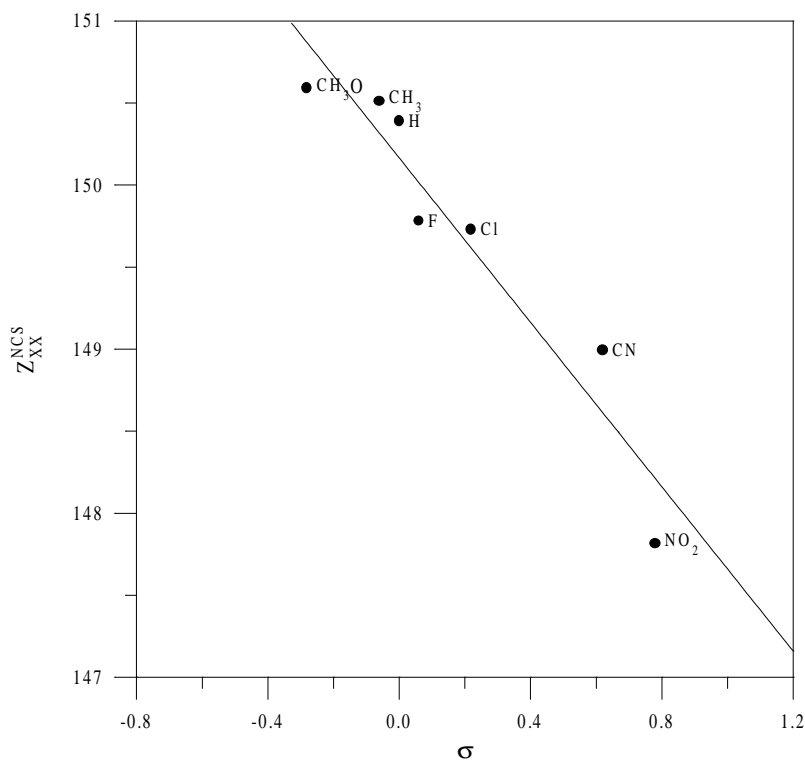


Figure 8. Dependence of calculated MQS-SM  $Z_{XX}^{NCS}$  for a series of substituted phenyl-isothiocyanate acids on the Hammett  $\sigma$  constants.

esters which have also been studied in the correlation of  $\log P$  values. In this way, the previously calculated MQS-SMs can also be re-used for the correlation with the biological data on narcosis. A comparison of experimental and calculated biological activities is summarized in Table 4. It is possible to observe in this case how the agreement between the experimental activities and MQS-SMs as theoretical descriptors of  $\log P$  again appears to be excellent.

## Conclusions

The present work complements previous studies of MQSM applied to determine QSPR, and confirms that this procedure appears as a useful tool for this purpose. The results in this paper encourage the consideration that in appropriate cases MQS-SM are able to provide an ab initio model to determine sound QSPR, as well as a plausible quantum mechanical justification background for such structure-properties linear relationships. In addition, the substitution of Hammett  $\sigma$  constants or  $\log P$  values by MQS-SM, due to the universal computational structure of these new molecular

Table 4. Comparison of experimental and calculated biological activities of aliphatic alcohols and acetic acid esters for the narcosis of tadpoles

Molecule	log 1/C	
	Obs <sup>a</sup>	Calc <sup>b</sup>
Methanol	0.24	0.19
Ethanol	0.54	0.58
1-Propanol	0.96	0.97
1-Butanol	1.42	1.35
2-Propanol	0.89	0.96
AcH methyl ester	1.10	1.11
AcH ethyl ester	1.52	1.52
AcH propyl ester	1.96	1.93
AcH butyl ester	2.30	2.32

<sup>a</sup> From ref. 44.

<sup>b</sup> Calculated using the equations:  $\log 1/C = 0.0225 Z_{AA^0}^{wA^0} - 1.4713$  for aliphatic alcohols, and  $\log 1/C = 0.0235 Z_{AA^0}^{wA^0} - 2.5942$  for acetic acid esters.

descriptors, opens broad horizons in QSPR studies, and permits other kinds of relationships to be attained, just as the presented correlation example related to  $pK_a$  shows. Also, according to the obtained calculation pattern, significant correlations may be estimated



using MQS-SM for a vast number of biological series of compounds.

### Acknowledgements

This work has been performed during the stay of one of us (R. Ponec) at the University of Girona, which has been made possible by a NATO grant. Thanks are also due to the Centre de Supercomputació de Catalunya (CESCA) and the Centre Europeu de Paral·lelisme de Barcelona (CEPBA) for a generous amount of computation time. This research has been partially supported by a CICYT grant: SAF 96-0158 and by the Fundació Maria Francisca de Roviralta. The authors gratefully acknowledge all these sources of support.

### References

- Hammett, L. P., *Trans. Faraday Soc.*, 34 (1938) 96.
- Taft, R. W., *J. Am. Chem. Soc.*, 75 (1953) 4231.
- Shorter, J., *Chem. Brit.*, 5 (1969) 269.
- Kubinyi, H. (Ed.) *3D QSAR in Drug Design: Theory Methods and Applications*, ESCOM, Leiden, 1993.
- Dean, P. M. (Ed.) *Molecular Similarity in Drug Design*, Blackie Academic & Professional, London, 1995.
- Hansch, C. and Leo, A., *Exploring QSAR. Fundamentals and Applications in Chemistry and Biology*, ACS Professional Reference Book, Washington, DC, 1995.
- Ponec, R. and Chvalovský, V., *Collect. Czech. Chem. Commun.*, 39 (1974) 3091.
- Ponec, R., *Collect. Czech. Chem. Commun.*, 45 (1980) 1646.
- Krygowski, T. M. and Perjessy, A., *Bull. Acad. Sci. Polon.*, 22 (1974) 437.
- Carbó, R., Leyda, L. and Arnau, M., *Int. J. Quantum Chem.*, 17 (1980) 1185.
- Carbó, R. (Ed.) *Molecular Similarity and Reactivity: From Quantum Chemical to Phenomenological Approaches*, Kluwer, Amsterdam, 1995.
- Carbó-Dorca, R. and Mezey, P. G. (Eds.) *Advances in Molecular Similarity*, JAI Press Inc., Greenwich, CT, 1996, Vol. 1.
- Cooper, D. L. and Allan, N. L., *J. Comput.-Aided Mol. Design*, 3 (1989) 253.
- Cooper, D. L. and Allan, N. L., *J. Am. Chem. Soc.*, 114 (1992) 4773.
- Cioslowski, J. and Fleischmann, E. D., *J. Am. Chem. Soc.*, 113 (1991) 64.
- Cioslowski, J. and Nanayakkara, A., *J. Am. Chem. Soc.*, 115 (1993) 11213.
- Burt, C., Richards, W. G. and Huxley, P., *J. Comput. Chem.*, 10 (1990) 1139.
- Good, A. C., So, S. S. and Richards, W. G., *J. Med. Chem.*, 36 (1993) 433.
- Ponec, R. and Strnad, M., *Int. J. Quantum Chem.*, 42 (1992) 501.
- Ponec, R., *J. Chem. Inf. Comput. Sci.*, 33 (1993) 805.
- Mezey, P. G., *J. Chem. Inf. Comput. Sci.*, 32 (1992) 650.
- Luo, X. and Mezey, P. G., *Int. J. Quantum Chem.*, 41 (1992) 557.
- Carbó, R., Besalú, E., Amat, L. and Fradera, X., *J. Math. Chem.*, 18 (1995) 237.
- Fradera, X., Amat, L., Besalú, E. and Carbó-Dorca, R., *Quant. Struct.-Act. Relat.*, 16 (1997) 25.
- Lobato, M., Amat, L., Besalú, E. and Carbó-Dorca, R., *Quant. Struct.-Act. Relat.*, 16 (1997) 465.
- von Neumann, J., *Mathematical Foundations of Quantum Mechanics*, Princeton University Press, Princeton, NJ, 1955.
- Bohm, D., *Quantum Theory*, Dover Pub. Inc., New York, NY, 1989.
- Constans, P., Amat, L. and Carbó-Dorca, R., *J. Comput. Chem.*, 18 (1997) 826.
- Constans, P. and Carbó, R., *J. Chem. Inf. Comput. Sci.*, 35 (1995) 1046.
- Amat, L. and Carbó-Dorca, R., *J. Comput. Chem.*, 18 (1997) 2023.
- Carbó-Dorca, R., *J. Mat. Chem.*, 22 (1997) 143; 23 (1998) 353; 23 (1998) 365.
- ASA coefficients and exponents can be seen and downloaded from the WWW site: <http://iqc.udg.es/cat/similarity/ASA/funcset.html>
- See for example: Saunders, V. R., *Computational Techniques in Quantum Chemistry and Molecular Physics*, D. Reidel Publ. Co., Dordrecht, Holland, 1975, pp. 347-424.
- Dewar, M. J. S., Zebisch, E. G., Healy, E. F. and Stewart, J. J. P., *J. Am. Chem. Soc.*, 107 (1985) 3902.
- Stewart, J. J. P. MOPAC6, QCPE 455, Indiana University, Bloomington, IN, 1993.
- Exner, O. and Simon, W., *Collect. Czech. Chem. Commun.*, 29 (1964) 2016.
- Fringuelli, F., Mario, G. and Taticchi, A., *J. Chem. Soc. Perkin Trans. II* (1972) 1738.
- Frisch, M. J., Trucks, G. W., Schlegel, H. B., Gill, P. M. W., Johnson, B. G., Robb, M. A., Cheeseman, J. R., Keith, T. A., 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, H. S., Defrees, D. J., Baker, H., Stewart, J. J. P., Head-Gordon, M., Gonzalez, C. and Pople, J. A., *GAUSSIAN 94, Revision A.1*, Gaussian, Inc.: Pittsburgh, PA, 1995.
- Kirkwood, J. G., *J. Chem. Phys.*, 2 (1934) 351.
- Miertus, S., Scrocco, E. and Tomasi, J., *Chem. Phys.*, 55 (1981) 117.
- Miertus, S. and Tomasi, J., *Chem. Phys.*, 65 (1982) 239.
- Hansch, C., Leo, A. and Hoekman, D., *Exploring QSAR. Hydrophobic, Electronic, and Steric Constants*, ACS Professional Reference Book, Washington, DC, 1995.
- Vlachová, D. and Drobnička, L., *Collect. Czech. Chem. Commun.*, 31 (1966) 997.
- Lipnick, R.L. In Suter II, G. W. and Lewis, M. A. (Eds.) *Aquatic Toxicology and Environmental Fate*, vol. II, ASTM, 1989, p. 468.

### 7.3.2 Correlacions entre els valors esperats de l'espai de moment i la constant $\mathbf{s}$

Dins el marc de col·laboració amb els investigadors N.L. Allan i D. L. Cooper, s'ha realitzat un estudi complementari a la deducció de les rectes de regressió  $Z_{XX,R}^{\text{COOH}}/\sigma$ . En l'article [39] s'analitzen les correlacions obtingudes entre el valor esperat  $\langle p^2 \rangle$ , definit en l'equació (2.18), i la constant  $\sigma$  per les 5 sèries d'àcids analitzades en l'article 7.1. En total s'han avaluat 12 substituents per sèrie, que coincideixen amb els llistats en la taula 7.2. També s'ha repetit els càlculs d'optimització de la geometria AM1 de tots els compostos. Malgrat millorar la descripció de les geometries moleculars, encara hi ha algun compost que no convergeix en un mínim d'energia, com és el substituent  $\text{NO}_2$  en la sèrie **II**, i els substituents  $\text{NO}_2$  i  $\text{N}(\text{CH}_3)_2$  en la sèrie **III**. Els valors de  $\langle p^2 \rangle$  s'han avaluat pel fragment  $\text{COOH}$  a partir de les equacions (2.16) i (2.18) emprant únicament funcions de base centrades en àtoms d'aquest grup. Com es pot observar en la taula 7.3, els coeficients de correlació són superiors a 0.95 en totes les reaccions, i molt similars als que s'obtenen amb les rectes de regressió  $Z_{XX,R}^{\text{COOH}}/\sigma$ . Per tant es pot concloure que les mesures  $\langle p^2 \rangle$  són uns bons descriptors teòrics dels efectes del substituent.

Sèrie	<b>I</b>	<b>II</b>	<b>III</b>	<b>IV</b>	<b>V</b>
$n$	12	11	10	12	12
$r^a$	0.976	0.963	0.956	0.978	0.976
$r^b$	0.979	0.965	0.957	0.979	0.979

**Taula 7.3** Coeficients de correlació de les rectes de regressió (a)  $\langle p^2 \rangle/\sigma$  [39] i (b)  $Z_{XX,R}^{\text{COOH}}/\sigma$  per les cinc sèries d'àcids.

## 7.4 Aplicacions *QSAR*

Havent demostrat que determinades *QS-SM* poden descriure els mateixos efectes que certs descriptors físico-químics clàssics, tal com  $\log P$  i la constant  $\sigma$  de Hammett, el següent estadi ha estat desenvolupar una aproximació *QSAR* basada en la semblança de fragments moleculars. Amb l'objectiu de validar la nova metodologia, el primer treball que es presenta consisteix en la reproducció d'alguns models de *QSAR* clàssica on hi intervenen els paràmetres  $\log P$  i  $\sigma$ . L'estudi se centre en la recerca del descriptor mecanicoquàntic apropiat que pot reemplaçar els paràmetres empírics emprats en la generació de les equacions *MLR* clàssiques.

La utilització de descriptors moleculars definits sobre fragments es fonamenta en part en el teorema hologràfic de la densitat electrònica,<sup>40</sup> segons el qual tota la informació continguda en la densitat electrònica global d'una molècula es troba també inclosa en la densitat local de qualsevol fragment amb volum no zero de la mateixa molècula. Basant-se en aquest teorema s'han proposat les *QS-SM* com una eina útil per extreure informació de la densitat electrònica de fragments. L'objectiu és escollir el fragment de la densitat electrònica amb volum no zero, on la predeterminada propietat molecular pot estar ampliada. L'exemple més evident és la utilització de la *QS-SM* del fragment COOH per quantificar els efectes electrònics de la substitució sistemàtica en sèries d'àcids carboxílics, com ho corroboren les excel·lents correlacions entre  $Z_{XX,R}^{\text{COOH}}/\sigma$  descrites en l'article 7.1 i en la taula 7.3.

### 7.4.1 Descripció de l'aproximació *QSAR* emprant *QS-SM* de fragments

En tots els estudis realitzats se segueixen unes pautes comunes, que formen part de l'esquema computacional desenvolupat per al mètode *QS-SM* de fragments. Els principals estadis del procés proposat es descriuen tot seguit.

**Optimització de les geometries moleculars.** Normalment se segueix el criteri d'agafar el conformer amb energia més baixa, a no ser que s'estudiï alguna propietat per la qual hi hagi dades de quina és la conformació activa amb què les molècules interaccionen amb el receptor.

**Construcció de les PASA DF.** A partir de les geometries moleculars i un conjunt de funcions ASA atòmiques, és fàcil construir la densitat de cada molècula com:

$$\mathbf{r}_A^{PASA}(\mathbf{r}) = \sum_a P_a \mathbf{r}_a^{ASA}(\mathbf{r}). \quad (7.9)$$

Normalment el coeficient  $P_a$  es defineix igual al nombre atòmic, però quan es fan estudis de QS-SM de fragments emprant l'aproximació PASA, apareixen algunes limitacions, en especial quan s'examinen fragments d'un sol àtom. La no variació dels paràmetres geomètrics i la definició unívoca de les funcions ASA atòmiques, fa que el valor de la QS-SM d'un fragment format per un sol àtom sigui constant indistintament de l'entorn molecular que l'envolta. Per evitar-ho, els coeficients  $P_a$  de l'equació (7.9) s'equiparen als valors de les càrregues atòmiques determinades prèviament en el càlcul d'optimització de la geometria molecular. Degut a la dimensió dels compostos considerats, normalment s'utilitzen mètodes de càlcul semiempíric, i llavors els coeficients  $P_a$  es defineixen només sobre els electrons de valència de cada àtom.

**Definició dels fragments moleculars.** L'elecció del fragment responsable de l'activitat biològica no és evident en la majoria d'estudis QSAR. Possiblement l'exemple més simple de càlcul sobre fragments ha estat l'ús de la QS-SM del grup COOH per quantificar els efectes electrònics de la substitució sistemàtica en sèries d'àcids carboxílics, presentat en l'apartat anterior. Però per altres sèries de molècules, la determinació del(s) fragment(s) rellevant d'una determinada activitat biològica serà una part fonamental de l'estudi. En els exemples que segueixen a continuació, s'ha utilitzat el coneixement a priori de les regions responsables de l'activitat biològica extretes de càlculs QSAR previs per localitzar els fragments. Mentre que en el darrer apartat del capítol es descriurà un mètode general que permet identificar els fragments moleculars importants d'una determinada activitat des del punt de vista de les QS-SM.

**Càlcul de les QS-SM.** Una vegada definits els fragments  $X$  comuns a totes les molècules que formen el conjunt estudiat, es calculen les respectives QS-SM:  $Z_{AA}^X(\Omega)$ , a més de la QS-SM de tota la molècula,  $Z_{AA}(\Omega)$ . Essent  $n$  el nombre de molècules que formen la sèrie analitzada, i  $m$  el nombre total de QS-SM, el resultat és una matriu de descriptors  $\mathbf{Z}$  de dimensió  $(n \times m)$ . Normalment s'avaluen dos tipus de mesures, les de solapament i les de Coulomb, per tant el resultat final són dues matrius de descriptors. Cadascuna d'elles es pot expressar com un vector filera de dimensió  $m$ ,  $\mathbf{Z} = \{\mathbf{z}_x\}$ , on els elements són les columnes  $(n \times 1)$  de la matriu  $\mathbf{Z}$ . Abans de fer l'anàlisi MLR, s'estandarditza cada vector columna  $\mathbf{z}_x$  de manera que s'obtenen unes noves variables amb mitjana zero i desviació estàndard u:

$$\mathbf{q}_x = s^{-1}(\mathbf{z}_x - \bar{z}_x \mathbf{1}) \quad (7.10)$$

de manera anàloga a l'equació (5.39).

**Construcció de models MLR.** Els models matemàtics es determinen sobre la nova matriu de descriptors estandarditzats. El procés de centrar les variables respecte a la mitjana i escalar-les en funció de la desviació estàndard no altera els valors del coeficient de correlació  $r$  si es compara amb el dels descriptors originals. En canvi, els coeficients de la regressió de les equacions multilineals seran comparables, i es poden entendre com la contribució de cada descriptor al model QSAR. El nombre total de models matemàtics que es construeixen ve donat pel nombre de paràmetres MLR,  $k$ , i el nombre de descriptors,  $m$ . De totes les possibles combinacions  $\binom{m}{k}$ , s'escull com a òptima la que dóna un valor del coeficient de predicció  $q^2$  més alt. Així es determinen les mesures QS-SM que donen un model MLR amb major capacitat predictiva. L'avaluació del coeficient  $q^2$  suposa una clara diferència amb referència a les anàlisis de QSAR clàssica, on normalment se selecciona el millor model en funció del coeficient de determinació  $r^2$ . Totes les possibles combinacions de  $m$  mesures  $\theta_{AA}^X$  amb  $k$  descriptors són generades mitjançant un algorisme de sumes aniuades NSS.<sup>42,43</sup>

**Validació dels models.** A més del coeficient de validació creuada, es pot calcular el test d'aleatorització del vector de les activitats biològiques o dividir el conjunt de dades en dos grups, la sèrie d'exploració amb la qual es construeix el model *MLR* que posteriorment s'aplica en els compostos del conjunt de test per predir-ne la seva activitat. Ambdues metodologies s'han descrit en l'apartat 5.5.

### 7.4.2 Exemples QSAR

En l'article 7.2 es reproduïxen alguns models *QSAR* que prèviament havien estat generats com una combinació lineal del paràmetre  $\log P$  i la constant  $\sigma$ , però ara emprant els seus equivalents teòrics. El principal objectiu és demostrar que els descriptors teòrics seleccionats correctament són més avantatjatsos que els descriptors empírics quan s'apliquen en anàlisis *QSAR* clàssiques. En l'article 7.2 es presenten tres exemples de sèries de compostos químics amb activitat biològica: derivats de la sulfonamida benzènica que presenten afinitat d'unió a l'anhidrasa carbònica; derivats de la benzilamina que són inhibidors competitiu de l'enzim proteolític tripsina; i derivats de l'indole amb capacitat de desplaçar el [3H] flunitrazepam en la reacció d'enllaç amb el receptor benzodiazepínic. Concretament en els derivats de l'indole s'estudia la inversa de la concentració molar necessària per produir un 50% d'inhibició en un assaig amb membrana de cervell boví. Totes les geometries moleculars s'han optimitzat a nivell semiempíric, emprant l'Hamiltonià AM1<sup>36</sup> implementat en el programa AMPAC.<sup>41</sup>

**Derivats de la sulfonamida benzènica.** El conjunt està format per 29 molècules, amb estructura comuna  $X-C_6H_4-SO_2NH_2$ , que es llisten en la *Table 1* de l'article 7.2. Hi ha 18 derivats *para*-substituïts, 5 *meta*, 5 *orto* i el compost genèric  $X=H$ . En l'estudi previ amb el mètode de Hansch,<sup>44</sup> primer es corrala l'activitat biològica de 19 molècules, corresponents als substituents en posició *para* més el compost genèric, amb els paràmetres  $\log P$  i  $\sigma$ . La inclusió dels substituents *orto* i *meta* en el model de Hansch final,<sup>44</sup> va anar acompanyada de la definició d'unes variables binàries com les definides en el model de Free-Wilson que indiquen la presència o absència d'un substituent en

aquestes posicions. Així, en el model final donat pel conjunt de 29 molècules,<sup>44</sup> es combinen quatre paràmetres:  $\log P$ ,  $\sigma$  i dues variables binàries amb valors 0 i 1 dependent de si els substituents en posició *orto* i *meta* són hidrogen o no. Quant als models emprant *QS-SM* de fragment es pressuposa inicialment que el grup  $\text{SO}_2\text{NH}_2$  és el principal exponent de l'activitat biològica. Analitzant el subgrup de 19 compostos substituïts en posició *para*, s'obté una excel·lent recta de regressió  $\theta_{AA}^{\text{SO}_2\text{NH}_2} / \sigma$ . Però en canvi no s'obté una bona correlació entre  $\theta_{AA}$  i  $\log P$ . Les mesures  $\theta_{AA}$  són molt sensibles a la sèrie estudiada. Observant la taula 7.1 es dedueix que les correlacions  $\theta_{AA} / \log P$  es donen només en sèries homòlogues de compostos. Si s'analitza els substituents *p*-X dels primers 19 compostos de la *Table 1*, s'observa que estan formats per hidrocarburs alifàtics, èsters i amides. Per obtenir una bona recta de regressió  $\theta_{AA} / \log P$  s'ha d'afegir una variable lògica al model que pren el valor 0 en les molècules amb substituent hidrocarbur, i 1 en la resta. El resultat és un empitjorant del model original proposat per Hansch, perquè són necessaris tres descriptors en lloc de dos. A més, el descriptor addicional és una variable binària, que és precisament un dels aspectes de les anàlisis de *QSAR* clàssica que es vol millorar amb el mètode *QS-SM* de fragments. Per solucionar aquest problema s'han definit nous fragments moleculars sobre l'estructura comuna de les 29 molècules, determinant els seus corresponents valors *QS-SM* i utilitzant-los en la generació de nous models *QSAR*. El resultat és una equació *MLR* amb quatre descriptors definida sobre el conjunt de 29 molècules que és estadísticament equiparable a la donada per Hansch en la referència [44]. Per exemple, s'ha analitzat el grup  $\text{SO}_2\text{NH}_2$  com si estès format per dos fragments independents,  $\text{SO}_2$  i  $\text{NH}_2$ , que se suposa interaccionen amb el receptor en llocs diferents. Altres fragments considerats en la construcció de les equacions *MLR* han estat l'anell benzènic,  $\text{C}_6\text{H}_4$ , i els carbonis de l'anell en les posicions *orto*, *meta* i *para*: *o*-C, *m*-C i *p*-C.

**Derivats de la benzilamina.** El segon exemple està format per un conjunt de 22 molècules, amb estructura comuna  $p\text{-X-C}_6\text{H}_4\text{-CH}_2\text{NH}_2$  i substituents llistats en la *Table 3* de l'article 7.2. Cada producte de la sèrie està caracteritzat pel substituent X en la posició *para*. En l'estudi previ emprant el mètode de Hansch,<sup>15</sup> s'havia generat un model *QSAR* amb dos paràmetres:  $\log P$  i  $\sigma$ , sobre una sèrie de 9 èsters del conjunt original.<sup>45</sup> En el primer estudi teòric que s'exposa en l'article 7.2, es reproduïx el

model obtingut per Hansch en [15] emprant els descriptors  $\theta_{AA}$  i  $\theta_{AA}^{CH_2NH_2}$ . Però aquest no ha estat la única anàlisi derivada de l'aproximació *QS-SM* de fragments. Examinant la sèrie completa de molècules,<sup>45</sup> s'entreveu que una de les mancances del mètode de Hansch és l'existència de molts substituents dels quals es desconeix el valor empíric de la constant  $\sigma$ . Aquestes limitacions no es donen en l'aproximació *QS-SM* de fragments, perquè coneguda la geometria i la densitat *PASA* d'una sèrie química, sempre es pot calcular el valor de la mesura de semblança. Igual que en l'exemple dels derivats de la sulfonamida benzènica, es generen les *QS-SM* de diferents fragments moleculars comuns en tots els compostos. S'ha analitzat el grup  $CH_2NH_2$  i els subgrups  $CH_2$  i  $NH_2$ , a més de l'anell benzènic  $C_6H_4$ . El resultat final dels derivats de la benzilamina és un model *MLR* amb tres paràmetres, ratificat pels tests de validació.

**Derivats de l'indole.** El tercer estudi el componen 23 derivats de l'indole que estan resumits en la *Table 5* de l'article 7.2. La principal diferència respecte als dos exemples anteriors és la major complexitat de les estructures moleculars considerades. L'esquelet comú dels indoles està format per tres anells i una cadena lineal. A més, presenten quatre possibles posicions on hi ha variació dels substituents, mentre que en les dues sèries anteriors hi havia un anell aromàtic amb un únic substituent variable. El model proposat per Hansch,<sup>46</sup> refereix a un subconjunt de 20 molècules i està format per tres paràmetres: la constant  $\sigma$  d'un dels substituents, i dues variables lògiques. Són variables que prenen el valor 0 o 1 per indicar la presència o absència de més d'una característica estructural referida a les quatre substitucions dels indoles, i no aporten informació sobre la interacció lligand–receptor. Paral·lelament a la reproducció del model proposat per Hansch, s'ha trobat en la bibliografia uns estudis que delimiten com ha de ser el farmacòfor per la reacció d'enllaç amb el receptor benzodiazepínic,<sup>47</sup> i identifiquen les possibles zones actives en els derivats de l'indole.<sup>48</sup> Amb aquesta informació s'han definit els fragments suposadament relacionats amb l'activitat, i s'han calculat les corresponents *QS-SM*. Fent una recerca sistemàtica sobre totes les *QS-SM* definides, s'obté una equació *MLR* òptima, amb bons resultats estadístics, que queden refutats mitjançant el test aleatori sobre la variable dependent. A més amb el model matemàtic resultant s'ha predit l'activitat de sis nous compostos que no havien participat en el model d'ajust i dels quals es coneix la propietat analitzada.



**Article 7.2**

---

**Autors:** *Lluís Amat, Ramon Carbó-Dorca, Robert Ponec.*

**Títol:** *Simple linear QSAR models based on quantum similarity measures*

**Revista:** *Journal Medicinal Chemistry*

**Volum:** 42      **Pàgines, inicial:** 5169    **final:** 5180    **Any:** 1999

---

## Simple Linear QSAR Models Based on Quantum Similarity Measures

Lluís Amat and Ramon Carbó-Dorca\*

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

Robert Ponec

*Institute of Chemical Process Fundamentals, Czech Academy of Sciences, Prague 6, Suchbát 2, 165 02 Czech Republic*

Received May 5, 1999

A novel QSAR approach based on quantum similarity measures was developed and tested in this paper. This approach consists of replacing the usual physicochemical parameters employed in QSAR analysis, such as octanol–water partition coefficient or Hammett  $\sigma$  constant, by appropriate quantum chemical descriptors. The methodological basis for this substitution is found in recent theoretical studies [*J. Comput. Chem.* **1998**, *19*, 1575–1583, *J. Comput.-Aided Mol. Des.* **1999**, *13*, 259–270], in which it was demonstrated that both molecular hydrophobic character and electronic substituent effect can be modeled by appropriately chosen quantum self-similarity measures (QS-SM). The most important aim of this study was to prove that selected QS-SM descriptors can be advantageously used in empirical QSAR analysis instead of classical descriptors. For this purpose several QSAR correlations are proposed, in which empirical descriptors such as Hammett  $\sigma$  constants or  $\log P$  values are replaced by the appropriate QS-SM. These examples involve: (i) a set of benzenesulfonamides which bind to human carbonic anhydrase, (ii) a set of benzylamines as competitive inhibitors of the enzyme trypsin, and (iii) a set of indole derivatives which are benzodiazepine receptor inverse agonist site ligands. Simple linear QSAR models were developed in order to obtain mathematical relationships between the biological activity and the pertinent quantum chemical descriptors. The validity of the obtained QSAR models is supported by comparison of the observed and predicted values of the biological activity and by a statistical analysis based on a randomization test.

### Introduction

In the past few years much effort has been devoted to applying the idea of quantum similarity measures (QSM) to rational drug design.<sup>1–15</sup> Because of its importance, this area of chemistry has experienced rapid growth. The mathematical background for this new expanding field was formulated some time ago by Carbó et al.,<sup>16</sup> who introduced the concept of QSM. Since then, great progress has been made not only in basic methodology but also in the formulation of robust computational schemes.<sup>17–28</sup> The basic idea of the above similarity approach to QSAR is to replace the traditional parameters in empirical QSAR analysis by selected theoretical descriptors based on QSM.

In keeping with this general philosophy, the present article reports an attempt to develop simple linear QSAR models based on quantum mechanical descriptors instead of empirical physicochemical parameters, characterizing the molecular hydrophobicity and electronic substituent effect in classical QSAR. The study is based on previous reports which described how the quantum self-similarity measure (QS-SM) of the whole molecule could be used as a descriptor of molecular hydrophobicity ( $\log P$ ).<sup>5,6</sup> Similarly, the electronic substituent effect may be appropriately modeled by fragment QS-SM corresponding to a functional group (re)active in a given process.<sup>6,7</sup> The fact that electronic phenomena such as

the substituent effect can be replaced by means of QS-SM attached to local molecular regions can be explained through the recently reported holographic electron density theorem.<sup>29</sup> This theorem states that all the information contained in the total electron density of the whole molecule is also contained in the density of any local fragment of the molecule. In consequence, the QS-SM characterizing the functional group (re)active in a given process can be used as an appropriate descriptor.

Several molecular sets were examined in this study: (i) a series of benzenesulfonamides that show some affinity to binding to the human carbonic anhydrase; (ii) a series of benzylamine derivatives as competitive inhibitors of the proteolytic enzyme trypsin; (iii) a set of indole derivatives which are benzodiazepine receptor inverse agonists. Indole derivatives are able to displace [<sup>3</sup>H]flunitrazepam from binding to bovine cortical membranes.

As will be shown, the theoretical QSAR models using QSM descriptors show statistical reliability comparable to descriptors derived from empirical correlations.<sup>30–32</sup>

### Theoretical Framework

The idea of QSM arises from the incorporation of the intuitive concept of molecular similarity into the framework of quantum mechanics. According to this mechanics, all the information concerning a quantum object (QO) is contained in the associated electron density

\* To whom correspondence should be addressed. E-mail: director@iqc.udg.es. Phone: 34 972 418359. Fax: 34 972 418356.

function obtained from the corresponding wave function square module. From this point of view, electron density can be regarded as an ultimate molecular descriptor. In consequence, the similarity of any two QO can be assessed quantitatively by comparing the similarity of the corresponding electron density clouds.<sup>27</sup>

In this way, a consistent expression of QSM between two QO *A* and *B*, described by the first-order density functions  $\{\rho_A(r), \rho_B(r)\}$ , is defined by the integral

$$Z_{AB}(\Omega) = \int \int \rho_A(r_1) \Omega(r_1, r_2) \rho_B(r_2) dr_1 dr_2 \quad (1)$$

where  $\Omega$  is a positive definite operator. The most commonly used form of the two-electron operator  $\Omega$  is a Dirac  $\delta$  function:  $\Omega(r_1, r_2) = \delta(r_1 - r_2)$ . By replacing this operator into eq 1, a general definition of the so-called overlap-like QSM is obtained:

$$Z_{AB} = \int \rho_A(r) \rho_B(r) dr \quad (2)$$

The values of similarity measures  $Z_{AB}$  defined by eq 2 depend on the relative translation and orientation of the compared QO *A* and *B* in 3D space. This implies that in order to get a meaningful and unique value of QSM, the relative mutual position of both QO has to be optimized so that a maximal value of the integral in eq 2 is reached.<sup>28</sup> However, such optimization depending of the relative position of both QO is no longer necessary when *A* and *B* are identical. In this case, definition 2 yields an invariant quantum self-similarity measure (QS-SM):

$$Z_{AA} = \int |\rho_A(r)|^2 dr \quad (3)$$

The fact that the relative position optimization becomes irrelevant for self-similarity measures is crucial from a computational point of view, not only because of substantial reduction of similarity integral measure computation time, but also because it reduces the conformational dependence of the results. As a consequence, because the 3D alignment procedure is avoided, QS-SM acquire special relevance as molecular descriptors in building up quantum mechanical equivalents of empirical QSAR Hansch-like models.

The procedure of generating these theoretical QSAR models is as follows: First, the appropriate QS-SM are computed for the whole series of compounds belonging to the studied set. These self-similarity measures are then arranged in the form of column vectors,  $\mathbf{z}$ , entering into linear regression analysis. But before this, each column vector  $\mathbf{z}$  is standardized so as to give new scaled variables with zero mean and unit variance. The idea underlying such statistical standardization is to ensure comparable weights of individual molecular descriptors in the final QSAR model. This standardization is described in the usual way

$$\theta = s^{-1}(\mathbf{z} - \langle \mathbf{z} \rangle \mathbf{1}) \quad (4)$$

where  $s$  and  $\langle \mathbf{z} \rangle$  are the standard deviation and the arithmetic mean of the original descriptors, respectively.

A great number of methodologies and computational algorithms have been developed for the practical implementation of QSM, which opens up the possibility of their application in many areas of theoretical chemistry.

Among these techniques, the most widely used is what is known as atomic shell approximation (ASA).<sup>24–26</sup> Since this approximation is also employed for the calculation of QS-SM in this study, we consider the basic idea of the ASA approach worth re-stating.

ASA density functions are constructed as a linear combination of spherical functions, with the restriction that all coefficients of expansion have to be real and positive. Thus, this constraint enables the statistical meaning of a correct probability distribution to be preserved. In addition, a *promolecular* model is employed, based on a plausible description of molecular density functions as a sum of individual atomic contributions. Then, the first-order density function under the *promolecular* ASA form for a molecular QO *A* may be expressed as

$$\rho_A^{\text{ASA}}(r) = \sum_{a \in A} P_a \rho_a^{\text{ASA}}(r) \quad (5)$$

where the coefficient  $P_a$  represents the atom *a* total charge, and  $\rho_a^{\text{ASA}}(r)$  the density function. In the present study, QS-SM were computed using weighting factors  $P_a$  equal to total valence atomic charge on individual atoms. Density for a given atom *a* is expressed as a linear combination of square normalized 1S-type GTO

$$\rho_a^{\text{ASA}}(r) = \sum_{i \in a} w_i |S_i(r - R_a; \zeta_i)|^2 \quad (6)$$

where the sum in eq 6 is performed over the functions associated to the atomic shells. Within this promolecular ASA model, only the coefficients  $w_i$  and exponents  $\zeta_i$  are needed to construct the density function. In this paper, one function is used to modulate density on H atoms, three functions are needed for C, O, and N atoms, and four functions are needed for Cl atoms. Coefficients  $w_i$  and exponents  $\zeta_i$  used here can be downloaded from a World Wide Web site.<sup>33</sup>

### Simple Linear QSAR Models Using QS-SM

Around 1960, Hansch and co-workers<sup>34,35</sup> introduced a new approach which proved to be especially fruitful in the field of rational drug design. Their approach was based on the application of linear-free energy relationships (LFER) to correlate biological activities with appropriate physicochemical descriptors. Since then, the application of what is known as QSAR became a respectable and widely used methodology in pharmacological research. A wealth of empirical descriptors relating to various physicochemical properties was introduced in consequence. The fundamental idea of the Hansch approach consists of the design of suitable QSAR models in the form of a multiple linear regression (MLR) between physicochemical descriptors and biological activities:

$$\text{biological activity} = f(\text{molecular or fragmental contributions}) = f(\log P, \sigma, E_s) \quad (7)$$

The most usual factors determining the biological activity are the hydrophobic character, characterized by  $\log P$  values, and the substituent electronic and steric effect represented by the Hammett and Taft constants.

On the basis of these parameters, the traditional drug design consists of combining these molecular descriptors in the form of an MLR so as to get the best statistical description of the biological data.

In order to place the above empirical process on a safer theoretical footing and so to provide a theoretical interpretation of the origins of QSAR, a mathematical formalism, based on the combination of the idea of molecular similarity with quantum mechanical postulates, has been proposed in a recent study.<sup>23</sup> In the following part, the basic idea of this rationalization will be briefly explained.

According to quantum mechanics, any observable property of a quantum system  $I$ ,  $\pi_I$ , for which the density function  $\rho_I(r)$  is known, can be calculated as the expectation value of an associated hermitean operator  $\Omega(r)$

$$\pi_I = \langle \omega \rangle = \int \Omega(r) \rho_I(r) dr \quad (8)$$

Equation 8 represents a continuous description of an observable property. However, such a continuous description is considerably different from the intrinsically discrete form of empirical QSAR. A clue to the resolution of this difference and to the theoretical formulation of QSAR lies precisely in the application of the idea of molecular similarity. For this purpose, given a set of molecules ( $A, B, C, \dots, M$ ) whose properties will be studied, first pairwise QSM for all possible molecular couples is calculated. These QSM can be conveniently arranged in the form of a matrix  $\mathbf{Z} = \{Z_{IJ}\}$ , which can be considered as a hypermatrix formed by column vectors as elements,  $\mathbf{Z} = \{\mathbf{z}_I\}$ . Using this symmetric matrix, the molecular property  $\pi_I$  can be approximated according to the general equation

$$\pi_I \approx \mathbf{a}^T \mathbf{Z}_I = \sum_K a_K Z_{KI} \quad (9)$$

where  $\mathbf{a}$  is an  $n$ -dimensional vector associated with the discrete representation of the unknown operator  $\Omega$ . This equation represents the discrete counterpart of eq 8. The unknown coefficients  $\mathbf{a}$ , characterizing the operator  $\Omega$ , can be determined in a least-squares manner.

Although eq 9 represents the most general form of theoretical QSAR models, in some cases the form of the correlation equation can be further simplified. Such a simplification is typical of a situation in which it is possible to extract the QS-SM,  $Z_{II}$ , from the rest of the elements of the similarity matrix,  $\{Z_{KI}, K \neq I\}$ , leading to the result:

$$\pi_I \approx a_I Z_{II} + \sum_{K \neq I} a_K Z_{KI} \quad (10)$$

In some cases, particularly when homogeneous series of QO are studied, the terms  $\alpha = a_I$  and  $\beta = \sum_{K \neq I} a_K Z_{KI}$  can be considered as constants. Then, a simple linear equation may be expressed by means of the QS-SM, which represents the theoretical counterpart of the simple one-parameter QSAR model, like the Hammett equation:

$$\pi_I \approx \alpha Z_{II} + \beta \quad (11)$$

The situation with the correlation of biological data is, however, more complex since the final biological effect is usually due to the combination of several different factors. This suggests that in this case the theoretical QSAR models should have the form of multilineal correlation equations. Another important factor, which plays an important role when correlating biological data, is that the majority of the processes responsible for the observed activity are usually restricted to certain more or less localized regions of the molecule (pharmacophore, binding site, etc). As a consequence, in some cases it is possible and more useful to focus just on the comparison of these active molecular regions,  $R$ . Under these circumstances the original  $\beta$  constant in eq 11 can be approximately rewritten in the form of a set of fragment self-similarities. A justification of this procedure may be obtained when inspecting the definition of  $\beta$  in eq 11 as follows

$$\beta \approx \sum_{K \neq I} a_K \sum_{b \in Ka \in I} P_b^K P_a^I Z_{ba}^{KI} \quad (12)$$

where  $\{Z_{ba}^{KI}\}$  are interatomic similarity contributions involving molecules  $K$  and  $I$  to the global molecular QSM,  $Z_{KI}$ , when comparing molecular densities written in the ASA approach as in eq 5. Reordering terms

$$\beta \approx \sum_{a \in I} P_a^I \vartheta_a^I = B \quad (13)$$

and the symbol in the sum is defined as:

$$\vartheta_a^I = \sum_{K \neq I} a_K \sum_{b \in K} P_b^K Z_{ba}^{KI} \quad (14)$$

while the whole result can be further approximated using the fact that  $B$  only depends explicitly on atomic contributions of molecule  $I$ , that is,

$$B \approx \sum_R \alpha_R Z_{I,RR} + \gamma \quad (15)$$

where  $R$  are groups of atoms present in molecule  $I$  as well as in the common skeleton shared by the rest of the studied molecular set. To smooth the successive approximation source of errors, the new coefficients,  $\{\alpha_R, \gamma\}$  in eq 15 are to be adapted to a specific molecular set, using a conventional fitting procedure.

Because  $\beta \approx B$ , eq 15, when substituted in eq 11, can be regarded as an alternative multilineal theoretical QSAR model. In this equation,  $\{Z_{I,RR}\}$  are the appropriate self-similarity measures for each individual fragment  $R$  contributing to the biological response. The fragment self-similarities constitute a simple way to take into account in some cases the variability of the supposed constant  $\beta$ . Moreover they provide information about the relevant parts of the common skeleton which can be taken as responsible for biological activity.

The basic goal of the present article is to demonstrate that in view of the analogy between empirical QSAR and theoretical equations, the physicochemical descriptors employed in classical QSAR studies can be replaced by appropriate theoretical descriptors based on QS-SM. In the following section some examples of such replacement will be presented, with the aim of showing that



theoretical QSAR models can be used to correlate biological data at least as successfully as the classical QSAR models.

## Results and Discussion

Having introduced the necessary theoretical background, its application to the construction of theoretical QSAR models for the series of biologically active molecules studied will be reported in this section. These molecular series involve three well-defined cases: (i) benzenesulfonamides which show binding affinity with human carbonic anhydrase (HCA); (ii) benzylamine derivatives as competitive inhibitors of the proteolytic enzyme trypsin; (iii) indole derivatives which are benzodiazepine receptor inverse agonists.

**1. Preliminary Considerations.** Before presenting the results, some general remarks concerning the computation of QS-SM are summarized together with the statistical aspects of the QSARs obtained:

- Molecular geometry of all involved molecules was fully optimized using the AMPAC program<sup>36</sup> and semiempirical AM1 Hamiltonian.<sup>37</sup>

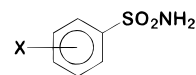
- The following QS-SM were calculated for each series of molecules: (a) QS-SM for the whole molecule as an alternative descriptor to log  $P$ ,<sup>5</sup> (b) the variation of the electronic structure of the fragment presumably responsible for the biological activity, induced by the systematic variation of substituents, was modeled by QS-SM for appropriate molecular fragments.

- For the statistical analysis of the QSAR models, two regression coefficients were calculated: conventional squared regression coefficient ( $r^2$ ) and the cross-validated (CV) coefficient for prediction ( $q^2$ ). This latter coefficient, which permits evaluation of the predictive power of the model, is defined as  $q^2 = (1 - \text{PRESS}/\text{SD})$ , where PRESS (predictive residual sum of squares) is the sum of squared errors of predictions in a leave-one-out (LOO) CV analysis, and SD is the squared sum of the difference of the observed values from their mean. A statistically reasonable QSAR model usually requires the  $q^2$  value to be greater than 0.6.<sup>38</sup>

- Using a nested summation symbol (NSS) algorithm,<sup>39,40</sup> all possible combinations of the computed QS-SM were generated and subsequently employed in QSAR models. Using this approach, the corresponding optimal QSAR model, in which a QS-SM set yields a maximal value of the  $q^2$  coefficient, was chosen. In this way, the study focused on determining which QS-SM descriptors produced the linear regression model with the best predictability.

- Finally, to verify that the results of the QSAR models designed are not due to accidental correlations or to over-parametrization of the model, a randomization test<sup>41</sup> was performed. This test consists of randomly rearranging the order of the components of the vector of biological activity data and correlating these rearranged vectors with the vector of QS-SM. This procedure was repeated 100 times for each chosen QS-SM set, keeping the coefficients  $r^2$  and  $q^2$  for each random run and recording all the obtained ( $r^2$ ,  $q^2$ ) pairs as points on a graph at the end. A consistent QSAR model is

**Chart 1.** Common Molecular Structure for Substituted Benzenesulfonamides



**Table 1.** Inhibitor Constants for the Binding of X-C<sub>6</sub>H<sub>4</sub>SO<sub>2</sub>NH<sub>2</sub> to HCA

	X	observed log $K^a$
1	H	6.69
2	4-CH <sub>3</sub>	7.09
3	4-C <sub>2</sub> H <sub>5</sub>	7.53
4	4-C <sub>3</sub> H <sub>7</sub>	7.77
5	4-C <sub>4</sub> H <sub>9</sub>	8.30
6	4-C <sub>5</sub> H <sub>11</sub>	8.86
7	4-CO <sub>2</sub> CH <sub>3</sub>	7.98
8	4-CO <sub>2</sub> C <sub>2</sub> H <sub>5</sub>	8.50
9	4-CO <sub>2</sub> C <sub>3</sub> H <sub>7</sub>	8.77
10	4-CO <sub>2</sub> C <sub>4</sub> H <sub>9</sub>	9.11
11	4-CO <sub>2</sub> C <sub>5</sub> H <sub>11</sub>	9.39
12	4-CO <sub>2</sub> C <sub>6</sub> H <sub>13</sub>	9.39
13	4-CONHC <sub>3</sub> H <sub>7</sub>	7.08
14	4-CONHC <sub>2</sub> H <sub>5</sub>	7.53
15	4-CONHC <sub>3</sub> H <sub>7</sub>	8.08
16	4-CONHC <sub>4</sub> H <sub>9</sub>	8.49
17	4-CONHC <sub>5</sub> H <sub>11</sub>	8.75
18	4-CONHC <sub>6</sub> H <sub>13</sub>	8.88
19	4-CONHC <sub>7</sub> H <sub>15</sub>	8.93
20	3-CO <sub>2</sub> CH <sub>3</sub>	5.87
21	3-CO <sub>2</sub> C <sub>2</sub> H <sub>5</sub>	6.21
22	3-CO <sub>2</sub> C <sub>3</sub> H <sub>7</sub>	6.44
23	3-CO <sub>2</sub> C <sub>4</sub> H <sub>9</sub>	6.95
24	3-CO <sub>2</sub> C <sub>5</sub> H <sub>11</sub>	6.86
25	2-CO <sub>2</sub> CH <sub>3</sub>	4.41
26	2-CO <sub>2</sub> C <sub>2</sub> H <sub>5</sub>	4.80
27	2-CO <sub>2</sub> C <sub>3</sub> H <sub>7</sub>	5.28
28	2-CO <sub>2</sub> C <sub>4</sub> H <sub>9</sub>	5.76
29	2-CO <sub>2</sub> C <sub>5</sub> H <sub>11</sub>	6.18

<sup>a</sup> From ref 30.

obtained when only the original arrangement of the activities produces a satisfactory regression model.

**2. Results.** The first two examples presented in this paper refer to QSAR studies of enzyme–ligand interactions, and the third one deals with the prediction of the ability of substituted indole derivatives to displace [<sup>3</sup>H] flunitrazepam from binding to bovine cortical membranes. The traditional approach to the description of such biological activity data is based on the construction of classical QSAR models using as descriptors hydrophobicity parameters (log  $P$ ), Hammett substituent constant, etc. As was explained earlier, the aim of this study was to propose a new universal methodology, based on the use of theoretical QS-SM based descriptors, which could serve as an alternative procedure for designing new QSAR models.

**(a) Benzenesulfonamide Derivatives.** A set of 29 substituted benzenesulfonamides, with a common structure shown in Chart 1 and substituents listed in Table 1, was studied as HCA inhibitors. Traditional studies have shown that the HCA inhibitory activity of substituted benzenesulfonamides is predominantly influenced by two basic factors: the hydrophobic interactions of these molecules with enzyme–receptor cavity, and the electronic structure of the active SO<sub>2</sub>NH<sub>2</sub> group reflecting the systematic variation of substituents within the series. On the basis of these findings, Hansch proposed empirical QSAR models<sup>30</sup> using log  $P$  and Hammett's substituent constant  $\sigma$  as appropriate descriptors. For the series of 19 *para*-substituted benzenesulfonamides

**Table 2.** QS-SM ( $Z_{AA}$ ) and Scaled QS-SM ( $\theta_{AA}$ ) Used To Derive Eqs 17–19 and 22–24 for the Binding of X-C<sub>6</sub>H<sub>4</sub>SO<sub>2</sub>NH<sub>2</sub> to HCA<sup>a</sup>

	$Z_{AA}$	$\theta_{AA}$	$Z_{AA}^{SO_2NH_2}$	$\theta_{AA}^{SO_2NH_2}$	$Z_{AA}^{SO_2}$	$\theta_{AA}^{SO_2}$	$Z_{AA}^{NH_2}$	$\theta_{AA}^{NH_2}$	$Z_{AA}^{m-C}$	$\theta_{AA}^{m-C}$
1	283.7310	-2.41624	189.7814	1.29909	151.9538	1.05833	37.8017	0.20978	15.0628	1.31094
2	298.9168	-2.13996	189.7945	1.56123	151.9679	1.29930	37.8008	0.15125	15.0831	1.44959
3	314.2117	-1.86171	189.7960	1.59287	151.9706	1.34686	37.7996	0.07627	15.0795	1.42482
4	329.5206	-1.58319	189.7943	1.55763	151.9689	1.31674	37.7996	0.07627	15.0860	1.46941
5	344.8184	-1.30488	189.7960	1.59231	151.9706	1.34650	37.7996	0.07544	15.0846	1.45950
6	360.1176	-1.02654	189.7960	1.59265	151.9706	1.34679	37.7996	0.07544	15.0853	1.46446
7	409.2965	-0.13184	189.6808	-0.72517	151.8467	-0.78089	37.8078	0.60322	14.6662	-1.39857
8	424.7788	0.14983	189.6858	-0.62495	151.8517	-0.69590	37.8079	0.60680	14.6690	-1.37903
9	440.0616	0.42787	189.6858	-0.62457	151.8516	-0.69606	37.8079	0.60782	14.6690	-1.37903
10	455.3495	0.70600	189.6857	-0.62594	151.8516	-0.69638	37.8079	0.60488	14.6697	-1.37414
11	470.6511	0.98438	189.6858	-0.62501	151.8517	-0.69568	37.8079	0.60571	14.6690	-1.37903
12	485.9479	1.26267	189.6857	-0.62582	151.8516	-0.69616	37.8079	0.60488	14.6690	-1.37903
13	391.2925	-0.45938	189.7171	0.00466	151.8892	-0.05090	37.8016	0.20620	14.9045	0.22969
14	406.7968	-0.17731	189.7189	0.04043	151.8909	-0.02171	37.8017	0.21252	14.9189	0.32820
15	422.0553	0.10028	189.7189	0.04081	151.8910	-0.02103	37.8017	0.21124	14.9204	0.33805
16	437.3506	0.37855	189.7188	0.04014	151.8910	-0.02113	37.8017	0.20933	14.9197	0.33312
17	452.6488	0.65687	189.7188	0.04039	151.8909	-0.02149	37.8017	0.21124	14.9197	0.33312
18	467.9466	0.93518	189.7188	0.04031	151.8909	-0.02140	37.8017	0.21124	14.9197	0.33312
19	483.2430	1.21346	189.7189	0.04083	151.8910	-0.02094	37.8017	0.21124	14.9197	0.33312
20	409.3961	-0.13003	189.6384	-1.57724	151.7981	-1.61525	37.8141	1.00305	14.9262	0.37747
21	424.8777	0.15163	189.6422	-1.50194	151.8031	-1.53016	37.8129	0.92608	14.9254	0.37254
22	440.1588	0.42964	189.6422	-1.50160	151.8031	-1.52985	37.8129	0.92608	14.9254	0.37254
23	455.4468	0.70777	189.6422	-1.50194	151.8031	-1.53020	37.8129	0.92608	14.9254	0.37254
24	470.7482	0.98615	189.6422	-1.50204	151.8031	-1.53018	37.8129	0.92608	14.9254	0.37254
25	409.2360	-0.13294	189.7274	0.21275	151.9359	0.75057	37.7646	-2.15843	14.7378	-0.90951
26	424.8852	0.15177	189.7396	0.45832	151.9468	0.93840	37.7658	-2.07770	14.7435	-0.87034
27	440.1338	0.42918	189.7381	0.42789	151.9453	0.91278	37.7658	-2.07936	14.7435	-0.87034
28	455.4154	0.70720	189.7381	0.42698	151.9453	0.91251	37.7658	-2.08140	14.7442	-0.86544
29	470.7187	0.98561	189.7401	0.46693	151.9473	0.94657	37.7658	-2.08127	14.7435	-0.87034

<sup>a</sup> Standardized values are obtained from eq 4.

(1–19 in Table 1), the following correlation equation was found:

$$\log K = 1.55\sigma + 0.65 \log P + 6.93$$

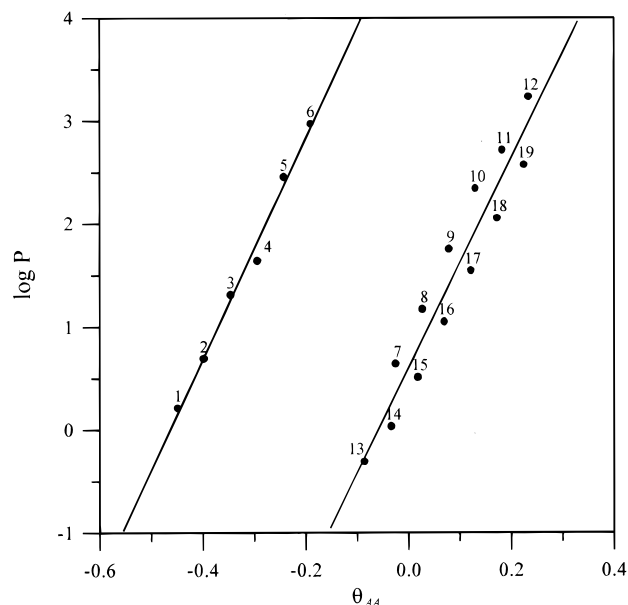
$$n = 19; \quad r^2 = 0.943 \quad (16)$$

The present approach to the design of alternative theoretical QSAR models arises from previously reported findings<sup>5–7</sup> that, in a series of structurally related molecules, both hydrophobic parameter  $\log P$  and the effect of the systematic substitution can be modeled by appropriate theoretical descriptors. Such quantum similarity descriptors have been chosen as the QS-SM  $\theta_{AA}$ , instead of  $\log P$ , and the fragment QS-SM  $\theta_{AA}^R$  replacing the substituent constant. Equation 17 is an example of such replacement: it describes the correlation of Hammett substituent constant with  $\theta_{AA}^{SO_2NH_2}$  (listed in Table 2) in a series of 19 *para*-substituted benzenesulfonamides. The correlation is fairly good.

$$\sigma = -0.2779\theta_{AA}^{SO_2NH_2} + 0.3160$$

$$n = 19; \quad r^2 = 0.966 \quad (17)$$

A slightly more complex situation arises when considering the correlation of  $\log P$  with  $\theta_{AA}$ , which in the same series splits into two regression lines with the same slope and different intercepts, as is shown in Figure 1. This specific form of correlation suggests that all these data can be described by a single regression line of the form:  $\log P = a\theta_{AA} + bI + c$ , where the variable  $I$  is the Boolean parameter, introduced to distinguish between alkyl- and nonalkyl-substituted derivatives ( $I = 0$  for molecules 1–6 and  $I = 1$  otherwise). The existence of this splitting may suggest that



**Figure 1.** Linear correlation between  $\log P$  and  $\theta_{AA}$  for a series of 19 *para*-substituted benzenesulfonamides.

the basic assumption in deriving eq 11, namely the requirement of constancy of the term  $\beta$ , is not apparently satisfied within the whole series. It thus seems quite plausible to regard this splitting as an indirect indication of the fact that the studied molecular set apparently does not form a homogeneous series and that there are in fact two different series, formed by the molecules 1–6 and 7–19. The actual form of the general MLR equation is given by

$$\log P = 1.9205\theta_{AA} - 4.2624I + 4.8523$$

$$n = 19; \quad r^2 = 0.943; \quad q^2 = 0.925 \quad (18)$$

Using such a unified correlation equation, the activity of the whole set of 19 *para*-substituted benzenesulfonamides can be described by eq 19, which can be regarded as a theoretical counterpart of the original empirical eq 16 given by Hansch:

$$\log K = 1.2511\theta_{AA} - 0.4879\theta_{AA}^{\text{SO}_2\text{NH}_2} - 2.7968I + 10.6089$$

$$n = 19; \quad r^2 = 0.947; \quad q^2 = 0.906 \quad (19)$$

Boolean variables were also used by Hansch when extending the applicability of his QSAR model to *ortho*- and *meta*-substituted benzenesulfonamides.<sup>30</sup> The corresponding Hansch QSAR equations take the form

$$\log K = 1.55\sigma + 0.62 \log P - 2.07I_m + 6.98$$

$$n = 24; \quad r^2 = 0.964 \quad (20)$$

$$\log K = 1.55\sigma + 0.64 \log P - 2.07I_m - 3.28I_o + 6.94$$

$$n = 29; \quad r^2 = 0.982 \quad (21)$$

where two additional Boolean parameters indicating the presence of *meta*- ( $I_m$ ) and *ortho*-substituents ( $I_o$ ) are included. While keeping the Hansch results, the original theoretical eq 19 can be generalized within the QS-SM procedure for the set of 24 *meta*- and *para*-derivatives, using the form

$$\log K = 1.2175\theta_{AA} - 0.4927\theta_{AA}^{\text{SO}_2\text{NH}_2} - 2.7321I - 2.6301I_m + 10.5585$$

$$n = 24; \quad r^2 = 0.971; \quad q^2 = 0.950 \quad (22)$$

and for the whole set of 29 *ortho*-, *meta*-, and *para*-substituted derivatives using

$$\log K = 1.2739\theta_{AA} - 0.4536\theta_{AA}^{\text{SO}_2\text{NH}_2} - 2.7847I - 2.5796I_m - 2.8896I_o + 10.5957$$

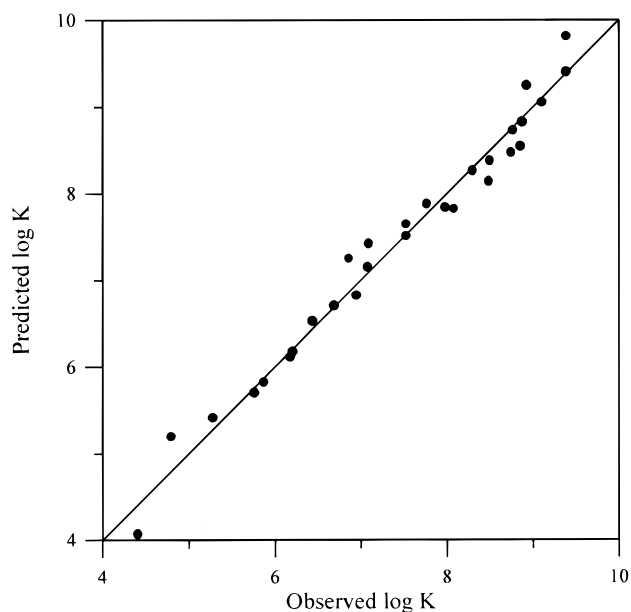
$$n = 29; \quad r^2 = 0.984; \quad q^2 = 0.973 \quad (23)$$

In addition to eqs 22 and 23, straightforwardly derived from the Hansch empirical equations, an alternative theoretical equation, which does not rely on mixing Boolean and real variables, can be proposed. This alternative equation, found by using systematically the NSS algorithm over the available QS-SM set, has the form

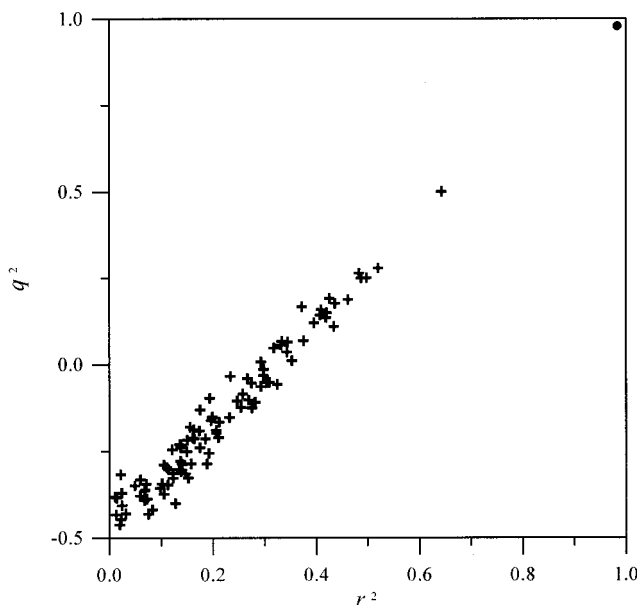
$$\log K = 1.2000\theta_{AA} + 2.3157\theta_{AA}^{\text{SO}_2\text{NH}_2} + 2.5149\theta_{AA}^{\text{NH}_2} - 0.6264\theta_{AA}^{m-C} + 7.4441$$

$$n = 29; \quad r^2 = 0.984; \quad q^2 = 0.976 \quad (24)$$

This equation points out the splitting of the fragment  $\text{SO}_2\text{NH}_2$ , expected to be responsible for the activity, into two independent subfragments:  $\text{SO}_2$  and  $\text{NH}_2$ . This splitting could suggest that the  $\text{SO}_2\text{NH}_2$  group binds to the receptor site in the pocket of the enzyme in two points—by oxygen of the  $\text{SO}_2$  fragment and by hydrogen bonding to the  $\text{NH}_2$  fragment. The existence of multisites as responsible for the receptor/ligand binding has been proposed in several hypothesized pharmacophore models, for example, in the binding of indole derivatives



**Figure 2.** Observed versus predicted log  $K$  values for benzenesulfonamide compounds obtained from a LOO CV analysis.

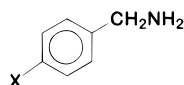


**Figure 3.** Representation of the  $r^2$  vs  $q^2$  statistical coefficients obtained from a random reordering test for the set of 29 benzenesulfonamides. Real (●) and random (+) QSAR models.

to benzodiazepine receptor. More detailed discussion of this phenomenon, together with its possible consequences for the construction of theoretical QSAR models, will be given in section c. As it will be shown there, the systematic NSS procedure permits the localization of individual interaction sites responsible for the biological activity in this particular case.

To verify the predictive power of the reported QSAR model corresponding to eq 24, a correlation between observed and predicted values of the inhibition constant log  $K$  is shown in Figure 2. The predicted values are computed employing a LOO CV analysis, yielding a  $q^2$  value of 0.976. Additionally, Figure 3 shows the results for a random reordering test of the vector containing HCA inhibition activities. As can be observed from this illustration, the correct arrangement of the vector

**Chart 2.** Common Molecular Structure for Benzylamine Derivatives



**Table 3.** Inhibitor Constants for the Binding of X-C<sub>6</sub>H<sub>4</sub>CH<sub>2</sub>NH<sub>2</sub> to the Enzyme Trypsin

	X	observed log 1/K <sub>i</sub> <sup>a</sup>
1	H	0.523
2	CH <sub>3</sub>	-0.176
3	Cl	0.155
4	OCH <sub>3</sub>	0
5	OCH <sub>2</sub> C <sub>5</sub> H <sub>6</sub>	0.398
6	NH <sub>2</sub>	0.301
7	COOH	-0.301
8	COOCH <sub>3</sub>	-0.362
9	COOCH <sub>2</sub> CH <sub>3</sub>	-0.447
10	COO(CH <sub>2</sub> ) <sub>2</sub> CH <sub>3</sub>	-0.301
11	COO(CH <sub>2</sub> ) <sub>3</sub> CH <sub>3</sub>	-0.041
12	COO(CH <sub>2</sub> ) <sub>4</sub> CH <sub>3</sub>	0.155
13	COO(CH <sub>2</sub> ) <sub>5</sub> CH <sub>3</sub>	0.523
14	COOCH <sub>2</sub> -C <sub>6</sub> H <sub>5</sub>	1.523
15	COOCH <sub>2</sub> -p-C <sub>6</sub> H <sub>4</sub> Cl	1.523
16	COO(CH <sub>2</sub> ) <sub>2</sub> C <sub>6</sub> H <sub>5</sub>	0.222
17	COO(CH <sub>2</sub> ) <sub>3</sub> C <sub>6</sub> H <sub>5</sub>	0.301
18	CONH <sub>2</sub>	-0.398
19	CONHC <sub>6</sub> H <sub>5</sub>	0.699
20	CONHCH <sub>2</sub> C <sub>6</sub> H <sub>5</sub>	0.398
21	CONH(CH <sub>2</sub> ) <sub>2</sub> C <sub>6</sub> H <sub>5</sub>	0.523
22	CONHC <sub>10</sub> H <sub>7</sub> (naphthalene)	1

<sup>a</sup> From ref 42.

containing activity values, which is depicted with a filled circle, corresponds to the best QSAR model.

In fact, to conclude this first example, it can be said that a satisfactory correlation between QS-SM and the binding constant *K* to HCA for this set of 29 benzenesulfonamides was found by means of eq 24. Results from these studies are comparable with those in a previous classical QSAR study. However, it must be stressed that in the present QS-SM model, which uses the same number of variables as the Hansch model, no mixing between Boolean and real variables is needed; in addition, the role of a pharmacophore position can easily be guessed, and the search for best fragment similarities is fully automated.

**(b) Benzylamine Derivatives.** In this example, the QSM theoretical approach was used to study the action of 22 benzylamine derivatives (Chart 2) as competitive inhibitors of the proteolytic enzyme trypsin. The activity of these derivatives was studied by Markwardt et al.,<sup>42</sup> and the corresponding data are summarized in Table 3. The biological activity of this series of substituted benzylamines was quantitatively studied by Hansch,<sup>31</sup> who reported the existence of empirical correlation with log *P* and Hammett  $\sigma$  constants as descriptors of a subset of nine molecules (**8–17** except phenyl derivative **14**):

$$\log 1/K_i = 0.41 \log P - 0.45\sigma - 1.07$$

$$n = 9; \quad r^2 = 0.955 \quad (25)$$

Such a form of empirical QSAR suggests again constructing the alternative theoretical QS-SM model in such a way that both traditional descriptors are replaced by their corresponding theoretical counterparts  $\theta_{AA}$  and  $\theta_{AA}^{CH_2NH_2}$ , respectively, listed in Table 4. Using

this replacement, the initial empirical equation can be rewritten in the form

$$\log 1/K_i = 0.6685\theta_{AA} - 3.7872\theta_{AA}^{SO_2NH_2} + 2.6124$$

$$n = 9; \quad r^2 = 0.964; \quad q^2 = 0.901 \quad (26)$$

which has a statistical importance similar to eq 25.

The ability to reproduce alternatively the traditional QSAR models is interesting, but certainly not the most important result of the present QS-SM approach. Its main advantage over traditional approaches is that the required theoretical QS-SM descriptors can be calculated easily. This is also true in situations where the traditional descriptors are either difficult to determine or completely unknown (for example,  $\sigma$  constants for some special substituents). Such is the situation with the whole series of 22 substituted benzylamines, where the lack of traditional descriptors restricted Hansch to studying only the subset of nine substituted derivatives. This limitation does not exist for the present theoretical approach and, in fact, a fairly good QSAR model describing the activity of the series of 21 derivatives has been found. It should be noted that the strongly deviating point **14** was excluded from the model, as it was by Hansch.<sup>31</sup> The resulting QSAR model takes the form of three-parameter correlation, with an additional descriptor, the fragment QS-SM  $\theta_{AA}^{C_6H_4}$ :

$$\log 1/K_i = 0.5572\theta_{AA} - 0.2608\theta_{AA}^{CH_2NH_2} +$$

$$0.2771\theta_{AA}^{C_6H_4} + 0.2220$$

$$n = 21; \quad r^2 = 0.828; \quad q^2 = 0.689 \quad (27)$$

The reason for the presence of this additional parameter is not yet completely clear. However, a plausible explanation could be proposed invoking possible specific interactions of the benzene ring with the enzyme cavity, which can become important in determining inhibition activity. This equation enables the experimental values of biological activity to be confronted with LOO CV values, as shown in Figure 4.

Figure 5 shows the results for the random reordering test performed over the set of 21 molecules in order to reject accidental correlation. As can be seen in this figure, the best model, with the highest values of  $r^2$  and  $q^2$ , does indeed correspond to the correct arrangement of log 1/*K<sub>i</sub>* values, so that accidental correlation can be excluded.

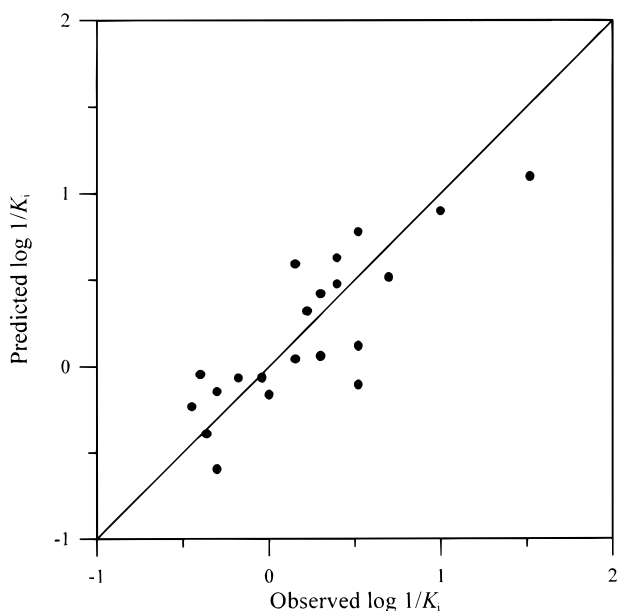
**(c) Indole Derivatives.** In this last example, a quantitative study of the relationships between the structure of a group of indole derivatives and their capacity to displace [<sup>3</sup>H] flunitrazepam from binding to bovine cortical membranes is presented. Molecular structures and experimental biological activity<sup>43</sup> for a set of 23 *N*-(indol-3-ylglyoxylyl)benzylamine derivatives (Chart 3) are listed in Table 5. These data were analyzed using the traditional QSAR approach,<sup>32</sup> providing some linear correlations with the main result that the biological activity for this molecular set does not depend on the hydrophobic parameter log *P*. In view of this, the decisive role in influencing biological activity in this series of substituted indole derivatives is presumably to be played by the substituent-induced variation in electronic structure of the active fragment, whose



**Table 4.** QS-SM ( $Z_{AA}$ ) and Scaled QS-SM ( $\theta_{AA}$ ) Used To Derive Eqs 26 and 27 for the Binding of X-C<sub>6</sub>H<sub>4</sub>CH<sub>2</sub>NH<sub>2</sub> to the Enzyme Trypsin<sup>a</sup>

	$Z_{AA}$	$\theta_{AA}$	$Z_{AA}^{CH_2NH_2}$	$\theta_{AA}^{CH_2NH_2}$	$Z_{AA}^{C_6H_4}$	$\theta_{AA}^{C_6H_4}$
1	134.9215	-1.87870	44.9282	-0.59640	89.8533	3.35123
2	149.9472	-1.70216	44.9221	-0.90111	89.4114	1.45386
3	301.8829	0.08296	44.9265	-0.68188	89.1607	0.37739
4	197.2198	-1.14674	44.8990	-2.03526	88.8875	-0.79540
5	286.0141	-0.10348	44.8990	-2.03561	88.8637	-0.89796
6	164.5553	-1.53053	44.8916	-2.40006	89.1915	0.50949
7	246.9712	-0.56221	44.9583	0.88469	88.8978	-0.75155
8	260.8403	-0.39925	44.9559	0.76648	88.9174	-0.66717
9	276.3086	-0.21751	44.9553	0.73528	88.9282	-0.62070
10	291.5922	-0.03795	44.9554	0.73765	88.9283	-0.62062
11	306.8786	0.14166	44.9553	0.73563	88.9283	-0.62057
12	322.1767	0.32140	44.9542	0.68066	88.9276	-0.62353
13	337.4751	0.50114	44.9542	0.68066	88.9275	-0.62366
14	349.6811	0.64455	44.9563	0.78464	88.9144	-0.67993
15	516.6186	2.60593	44.9558	0.75723	88.9015	-0.73549
16	365.1546	0.82635	44.9560	0.76939	88.9268	-0.62683
17	380.4667	1.00626	44.9560	0.76796	88.9245	-0.63660
18	228.6429	-0.77755	44.9422	0.09151	89.2038	0.56246
19	315.9116	0.24779	44.9416	0.05874	89.2277	0.66494
20	331.6791	0.43304	44.9425	0.10554	89.2149	0.61030
21	347.0398	0.61352	44.9411	0.03719	89.2330	0.68766
22	374.1008	0.93147	44.9415	0.05707	89.2318	0.68267

<sup>a</sup> Standardized values are obtained from eq 4.

**Figure 4.** Observed versus predicted  $\log 1/K_i$  values for benzylamine compounds obtained from a LOO CV analysis.

structure is not known. In keeping with this expectation, Hadjipavlou-Litina and Hansch proposed the correlation of the biological activity for a set of 20 derivatives (points **2**, **13**, and **14** were excluded) with the Hammett substituent constant  $\sigma$  of the substituent R:<sup>32</sup>

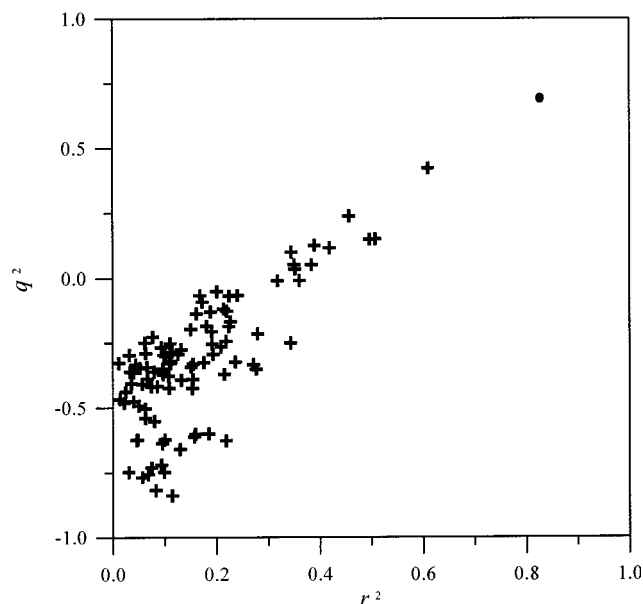
$$\log 1/K_i = 1.00\sigma + 6.60$$

$$n = 20; \quad r^2 = 0.498 \quad (28)$$

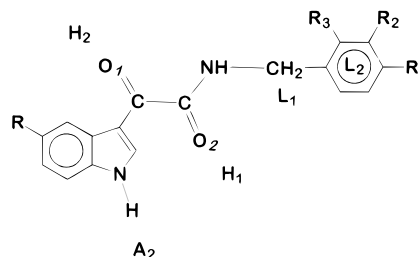
This correlation is not very satisfactory, but the graphical form of this dependence, shown in Figure 6, suggests that the description of the activity of the indole derivatives could be improved by adding two Boolean variables,  $I_2$  and  $I_3$ , as was proposed by Hadjipavlou-Litina and Hansch.<sup>32</sup>

$$\log 1/K_i = 1.01\sigma + 0.60I_2 - 0.40I_3 + 6.56$$

$$n = 20; \quad r^2 = 0.810 \quad (29)$$

**Figure 5.** Representation of the  $r^2$  vs  $q^2$  statistical coefficients obtained from a random reordering test for the set of 21 benzylamines. Real (●) and random (+) QSAR models.

### Chart 3. Common Molecular Structure for Indole Derivatives



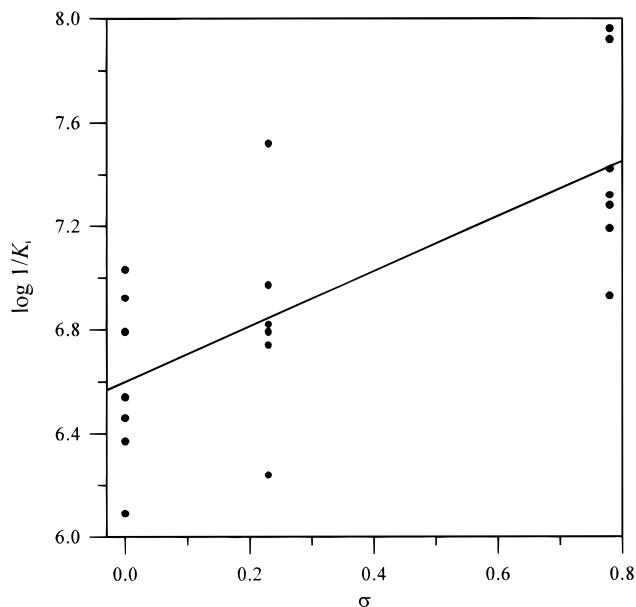
In this equation,  $\sigma$  is related to the substituent R, Boolean variable  $I_2$  is defined as 1 when both  $R_1$  and  $R_2$  are the CH<sub>3</sub>O group and as 0 otherwise, while  $I_3 = 1$  for the cases  $R_2 = OH/R_1 = H$  and 0 otherwise.

Although eq 29 provides a reasonable description of the biological data, it is not completely satisfactory from a theoretical point of view. Namely, it is clear that,

**Table 5.** Benzodiazepine Receptor Affinity of Indole Derivatives

	R	R <sub>1</sub>	R <sub>2</sub>	R <sub>3</sub>	observed log 1/K <sub>i</sub> <sup>a</sup>
1	H	H	H	H	6.92
2	Cl	H	H	H	6.31
3	NO <sub>2</sub>	H	H	H	6.93
4	H	OCH <sub>3</sub>	H	H	6.79
5	Cl	OCH <sub>3</sub>	H	H	6.97
6	NO <sub>2</sub>	OCH <sub>3</sub>	H	H	7.28
7	H	H	OCH <sub>3</sub>	H	6.54
8	Cl	H	OCH <sub>3</sub>	H	6.79
9	NO <sub>2</sub>	H	OCH <sub>3</sub>	H	7.42
10	H	OCH <sub>3</sub>	OCH <sub>3</sub>	H	7.03
11	Cl	OCH <sub>3</sub>	OCH <sub>3</sub>	H	7.52
12	NO <sub>2</sub>	OCH <sub>3</sub>	OCH <sub>3</sub>	H	7.96
13	H	Cl	H	H	7.17
14	H	H	H	Cl	5.59
15	H	OH	H	H	6.37
16	Cl	OH	H	H	6.82
17	NO <sub>2</sub>	OH	H	H	7.92
18	H	H	OH	H	6.09
19	Cl	H	OH	H	6.24
20	NO <sub>2</sub>	H	OH	H	7.19
21	H	OH	OH	H	6.46
22	Cl	OH	OH	H	6.74
23	NO <sub>2</sub>	OH	OH	H	7.32

<sup>a</sup> From ref 43.



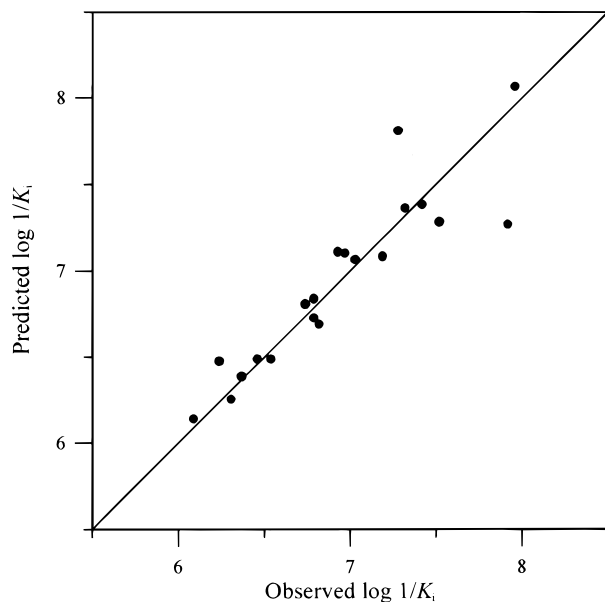
**Figure 6.** Dependence of Hammett constant  $\sigma$  for a series of indole derivatives on observed  $\log 1/K_i$  values.

whatever the biologically active fragment may be, its structure will certainly be affected by the cumulative effect of all substituents (R, R<sub>1</sub>, and R<sub>2</sub>). Thus, it would be much more realistic to seek the QSAR model in the form of a linear combination of  $\sigma$  constants for all substituents, rather than focusing on the effect of the single isolated substituent R. In analyzing such possible equations, an excellent correlation was found between the linear combination of  $\sigma$  constants of substituents R, R<sub>1</sub>, and R<sub>2</sub> and the fragment QS-SM  $\theta_{AA}^{COCONHCH_2}$

$$\theta_{AA}^{COCONHCH_2} = -3.0455\sigma_p(R) - 0.1714\sigma_p(R_1) - 0.5838\sigma_m(R_2) + 0.9458$$

$$n = 23; \quad r^2 = 0.991; \quad q^2 = 0.987 \quad (30)$$

which indicates that the biological activity of these



**Figure 7.** Observed versus predicted  $\log 1/K_i$  values for indole derivatives obtained from a LOO CV analysis.

molecules could be due to the presence of the COCONHCH<sub>2</sub> fragment. But the correlation of experimental activity with this selected theoretical descriptor is not very satisfactory:

$$\log 1/K_i = -0.3900\theta_{AA}^{COCONHCH_2} + 6.8856$$

$$n = 23; \quad r^2 = 0.491; \quad q^2 = 0.393 \quad (31)$$

This result shows that the problem of determining the fragment responsible for biological activity is more complex. A solution to this problem seems to be offered by a recent study,<sup>44</sup> in which a mechanism of benzodiazepine receptor (BzR) activity was proposed. According to this study, the biological activity of BzR ligands relies on the presence of four interaction sites: (i) a hydrogen bond acceptor site (A<sub>2</sub>), (ii) a hydrogen bond donor site (H<sub>1</sub>), (iii) a "bifunctional" hydrogen bond donor/acceptor site (H<sub>2</sub>/A<sub>3</sub>), and (iv) three lipophilic pockets (L<sub>1</sub>, L<sub>2</sub>, and L<sub>3</sub>). In particular, for the series of *N*-(indol-3-ylglyoxylyl)benzylamine derivatives, these (re)active sites were identified as<sup>43</sup> follows: A<sub>2</sub> = NH indole group, H<sub>1</sub> = C=O<sub>2</sub> group, H<sub>2</sub> = C=O<sub>1</sub> group, L<sub>1</sub> = CH<sub>2</sub> group, and L<sub>2</sub> = phenyl ring. All these crucial molecular fragments are depicted in Chart 3. It is worth mentioning that three of these sites are also present in the COCONHCH<sub>2</sub> fragment.

According to the previously described pharmacophore model, five QS-SM fragments were selected and tested as possible molecular descriptors for modeling the action of indole derivatives to the BzR:  $\theta_{AA}^{NH}$ (A<sub>2</sub>) = QS-SM for the indole group NH,  $\theta_{AA}^{C=O_2}$ (H<sub>1</sub>) = QS-SM for the group C=O<sub>2</sub>,  $\theta_{AA}^{C=O_1}$ (H<sub>2</sub>) = QS-SM for the group C=O<sub>1</sub>,  $\theta_{AA}^{CH_2}$ (L<sub>1</sub>) = QS-SM for the group CH<sub>2</sub>, and  $\theta_{AA}^{Ph}$ (L<sub>2</sub>) = QS-SM for the phenyl ring plus R<sub>1</sub>, R<sub>2</sub>, and R<sub>3</sub> substituents.

The corresponding QS-SM are listed in Table 6. On the basis of the above list, the theoretical QSAR model was searched for in the form of a linear combination of theoretical descriptors corresponding to individual interaction sites which give the best statistical description

**Table 6.** QS-SM ( $Z_{AA}$ ) and Scaled QS-SM ( $\theta_{AA}$ ) Used To Derive Eqs 30–33 for the BzR Affinity of Indole Derivatives

	$Z_{AA}^{COCONHCH_2}$	$\theta_{AA}^{COCONHCH_2}$	$Z_{AA}^{NH}$	$\theta_{AA}^{NH}$ <sup>a</sup>	$Z_{AA}^{C=O_2}$	$\theta_{AA}^{C=O_2}$ <sup>a</sup>	$Z_{AA}^{C=O_1}$	$\theta_{AA}^{C=O_1}$ <sup>a</sup>	$Z_{AA}^{CH_2}$	$\theta_{AA}^{CH_2}$ <sup>a</sup>	$Z_{AA}^{Ph}$	$\theta_{AA}^{Ph}$ <sup>a</sup>
1	169.8903	0.88904	28.7523	1.32205	63.3717	0.42105	61.8345	0.55964	14.0570	0.72942	89.7885	-1.75575
2	169.8128	0.19186	28.7319	-1.21936	63.3677	0.20432	61.7738	-0.11068	14.0590	0.84100	89.7804	-1.75593
3	169.6365	-1.39478	28.7407	-0.12213	63.3725	0.46295	61.6323	-1.67314	14.0650	1.17221	89.7585	-1.75641
4	169.8982	0.96021	28.7523	1.32005	63.3928	1.57312	61.8559	0.79688	14.0155	-1.55821	152.1434	-0.37822
5	169.8338	0.38107	28.7330	-1.07917	63.3970	1.80317	61.8034	0.21634	14.0189	-1.37179	152.1318	-0.37847
6	169.6342	-1.41534	28.7407	-0.12238	63.3876	1.28778	61.6547	-1.42621	14.0243	-1.07726	152.1108	-0.37894
7	169.8915	0.90029	28.7523	1.32043	63.3677	0.20416	61.8301	0.51089	14.0621	1.01270	152.1492	-0.37809
8	169.8245	0.29744	28.7330	-1.08104	63.3720	0.43580	61.7744	-0.10384	14.0648	1.15974	152.1419	-0.37825
9	169.6223	-1.52232	28.7407	-0.12325	63.3590	-0.27081	61.6264	-1.73895	14.0750	1.72239	152.1219	-0.37869
10	169.8694	0.70160	28.7523	1.32255	63.3749	0.59325	61.8392	0.61172	14.0272	-0.91306	213.7586	0.98298
11	169.7975	0.05443	28.7330	-1.07854	63.3741	0.54976	61.7832	-0.00648	14.0292	-0.80324	213.7489	0.98277
12	169.6082	-1.64916	28.7407	-0.12362	63.3719	0.43165	61.6377	-1.61417	14.0366	-0.39542	213.7180	0.98208
13	169.8838	0.83044	28.7433	0.20721	63.3219	-2.29971	61.8627	0.87116	14.0551	0.62545	256.7471	1.93268
14	169.9046	1.01837	28.7522	1.31568	63.3489	-0.82597	61.9016	1.30078	14.0599	0.88910	257.1127	1.94075
15	169.9007	0.98283	28.7500	1.03828	63.3813	0.94529	61.8694	0.94556	14.0196	-1.33638	138.1157	-0.68811
16	169.8339	0.38151	28.7308	-1.35969	63.3840	1.09236	61.8134	0.32730	14.0209	-1.26456	138.1048	-0.68836
17	169.6404	-1.35953	28.7385	-0.40240	63.3780	0.76457	61.6653	-1.30869	14.0248	-1.04521	138.0710	-0.68910
18	169.8903	0.88955	28.7500	1.04415	63.3540	-0.54369	61.8386	0.60486	14.0662	1.23641	138.1384	-0.68761
19	169.8191	0.24885	28.7308	-1.35881	63.3551	-0.48802	61.7804	-0.03794	14.0682	1.34496	138.1280	-0.68784
20	169.6321	-1.43456	28.7384	-0.40365	63.3513	-0.69142	61.6364	-1.62845	14.0756	1.75498	138.1006	-0.68845
21	169.9051	1.02277	28.7478	0.76201	63.3627	-0.07107	61.8726	0.98134	14.0350	-0.48704	186.5875	0.38272
22	169.8369	0.40908	28.7286	-1.63746	63.3647	0.03835	61.8159	0.35504	14.0363	-0.41506	186.5741	0.38242
23	169.6377	-1.38365	28.7362	-0.68279	63.3521	-0.65142	61.6685	-1.27304	14.0416	-0.11827	186.5384	0.38163

<sup>a</sup> Standardized similarity measure values are obtained from eq 4 and have been calculated together with QS-SM of molecules listed in Table 7.

**Table 7.** Benzodiazepine Receptor Affinity of Indole Derivatives

	R	R <sub>1</sub>	R <sub>2</sub>	R <sub>3</sub>	observed log 1/ <i>K</i> <sub>1</sub> <sup>a</sup>	predicted log 1/ <i>K</i> <sub>1</sub> <sup>b</sup>
24	H	H	Cl	H	6.80	6.52
25	H	F	H	H	7.28	5.82
26	H	H	H	F	6.18	5.77
27	H	OH	OCH <sub>3</sub>	H	6.85	6.74
28	Cl	OH	OCH <sub>3</sub>	H	7.57	7.04
29	NO <sub>2</sub>	OH	OCH <sub>3</sub>	H	7.89	7.67

<sup>a</sup> From ref 43. <sup>b</sup> Calculated from eq 33 using scaled QS-SM described in Table 8.

of the data. Such a search can best be performed by means of the NSS algorithm. With this approach, the following QSAR models were obtained for a set of 20 compounds, where molecules **1**, **13**, and **14** are rejected:

$$\log 1/K_1 = -0.4086\theta_{AA}^{C=O_1} + 0.3541\theta_{AA}^{Ph} + 6.9237$$

$$n = 20; \quad r^2 = 0.751; \quad q^2 = 0.683 \quad (32)$$

$$\log 1/K_1 = 0.2767\theta_{AA}^{C=O_2} - 0.4573\theta_{AA}^{C=O_1} + 0.3697\theta_{AA}^{Ph} + 6.8085$$

$$n = 20; \quad r^2 = 0.886; \quad q^2 = 0.823 \quad (33)$$

Not only do these equations provide a better statistical description of the biological activity than the original Hansch equation (eq 29) but also the physical meaning of individual descriptors is clearer than that for the Boolean variables *I*<sub>2</sub> and *I*<sub>3</sub>. In addition, eq 33 suggests

that the importance of potential interaction sites in determining biological activity is not the same for all fragments. The interactions for the hydrogen bond donor sites H<sub>1</sub>, H<sub>2</sub>, and L<sub>2</sub> probably dominate the rest. In conclusion, the presented theoretical procedure not only provides a satisfactory statistical description of the biological activities but also permits the localization and identification of the possible reaction sites responsible for the biological activity in a given series of compounds. Consequently, the reported theoretical approach can be a reasonable and reliable alternative to classical QSAR approaches.

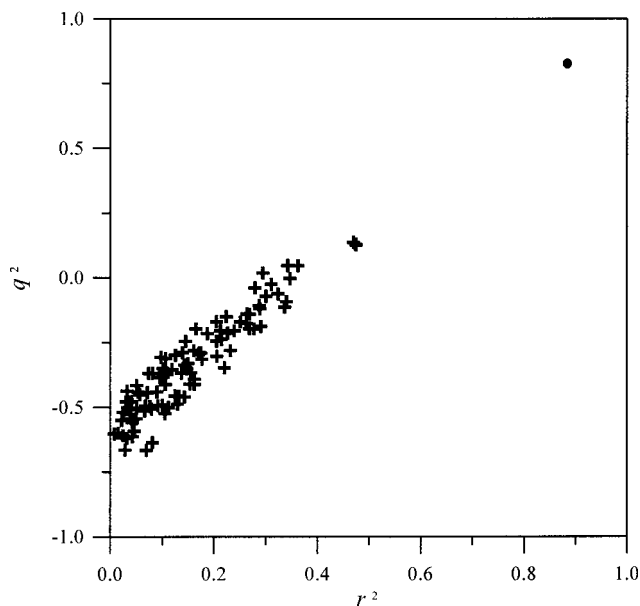
The predictive *q*<sup>2</sup> value obtained from a LOO CV analysis in eq 33 is satisfactory. This assertion is confirmed by the representation of predicted and observed values of log 1/*K*<sub>1</sub> for the set of 20 indole derivatives, shown in Figure 7.

Finally, an independent validation study of QSAR model<sup>33</sup> was carried out with the aim of predicting the activity for a new subset of compounds which were not considered in the construction of the model. In reference 43, the activities of six new indole derivatives (listed in Table 7) with the same common skeleton given in Chart 3 where reported. Fragment QS-SM for these compounds are listed in Table 8. Using eq 33, the theoretical values of log 1/*K*<sub>1</sub> were calculated which are also presented in Table 7. As can be seen, most of the predictions are correct. The only deviation concerns the molecule **25**, which has a fluorine atom as a substituent. However, when predicted and observed values for the

**Table 8.** Benzodiazepine Receptor Affinity of Indole Derivatives

	$Z_{AA}^{NH}$	$\theta_{AA}^{NH}$ <sup>a</sup>	$Z_{AA}^{C=O_2}$	$\theta_{AA}^{C=O_2}$ <sup>a</sup>	$Z_{AA}^{C=O_1}$	$\theta_{AA}^{C=O_1}$ <sup>a</sup>	$Z_{AA}^{CH_2}$	$\theta_{AA}^{CH_2}$ <sup>a</sup>	$Z_{AA}^{Ph}$	$\theta_{AA}^{Ph}$ <sup>a</sup>
24	28.7433	0.20521	63.3266	-2.04092	61.8707	0.95939	14.0535	0.43682	256.8588	1.93514
25	28.7433	0.20209	63.3308	-1.81120	61.8700	0.95187	14.0378	-0.32852	162.4339	-0.15088
26	28.7533	1.45239	63.3359	-1.53684	61.8946	1.22355	14.0377	-0.33514	162.5210	-0.14895
27	28.7500	1.04265	63.3682	0.22929	61.8610	0.85266	14.0309	-0.71003	200.4366	0.68867
28	28.7308	-1.35657	63.3693	0.28769	61.8043	0.22669	14.0329	-0.60049	200.4240	0.68839
29	28.7384	-0.40390	63.3623	-0.09347	61.6594	-1.37409	14.0390	-0.26553	200.3974	0.68781

<sup>a</sup> Standardized similarity measures values are obtained from eq 4 and have been calculated together with QS-SM of molecules listed in Table 5.



**Figure 8.** Representation of the  $r^2$  vs  $q^2$  statistical coefficients obtained from a random reordering test for the set of 20 indole derivatives. Real (●) and random (+) QSAR models.

rest of the five compounds are analyzed, a squared regression coefficient of 0.945 is obtained.

As in previous examples, a randomization test was carried out to estimate statistical reliability of the QSAR model given in eq 33. This validation test is presented in Figure 8. As can be seen, only the correct arrangement of biological data provides a satisfactory QSAR model.

## Conclusions

The examples put forward here clearly show that QS-SM for the whole molecule and the appropriate molecular fragments can advantageously be used as efficient descriptors for predicting biological and pharmacological activities. We can thus believe that because of its relative simplicity and complete generality the method opens a new interesting possibility to enrich the traditional QSAR approaches by describing a systematic procedure of constructing new theoretical QSAR models. Moreover, the possibility of identification of individual interaction sites responsible in each particular case for the observed biological activity could also be of considerable importance for the rational design of new biologically active molecules.

**Acknowledgment.** R.P. acknowledges the support of this work by the grant of the Grant Agency of the Czech Republic No. 203/00/1289. The research was also partly funded by a CICYT grant (SAF 96-0158), the Fundació Maria Francisca de Roviralta, and an European Commission contract (#ENV4-CT97-0508). The authors also thank the referees for their constructive criticism, which improved several aspects of this work.

## References

- 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.; Besalú, E.; 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.
- Amat, L.; Carbó-Dorca, R.; Ponec, R. Molecular Quantum Similarity Measures as an Alternative to Log  $P$  Values in QSAR Studies. *J. Comput. Chem.* **1998**, *19*, 1575–1583.
- Ponec, R.; Amat, L.; Carbó-Dorca, R. Molecular basis of quantitative structure-properties relationships (QSPR): A quantum similarity approach. *J. Comput.-Aided Mol. Des.* **1999**, *13*, 259–270.
- Ponec, R.; Amat, L.; Carbó-Dorca, R. Quantum Similarity Approach to LFER: Substituent and Solvent Effects on the Acidities of Carboxylic Acids. *J. Phys. Org. Chem.* **1999**, *12*, 447–454.
- Good, A. C.; Hodgkin, E. E.; Richards, W. G. Similarity screening of molecular data sets. *J. Comput.-Aided Mol. Des.* **1992**, *6*, 513–520.
- 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.
- Cooper, D. L.; Allan, N. L. A novel approach to molecular similarity. *J. Comput.-Aided Mol. Des.* **1989**, *3*, 253–259.
- Measures, P. T.; Mort, K. A.; Allan, N. L.; Cooper, D. L. Applications of momentum-space similarity. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 331–340.
- Benigni, R.; Cotta-Ramusino, M.; Giorgi, F.; Gallo, G. Molecular similarity matrixes and quantitative structure–activity relationships: a case study with methodological implications. *J. Med. Chem.* **1995**, *38*, 629–635.
- Mestres, J.; Rohrer, D. C.; Maggiora, G. M. A molecular field-based similarity approach to pharmacophoric pattern recognition. *J. Mol. Graphics Modelling* **1997**, *15*, 114–121.
- Mestres, J.; Rohrer, D. C.; Maggiora, G. M. A molecular-field-based similarity study of nonnucleoside HIV-1 reverse transcriptase inhibitors. *J. Comput. Aided-Mol. Des.* **1999**, *13*, 79–93.
- Carbó, R.; Leyda, L.; 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.
- Carbó, R.; Domingo, L. LCAO-MO Similarity Measures and Taxonomy. *Int. J. Quantum Chem.* **1987**, *23*, 517–545.
- Carbó, R.; Calabuig, B. Quantum molecular similarity measures and the  $n$ -dimensional representation of a molecular set: phenyldimethylthiazines. *J. Mol. Struct. (THEOCHEM)* **1992**, *254*, 517–531.
- Carbó, R.; Calabuig, B.; Vera, L.; Besalú, E. Molecular Quantum Similarity: Theoretical Framework, Ordering Principles, and Visualization Techniques. *Adv. Quantum Chem.* **1994**, *25*, 253–313.
- Besalú, E.; Carbó, R.; Mestres, J.; Solà, M. Foundations and Recent Developments on Molecular Quantum Similarity. *Topics Curr. Chem.* **1995**, *173*, 31–62.
- Molecular Similarity and Reactivity: From Quantum Chemical to Phenomenological Approaches*; Carbó, R., Ed.; Kluwer Academic: Amsterdam, 1995.
- Advances in Molecular Similarity*; Carbó-Dorca, R., Mezey, P. G., Eds.; JAI Press Inc.: Greenwich, CT, 1996; Vol. 1. and 1998; Vol. 2.
- 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 (QSPR) *J. Math. Chem.* **1995**, *18*, 237–246.
- Constans, 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-diamminedichloroplatinum(II) complex as an Application Example. *J. Comput. Chem.* **1999**, *20*, 911–920.



- (27) Carbó-Dorca, R.; Besalú, E. A general survey of molecular quantum similarity. *J. Mol. Struct. (THEOCHEM)* **1998**, *451*, 11–23.
- (28) Constans, 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.
- (29) Mezey, P. G. The Holographic Electron Density Theorem and Quantum Similarity Measures. *Mol. Phys.* **1999**, *96*, 169–178.
- (30) Hansch, C.; McClarin, J.; Klein, T.; Langridge, R. A Quantitative Structure–Activity Relationship and Molecular Graphics Study of Carbonic Anhydrase Inhibitors. *Mol. Pharmacol.* **1985**, *27*, 493–498.
- (31) Hansch, C.; Leo, A. *Exploring QSAR. Fundamentals and Applications in Chemistry and Biology*; American Chemical Society: Washington, DC, 1995.
- (32) Hadjipavlou-Litina, D.; Hansch, C. Quantitative Structure–Activity Relationships of the Benzodiazepines. A Review and Reevaluation. *Chem. Rev.* **1994**, *94*, 1483–1505.
- (33) ASA coefficients and exponents can be seen and downloaded from the following WWW site: <http://iqc.udg.es/cat/similarity/ASA/funcset.html>.
- (34) Hansch, C.; Fujita, T.  $\rho$ - $\sigma$ - $\pi$  Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *J. Am. Chem. Soc.* **1964**, *86*, 1616–1626.
- (35) Fujita, T.; Iwasa, J.; Hansch, C. A New Substituent Constant,  $\pi$ , Derived from Partition Coefficients. *J. Am. Chem. Soc.* **1964**, *86*, 5175–5180.
- (36) AMPAC 6.01, Semichem, Inc., 7128 Summit, Shawnee, KS 66216. D.A.
- (37) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. AM1: A New General Purpose Quantum Mechanical Molecular Model. *J. Am. Chem. Soc.* **1985**, *107*, 3902–3909.
- (38) Clementi, S.; Wold, S. How to Choose the Proper Statistical Method. In: *Chemometric Methods in Molecular Design*; van de Waterbeemd, H., Ed.; VCH Publishers Inc.: Weinheim, 1995; Vol. 2; pp 319–338.
- (39) Carbó, R.; Besalú, E. Definition, mathematical examples and quantum chemical applications of nested summation symbols and logical Kronecker deltas. *Comput. Chem.* **1994**, *18*, 117–126.
- (40) Carbó, R.; Besalú, E. Definition and quantum chemical applications of nested summations symbols and logical functions: Pedagogical artificial intelligence devices for formulae writing, sequential programming and automatic parallel implementation. *J. Math. Chem.* **1995**, *18*, 37–72.
- (41) Wold, S.; Eriksson, L. Statistical Validation of QSAR Results. In: *Chemometric Methods in Molecular Design*; van de Waterbeemd, H., Ed.; VCH Publishers Inc.: Weinheim, 1995; Vol. 2; pp 309–318.
- (42) Markwart, F.; Landmann, H.; Walsmann, P. Comparative Studies on the Inhibition of Trypsin, Plasmin, and Thrombin by Derivatives of Benzylamine and Benzamidine. *Eur. J. Biochem.* **1968**, *6*, 502–506.
- (43) Da Settimo, A.; Primofiore, G.; Da Settimo, F.; Marini, A. M.; Novellino, E.; Greco, G.; Martini, C.; Giannaccini, G.; Lucacchini, A. Synthesis, Structure–Activity Relationships, and Molecular Modeling Studies of *N*-(Indole-3-ylglyoxylyl)benzylamine Derivatives Acting at the Benzodiazepine Receptor. *J. Med. Chem.* **1996**, *39*, 5083–5091.
- (44) Zhang, W.; Koehler, K. F.; Zhang, P.; Cook, J. M. Development of a Comprehensive Pharmacophore Model for the Benzodiazepine Receptor. *Drug Des. Discovery* **1995**, *12*, 193–248.

JM9910728

## 7.5 Identificació dels fragments moleculars responsables de l'activitat biològica

No sempre les interaccions entre el fàrmac i el receptor són prou senzilles per ser caracteritzades mitjançant una equació *MLR* amb un sol paràmetre, ni tampoc és usual conèixer a priori els fragments moleculars transcendents d'una determinada propietat. El següent avenç en els càlculs de *QS-SM* de fragments ha estat proposar un mètode general capaç d'identificar les regions característiques d'una sèrie homogènia de compostos que millor descriuen una propietat molecular, sense cap restricció o especificació imposada a priori. El procés desenvolupat permet la detecció de les regions moleculars comunes a tota la sèrie molecular que són responsables d'una alta resposta biològica, i permet així dissenyar un patró on se senyalitzen les regions actives. El principal avantatge de l'aproximació de *QS-SM* de fragments respecte als estudis de *MQSM* que s'han mostrat en el capítol 6 és la supressió del procés de selecció de la superposició molecular, perquè les distribucions electròniques comparades pertanyen a una mateixa molècula.

### 7.5.1 Generalització del mètode *QS-SM* de fragments

El procediment consta bàsicament de tres estadis. En el primer es determinen els fragments moleculars i es calculen les corresponents *QS-SM*. El mètode és únicament aplicable en sèries de compostos derivats d'una estructura molecular comuna. Donat un conjunt de molècules amb un esquelet comú i uns substituents variables que són característics de cada producte de la sèrie, s'ha codificat un programa que identifica de manera automàtica tots els possibles fragments moleculars definits per un nombre determinat d'àtoms consecutius pertanyent al nucli comú i en determina els respectius valors de *QS-SM*. El segon procés consisteix a generar tots els models *QSAR* utilitzant en qualitat de descriptors moleculars les *QS-SM* dels fragments definits anteriorment. De tots els models se seleccionen els que donen un coeficient de predicció satisfactori segons uns barems establerts en la bibliografia, i la resta es descarten. Finalment, es representa en un histograma la freqüència que apareixen els àtoms de l'esquelet comú en les *QS-SM* de fragment dels models *QSAR* seleccionats. Els àtoms amb major

aparença en els histogrames són els que tenen més rellevància en la descripció de la propietat.

Amb l'objectiu de fer general l'aproximació, tot seguit es descriuen els processos comuns que es repeteixen en qualsevol anàlisi *QS-SM* de fragments.

**Definició dels fragments moleculars i les *QS-SM* relacionades.** La primera fase és la identificació d'una base estructural comuna en el conjunt de compostos estudiats, que de moment és l'únic procés que necessita de la intervenció externa de l'usuari. Després, a partir de l'esquelet comú, es defineixen tots els fragments moleculars normalment formats per un, dos i tres àtoms consecutius de manera automàtica. Els àtoms d'hidrogen no es consideren individualment com a simples fragments d'un sol àtom perquè donen un valor molt petit de *QS-SM*, el qual és difícil de diferenciar de possibles errors en el càlcul. El programa també calcula la *QS-SM* de la molècula entera i permet tenir en compte altres fragments d'interès, com ara anells dins l'estructura comuna o els fragments moleculars corresponents als substituents que varien en cada molècula.

**Selecció dels descriptors moleculars.** Una vegada calculats els valors de les *QS-SM* dels  $m$  fragments definits per les  $n$  molècules existents, s'agrupen en una matriu,  $\mathbf{Z}$ , de dimensió  $(n \times m)$ . Llavors per cada nombre  $k$  de paràmetres de la regressió multilinear que es vol analitzar es generen totes les possibles combinacions de  $m$  fragments amb  $k$  descriptors mitjançant un algorisme *NSS*, i es guarden els fragments de les equacions que donen el valor del coeficient de predicció de la validació creuada superior al límit establert. Malgrat no existir una unanimitat sobre quin hauria de ser el valor de la divisòria entre model seleccionat i model refusat, un bon criteri és escollir els models *MLR* amb coeficient  $r_{cv}$  superior o igual a 0.75, essent  $r_{cv}$  el coeficient de correlació entre el vector de propietats observades  $\mathbf{y}$  i el de predites  $\hat{\mathbf{y}}$ . Una altra possibilitat és utilitzar criteris referents a la importància estadística dels models *MLR*, com és la probabilitat de correlar accidentalment variables aleatòries que ha estat definida en l'equació (5.27).

**Representacions gràfiques.** Finalment es representa en un histograma la freqüència dels àtoms de l'esquelet comú que apareixen en les *QS-SM* de fragment dels models *QSAR* seleccionats. Els àtoms amb major aparença mostraran una major rellevància en la descripció de la propietat. L'anàlisi qualitativa dels histogrames dóna una idea de quins són els àtoms més rellevants de l'activitat estudiada, i permet construir un patró estructural amb les regions moleculars actives senyalitzades. El resultat és un intent de predir les característiques estructurals que produeixen o són responsables de l'activitat biològica, sense especificacions a priori.

En l'article 7.3 es descriu detalladament tot el procediment i es presenten dos exemples il·lustratius. El primer fa referència a les propietats electròniques degudes a l'efecte dels substituents en els derivats de l'àcid benzoic. L'objectiu és corroborar que les *QS-SM* que millor descriuen la constant  $\sigma$  de Hammett són les derivades del fragment COOH. El segon exemple d'aplicació de la nova aproximació *QS-SM* de fragments ha estat el conjunt d'esteroides de Cramer o Tripos.<sup>49</sup> Igual que en el desenvolupament d'anteriors metodologies *QSAR*, el conjunt d'esteroides de Cramer ha servit de test dels nous descriptors moleculars.

### 7.5.2 Correlació de la constant $\sigma$ amb altres fragments moleculars

En l'apartat 7.3.1 s'ha estudiat la correlació entre la constant  $\sigma$  i la *QS-SM* del fragment COOH en diferents sèries de compostos químics. Els coeficients de regressió obtinguts permeten afirmar que existeix una relació lineal entre la constant empírica  $\sigma$  i els descriptors teòrics basats en *QS-SM*. Però no es pot assegurar que la *QS-SM* del fragment COOH sigui el millor descriptor per descriure la constant de Hammett, perquè podria haver altres fragments moleculars que correlessin millor amb els valors de  $\sigma$ . Seguint el raonament exposat, en el primer estudi presentat en l'article 7.3 s'ha calculat les regressions lineals entre  $\sigma$  i un ampli conjunt de *QS-SM* de fragments moleculars definits sobre l'estructura comuna dels derivats de l'àcid benzoic. S'han calculat les mesures  $Z_{II}^X$ , essent  $X$  qualsevol fragment format per un, dos o tres àtoms enllaçats de l'esquelet comú de la sèrie dels derivats de l'àcid benzoic, més el fragment COOH. De



manera similar a les anàlisis de l'apartat precedent, els hidrògens no s'han considerat com a fragment d'un sol àtom. En total s'han definit 44 fragments.

A diferència de l'article 7.1, s'han optimitzat les geometries dels derivats de l'àcid benzoic amb el mètode HF emprant el conjunt de funcions de base 3-21G\* existent en el programa GAUSSIAN.<sup>32</sup> La sèrie està composta per 12 molècules, corresponents als substituents llistats en la taula 7.2. Els càlculs que es presenten en l'article 7.3 corresponen a les mesures *QS-SM* HF/3-21G\* emprant els operadors de recobriment i de Coulomb. S'han determinat les rectes de regressió  $Z_{ii}^x/\sigma$  pels 44 fragments moleculars definits. Els resultats mostrats en l'article 7.3 confirmen que les *QS-SM* del fragment COOH o de subfragments que contenen algun dels seus àtoms, són els millors descriptors teòrics que expliquen l'efecte de la substitució sistemàtica en la sèrie d'àcids carboxílics. El fet que hi hagi més d'un fragment que pot descriure els efectes electrònics és una conseqüència del teorema hologràfic de la densitat electrònica.<sup>40</sup> Però també és interessant observar que no tots els fragments donen bones correlacions, sinó bàsicament els construïts a partir dels àtoms que formen el grup COOH. Es pot concloure que el mètode *QS-SM* de fragments és una aproximació útil per determinar les regions de la densitat electrònica on la predeterminada propietat molecular està ampliada.

### 7.5.3 Estudi de l'activitat dels esteroides de Cramer

La sèrie molecular elegida per validar la nova metodologia és una família molt coneguda entre els investigadors de l'àrea *QSAR*, l'anomenat conjunt d'esteroides de Cramer o de Tripos.<sup>49</sup> Amb el temps ha esdevingut un conjunt estàndard en la majoria d'aproximacions *QSAR* com ho evidencien els nombrosos treballs que es poden trobar en la bibliografia, algun dels quals se citen en l'article 7.3. La sèrie està formada per 31 esteroides que tenen afinitat pel receptor de l'enllaç corticosteroide de la globulina (*Corticosteroid-Binding Globulin receptor, CBG*). Les geometries de les molècules s'han optimitzat completament amb el mètode AM1<sup>36</sup> i el programa AMPAC,<sup>41</sup> i sobre l'estructura de mínima energia s'ha fet un càlcul puntual HF/3-21G\* amb el programa GAUSSIAN.<sup>32</sup> No és usual dur a terme anàlisis *QSAR* emprant funcions densitat *ab*

*initio*, però la no necessitat de determinar l'alineament molecular fa possible la realització de càlculs exactes.

Amb el mètode *QS-SM* de fragments s'ha generat un model *QSAR* estadísticament millor que els obtinguts en anteriors aproximacions basades en matrius de semblança quàntica,<sup>50,51</sup> que ha permès dilucidar quines regions moleculars són importants per descriure l'activitat biològica mitjançant les *QS-SM*. Un altre aspecte important a ressaltar és la correcta descripció del compost 21-metil-2a-fluorcortisol, identificat amb el número **31** en la *Figure 1* de l'article 7.3, amb els models derivats de les *QS-SM* de fragments. En la majoria de les aproximacions *QSAR* que han estudiat el conjunt d'esteroides de Cramer s'ha observat que el compost **31** presenta l'error relatiu més gran de tota la sèrie i es pot considerar un *outlier*. La desviació que experimenta aquesta molècula es pot atribuir a la presència d'un àtom de fluor en la seva estructura. Aquesta característica influeix clarament en l'afinitat amb l'enllaç *CBG*, doncs les estructures de les molècules **30** i **31** es diferencien únicament en l'àtom de fluor en canvi les seves activitats són completament oposades.

**Article 7.3**

---

**Autors:** *Lluís Amat, Emili Besalú, Ramon Carbó-Dorca, Robert Ponec.*

**Títol:** *Identification of active molecular sites using quantum-self-similarity measures*

**Revista:** *Journal of Chemical Information and Computer Sciences*

**Volum:** *41*      **Pàgines, inicial:** *978*    **final:** *991*    **Any:** *2001*

---

# Identification of Active Molecular Sites Using Quantum-Self-Similarity Measures

Lluís Amat, Emili Besalú, and Ramon Carbó-Dorca\*

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

Robert Ponec

Institute of Chemical Process Fundamentals, Czech Academy of Sciences, Prague 6,  
Suchbát 2, 165 02, Czech Republic

Received December 7, 2000

A novel approach to construct theoretical QSAR models is proposed. This technique, based on the systematic use of quantum similarity measures as theoretical molecular descriptors, opens the possibility to localize and to identify the position of the bioactive part of drug molecules in situations, where the nature of the pharmacophore is not known. To test the reliability of this new approach, the method has been applied to the study of steroids binding to corticosteroid-binding human globulin. The studied molecules involved the set of 31 Cramer's steroids, often used as a benchmark set in QSAR studies. It has been shown that theoretical QSAR models based on the present procedure are superior to those derived from alternative existing approaches. In addition, a new method to measure the statistical significance of multiparameter QSAR models is also proposed.

## INTRODUCTION

In the last two decades quantum similarity theory<sup>1–22</sup> has increasingly been applied as a new means of the construction of theoretical QSAR models.<sup>23–44</sup> Some molecular similarity techniques have been focused on characterizing biological mechanisms by means of identifying molecular fragments, which are important for recognition by the enzyme.<sup>31–34,38,39,42,43</sup> Our own approach to quantum similarity measures (QSM) have clearly shown that the appropriate theoretical measures correlate remarkably well with empirical molecular descriptors such as  $\log P^{30}$  or Hammett  $\sigma$  constants<sup>31,32</sup> and, consequently, can replace them in QSAR equations. Thus, for example, the Hammett  $\sigma$  constants were found to correlate with the quantum self-similarity measures (QS-SM) of the fragment COOH in a series of substituted aromatic carboxylic acids,<sup>31,32</sup> so that the theoretical LFER equations equivalent to empirical Hammett  $pK$  vs  $\sigma$  plots could be straightforwardly formulated. In a similar way it was also possible to obtain alternative theoretical QSAR models for the correlation of biological activities.<sup>33,34</sup> Although this approach proved to be useful for many systems, it is nevertheless true that it can be straightforwardly applied only to systems in which the bioactive part of the molecule (pharmacophore) is known beforehand, so that the identification of the appropriate theoretical descriptor is self-evident. This, however, is not often the case, and in this situation the design of the appropriate QSAR model, whether empirical or theoretical, is extremely difficult. Because of the importance of these models for rational drug design, several approaches were proposed in recent years in which the problem of identification of the unknown pharmacophore was addressed. Two most commonly used approaches, GRID<sup>45</sup> and CoMFA,<sup>46</sup> are based on the calculations of interaction energies at grid

points in the space surrounding the target structure. Two interaction energy contributions are computed: the steric by measures of Lennard-Jones potentials and the electrostatic based on the Coulomb potential. From these approaches other methods were derived, such as CoMSIA,<sup>47</sup> CoMMA,<sup>48</sup> or more recently SOMFA.<sup>49</sup>

Our aim in this study is to complement the above-mentioned techniques by a simple procedure based on the systematic use of the so-called fragment QS-SM as a source of new theoretical molecular descriptors. To demonstrate the basic idea of this approach, the method will be first applied to the dissociation of substituted benzoic acids, where the chemical process is clearly localized into the COOH group, and as it will be shown, this molecular fragment is also correctly identified as the active, reaction center by the present procedure. This molecular set was previously studied using semiempirical methods and only evaluating the correlation between the QS-SM of the fragment COOH and the  $\sigma$  constant.<sup>31</sup> Now, the study is extended to ab initio calculations, and several molecular fragments are analyzed in order to find the best ones which better correlates with the experimental values.

Based on this result, the approach is then applied to the study of the affinity of a series of 31 steroids interacting with the corticosteroid binding globuline (CBG). This steroid series, known also as the Cramer's set, is widely used as a benchmark for testing various theoretical QSAR models,<sup>23,25,36,46–63</sup> and it has also been used in two previous studies,<sup>23,25</sup> based on QSM. In the first work,<sup>23</sup> the theoretical QSAR models were derived from Topological Quantum Similarity Indices (TQSI) and Molecular Quantum Similarity Measures (MQSM), while in the second one,<sup>25</sup> the formalism was based on the so-called tuned MQSM derived from the incorporation of MQSM into the general convex set theory.<sup>19</sup> Within this approximation, two or more quantum similarity

\* Corresponding author fax: 34 972 418356; e-mail: director@iqc.udg.es.

matrices are combined so as to optimize the precision and the predicting power of the corresponding QSAR model. However, this approach is computationally demanding since the calculation of the individual elements of quantum similarity matrices  $Z_{AB}$  requires the optimization of the relative position of the corresponding molecular pairs  $A$  and  $B$ .<sup>17</sup> In the present approach, these computational demands are considerably reduced by using fragment intramolecular QS-SM as molecular descriptors which allows to avoid the problems of the molecular alignment completely.

### THEORETICAL

Although theoretical considerations underlying the introduction of the concept of QSM were sufficiently described in several previous studies, it is worthwhile to be reminded of briefly the basic ideas, which provide an appropriate theoretical background for the theory of empirical structure–activity relationships.<sup>18–22</sup> The crucial role in introducing the concept of QSM has played an effort in finding an appropriate theoretical tool for the quantitative exploitation of the old, intuitive, but for chemistry extremely fruitful idea that similar molecules also have similar properties. As electron distribution, characterized by the quantum chemically generated density function  $\rho(\mathbf{r})$ , is in fact the ultimate molecular descriptor, it seems quite natural to base the definition of the QSM on this simplest observable quantum chemical quantity. Within this approach, the QSM of two molecules  $A$  and  $B$ , described by the corresponding density functions  $\rho_A(\mathbf{r})$  and  $\rho_B(\mathbf{r})$ , can be quantitatively characterized by the value of the integral

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

where  $\Omega(\mathbf{r}_1, \mathbf{r}_2)$  is an arbitrary positive definite two-electron operator which acts in the formula as a general weighting factor. Several kinds of QSM can be introduced depending on the actual choice of the operator  $\Omega$ . Thus, for example, the identification of  $\Omega$  with the Dirac delta function reduces the general expression to the formula

$$Z_{AB} = \int \rho_A(\mathbf{r}) \rho_B(\mathbf{r}) d\mathbf{r} \quad (2)$$

which represents the so-called overlap-like QSM.<sup>1</sup>

Another possible option is to identify  $\Omega$  with the Coulomb repulsion term  $\mathbf{r}_{12}^{-1}$ , and in this case the general formula (1) transforms into the so-called Coulomb-like QSM:

$$Z_{AB} = \int \int \rho_A(\mathbf{r}_1) |\mathbf{r}_1 - \mathbf{r}_2|^{-1} \rho_B(\mathbf{r}_2) d\mathbf{r}_1 d\mathbf{r}_2 \quad (3)$$

Although it is in general possible to introduce also other types of QSM, only overlap-like and Coulomb-like similarity measures will be considered in this study.

**Connection between MQSM and QSAR.** One of the most interesting and the most important applications for the above introduced QSM consists of the possibility of their use in the design of theoretical QSAR models. This use is based on the possibility to transform the formula for the expectation value of the nondifferential operator  $\omega$  associated with the measured molecular property  $y_I$

$$y_I = \int \omega(\mathbf{r}) \rho_I(\mathbf{r}) d\mathbf{r} \quad (4)$$

into a discrete form, analogous to traditional multilinear QSAR equations.<sup>18–22</sup> This transformation relies on the concept of molecular similarity. The basic idea of this procedure will be briefly reminded below. For this purpose, imagine that we have a series of molecules  $\{A, B, \dots, N\}$  and the pairwise QSM  $Z_{IJ}$  for all possible pairs are calculated from the set. The quantities  $Z_{IJ}$  form the square  $n \times n$  matrix  $\mathbf{Z}$ , the so-called similarity matrix, in terms of which the formula (4) can be rewritten as

$$y_I = \sum_K c_K Z_{KI} \quad (5)$$

where  $c_K$  are the components of the vector representing the operator  $\omega$  in the discrete basis of density functions  $\{\rho_I\}$ . This equation, which represents in fact the discrete counterpart of the continuous formulation (4), can be regarded as the most general form of QSAR equations. Although the above formalism represents the most direct approach to the design of theoretical QSAR models, the straightforward application of this approach has one unpleasant side-effect—it is computationally very demanding. This is due to the fact that the values of the pairwise similarity measures  $Z_{IJ}$ , forming the molecular similarity matrix, depend on the distance and the mutual orientation of the corresponding molecules. This implies that, in order to get meaningful results, the position of the molecules has to be optimized so as to give the maximum similarity for each pair  $I$  and  $J$ .<sup>17</sup>

In this study we propose a new approach, which to a considerable extent reduces the computational requirements of the original formalism, while still retaining the sufficient accuracy of the corresponding to theoretical QSAR models. This approach avoids the lengthy process of molecular alignment by considering only the diagonal elements  $Z_{II}$  of the similarity matrix  $\mathbf{Z}$ . These elements, the so-called quantum self-similarity measures (QS-SM), play, within this simplified approach, the role of theoretical QSAR descriptors, and as it has been shown in several previous published papers,<sup>30–33</sup> these QS-SM can be successfully used as an alternative to traditional QSAR descriptors. Thus, for example, the QS-SM  $Z_{II}$  was found to correlate with the molecular hydrophobicity empirical descriptor, the log P.<sup>30</sup> Similarly, it was also possible to describe the substituent effect, traditionally characterized by the Hammett  $\sigma$  constant, by the so-called fragment QS-SM.<sup>31,32</sup>

**Fragment Self-Similarity Measures.** The definition of these QSM is analogous to the original formula (1) from which it differs only in that the electron densities of an appropriate fragment  $X$  are compared instead of the total densities of the whole molecule

$$Z_{II}^X(\Omega) = \int \int \rho_I^X(\mathbf{r}_1) \Omega(\mathbf{r}_1, \mathbf{r}_2) \rho_I^X(\mathbf{r}_2) d\mathbf{r}_1 d\mathbf{r}_2 \quad (6)$$

The choice of the appropriate fragment is always a matter of certain arbitrariness, but intuitively one feels that the best chance to describe properly the effect of the systematic structural variation on a certain molecular property is when the corresponding fragment QS-SM is associated with the functional group responsible in any given case for the observed property. The correctness of this intuition was clearly confirmed in several previous studies, where the original empirical QSAR models were substituted by the

corresponding to theoretical counterparts.<sup>30–33</sup> However, such a straightforward approach is applicable only for systems and processes, in which the fragment responsible for the observed activity is known. An example in this respect can be constituted by the study of the biological activity of substituted phenylisothiocyanates<sup>31</sup> and benzensulfonamides,<sup>33</sup> whose activity is most probably due to the presence of NCS or SO<sub>2</sub>NH<sub>2</sub> groups, respectively. Unfortunately such a situation is not often the case and our aim here is to propose a simple general method, allowing for the identification and the localization of the position of the active fragment in the molecule in any given case. The method is based on the generation of various molecular fragments and on the subsequent evaluation of the quality of all possible QSAR models, based on the use of the corresponding fragment QS-SM as theoretical molecular descriptors. In this initial development, the proposed technique only applies to molecules which present common structural features, but it can be adapted to structurally diverse molecules following some schemes given in the literature.<sup>42</sup>

The QS-SM QSAR models are generated using the following systematic procedure:

1. The set of considered molecular fragments is defined. In this study, the fragments were defined as a group of several (1, 2, or 3) neighboring atoms contributing to the molecular skeleton. The corresponding fragment electron densities,  $X$ , can then be obtained from the total electron density of the whole molecule  $I$

$$\rho_f(\mathbf{r}) = \sum_{\mu} \sum_v D_{\mu v} \chi_{\mu}^*(\mathbf{r}) \chi_v(\mathbf{r}) \quad (7)$$

by appropriately restricting the summations over the basis functions. Thus for example, if the molecular fragment  $X$  consists of atoms  $a_1, a_2, \dots, a_n$ , then the corresponding density is given by

$$\rho_f^X(\mathbf{r}) = \sum_{i=1}^n \sum_{\mu \in a_i} \sum_{v \in I} D_{\mu v} \chi_{\mu}^*(\mathbf{r}) \chi_v(\mathbf{r}) \quad (8)$$

Based on these densities, the associated fragment QS-SM are computed according to eq 6.

2. Having defined the set of  $m$  molecular fragments and the corresponding QS-SM for each fragment, the next step consists of the systematic evaluation of the quality of all possible QSAR models based on the fragment QS-SM as descriptors. Here it is necessary to stress that the generated QSAR models do not need to be only one-dimensional but can be constructed and analyzed using any of the available multilinear correlation equations, with the number of independent descriptors ranging from 1 to any selected value  $k$  ( $k < m$ ). The total number of  $k$ -parameter correlation equations emerging from this scheme is  $\binom{m}{k}$ . All these alternatives were systematically generated using a nested summation algorithm.<sup>64,65</sup> The quality of any individual correlation was characterized by the value of the regression coefficient  $r$ . In addition, and in order to estimate the predictive power of the model, cross-validation (CV) following the leave-one-out (LOO) scheme was performed. The representation of experimental values against the cross-validated values gives the cross-validation regression coefficient  $r_{cv}$ . Recently, in our laboratory, a new basic approach

to directly obtain  $r_{cv}$  has been employed. See the Appendix for more details.

**Evaluation of the Statistical Significance of Generated Theoretical QSAR Models.** As it was stressed above, the proposed method of detection and localization of active molecular fragments is based on the evaluation of the quality of all systematically generated QSAR models and on their subsequent selection. This would be a trivial problem if the compared QSAR models were based on the same number of parameters (descriptors), since in this case the quality of the correlation can unequivocally be evaluated by the value of the correlation coefficient  $r$ . However, such a simple evaluation is not applicable in the present case, since the above-reported systematic procedure always generates not only linear,  $k = 1$ , but also all possible  $k$ -parameter multilinear correlation equations for any selected value of  $k$ . It is apparent, that when comparing QSAR models which differ in the number of parameters, the comparison of correlation coefficients is useless. This is due to the fact that the inclusion of any new additional parameter into a QSAR model always increases the value of the correlation coefficient, but obviously such an increase does not necessary guarantee the increase of the statistical importance of the correlation. The solution of this important problem of QSAR analysis was recently addressed by one of us,<sup>66</sup> and the basic idea of this approach will be briefly summarized below.

Suppose a general multilinear correlation equation

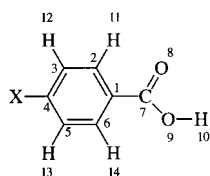
$$y_i = \sum_j a_j x_{ij} + b \quad | j \in 1, 2, \dots, k \wedge i \in 1, 2, \dots, n \quad (9)$$

is based on the set of variables  $\{y_1, y_2, \dots, y_n\}$   $\{x_{i1}, x_{i2}, \dots, x_{ik} | i = 1, 2, \dots, n\}$  and assume that the actually observed correlation coefficient is  $r$ . Then, instead of using the actual values of variables  $y_i$  and  $x_{ij}$ , one can analyze the same correlation using another set of randomly generated variables  $(\lambda_i, \nu_{ij})$ . It is clear, that the correlation coefficient  $R$  of this randomly generated correlation will be, with high probability, quite low. But, repeating the same random experiment many times, there is a certain (nonzero) probability  $P$  that the correlation coefficient of such an accidental randomly generated correlation will be equal to or greater than  $r$ . Such probability depends on the number of points  $n$ , on the value of the correlation coefficient  $r$ , and on the number of parameters  $k$ . It is clear that the lower the value of the probability  $P$  is, the more difficult it will be to obtain the correlation with  $R > r$  accidentally or, in other words, the higher is the statistical significance of the original correlation. This implies that the values of the probability  $P$  (or its negative logarithm  $pP$ ) can be used as a simple universal measure of the statistical significance of the correlations, irrespective of how many parameters they involve. This probability is closely related to the so-called confidence level of the correlation. The relation of these quantities is given by the formula

$$CL(\text{in } \%) = (1 - P)100 \quad (10)$$

The main goal of the study<sup>66</sup> in which the above criterion was introduced is that it was possible to propose a simple geometrical model allowing the analytical calculation of the corresponding probability for any values of  $n$ ,  $k$ , and  $r$ . This



**Scheme 1.** Numbering of Atoms for Benzoic Acids

probability is given by the formula

$$P = \frac{\int_0^{\arccos(r)} \cos^{k-1} \theta \sin^{n-k-2} \theta d\theta}{\int_0^{\pi/2} \cos^{k-1} \theta \sin^{n-k-2} \theta d\theta} \quad (11)$$

The numerical calculation of this probability is not difficult and can be done using any existing mathematical programs. Here, an original Fortran code has been used and can be obtained upon request. The quality of the empirical correlations can also be characterized by the value of the cross-validated correlation coefficient  $r_{cv}$ . In terms of this approach, acceptable correlations are characterized by  $r_{cv} > 0.75$ , and this simple, but widely used, criterion has been also used in this study.

**Computations.** Several kinds of calculation steps were followed in this study. The first of them consists of the quantum chemical generation of wave functions and electron densities for the studied series of molecules. These calculations were performed by the Hartree–Fock method using a 3-21G\* basis set for fully optimized molecular geometries. These calculations were carried out using Gaussian 98 software package.<sup>67</sup> In the next step, fragment densities for a series of systematically generated fragments were calculated using eq 8. Finally, the above generated data were employed for the calculation of the corresponding fragment QS-SM, which served afterward as descriptors for the construction of theoretical QSAR models. These models were systematically generated using a nested summation algorithm,<sup>64,65</sup> and the statistical parameters of these correlations were analyzed using the above-reported criteria.

## RESULTS AND DISCUSSION

**(A) Para-Substituted Benzoic Acids.** To convincingly demonstrate the ability of the present approach to detect and localize a molecular fragment responsible, in a given case, for the observed activity, the results of the application of the above formalism to the dissociation of substituted para-substituted benzoic acids are reported first. It is apparent that the active molecular fragment in this example is the COOH group, and, as it will be shown, the systematic search of the best theoretical descriptor also confirms this intuitive expectation. The application of the above-reported approach will be described in detail next.

**1. Molecular Set.** Twelve para-substituted benzoic acids, with substituents, are as follows: NO<sub>2</sub>, CN, CF<sub>3</sub>, CCl<sub>3</sub>, Br, Cl, F, H, CH<sub>3</sub>, CH<sub>2</sub>CH<sub>3</sub>, OCH<sub>3</sub>, and N(CH<sub>3</sub>)<sub>2</sub>.

**2. Fragments for the Calculation of QS-SM.** The classification of fragments used for the construction of theoretical QSAR descriptors is based on the numbering of individual atoms as shown in Scheme 1.

Based on this numbering, the set of the molecular fragments was arbitrarily selected according to the following strategy: (a) all monatomic fragments corresponding to

individual C and O atoms—nine fragments in total: C1–C7, O8, and O9; (b) all biatomic fragments between pairs of directly bonded atoms—14 fragments in total; (c) all triatomic fragments involving directly bonded triads of atoms—20 fragments in total; and (d) the fragment COOH. The total number of such generated fragments was  $m = 44$ .

**3. Selection of the Best Theoretical QSAR Model.** As the dissociation of substituted benzoic acids is empirically described by the Hammett equation, involving only one parameter, the  $\sigma$  constant, a restriction of only one-parameter QSAR models,  $k = 1$  was adopted. The theoretical QSAR models are based on the computation of the correlation between the  $\sigma$  constant and the QS-SM of each fragment. As the total number of generated theoretical descriptors,  $m$ , is 44, the total number of all generated and analyzed linear equations is also 44. The results of these calculations for QSAR models derived from overlap and Coulomb-like fragment QS-SM are summarized in Tables 1 and 2, respectively.

The first, and most important, conclusion resulting from Tables 1 and 2 is that the best QSAR models are indeed generated from QS-SM associated with COOH fragment or any of its subfragments. This fact is preferably observed for Coulomb QS-SM, presented in Table 2. In this respect, the obtained results confirm the correctness of the original expectation. However, the situation is slightly more complex. This is due to the fact that in addition to “expected” correlations with the QS-SM related to the COOH fragment, there is also a considerable number of QSAR models whose theoretical descriptors seem to be in no relation to this group, but whose precision is comparable or only slightly lower. This result seems to be unexpected since in traditional QSAR models there is usually just one empirical descriptor appropriate for the evaluation of the given property, and it can be hardly imagined that the same property could be described by so many different descriptors. To explain the above findings, it has to be realized that a strict specificity characteristic of the empirical parameters does not exist for theoretical descriptors, especially if they are based, as in this case, on quantities derived from electron density. The reason for this greater flexibility of quantum chemical QSAR descriptors, compared to classical ones, may be related to the recently formulated “holographic electron density theorem”.<sup>68</sup> This theorem states that any finite fragment of the electron density, considered to be the ultimate molecular descriptor, contains the same amount of structural information as the total electron density of the whole molecule. This implies that all possible fragments of electron density are equivalent in their information content, so that any of them could, in principle, be used for the generation of the appropriate theoretical descriptor. However, the actual situation is slightly less favorable, and some differences between individual fragments and their associated theoretical descriptors can be observed. This is due to the fact that although the holographic electron density theorem guarantees the same information content within any molecular fragment, it says nothing about how this information could be extracted. The present approach, based on the application of QS-SM, represents one of these possibilities. It is just here, in the selection of the particular method of extracting the structural information, where the differences between individual fragments enter into play.

**Table 1.** Molecular Fragments and Statistical Parameters of QSAR Models for the Dissociation of Para-Substituted Benzoic Acids Based on Overlap QS-SM

fragment <sup>a</sup>	r <sup>b</sup>	r <sub>cv</sub> <sup>c</sup>	CL <sup>d</sup>	pP <sup>e</sup>
O9H10	0.980	0.971	100.000	7.609
O9	0.974	0.961	100.000	7.079
C1	0.957	0.945	100.000	5.950
C7O8	0.948	0.905	100.000	5.555
C7O8O9	0.937	0.892	99.999	5.138
C2C1C6	0.911	0.885	99.996	4.411
C1C7O8	0.894	0.857	99.991	4.048
C7O8O9H10	0.893	0.835	99.991	4.029
C1C6	0.875	0.830	99.981	3.712
C2C1	0.873	0.806	99.979	3.687
C7	0.820	0.748	99.890	2.957
C2C1H11	0.808	0.712	99.854	2.836
O8	0.803	0.719	99.835	2.782
C1C6H14	0.776	0.698	99.701	2.525
C7O9	0.751	0.669	99.515	2.314
C2C1C7	0.665	0.514	98.180	1.740
C3C2H12	0.652	0.525	97.841	1.666
C3H12	0.563	0.365	94.316	1.245
C3C2H11	0.523	0.326	91.929	1.093
C1C7	0.475	0.058	88.092	0.924
C5H13	0.467	0.210	87.381	0.899
C1C6C7	0.459	-0.146	86.701	0.876
C1C6C5	0.451	-0.042	85.910	0.851
C3C2	0.422	0.121	82.860	0.766
C6C5H13	0.399	0.112	80.144	0.702
C3C2C4	0.382	0.024	77.968	0.657
C3C4	0.380	-0.017	77.665	0.651
C7O9H10	0.376	-0.379	77.125	0.641
C6C5C4	0.347	-0.053	73.027	0.569
C3	0.335	-0.061	71.342	0.543
C4	0.333	0.011	71.008	0.538
C6	0.318	-0.401	68.573	0.503
C3C4H12	0.311	-0.199	67.405	0.487
C5C4	0.304	-0.213	66.330	0.473
C1C7O9	0.303	-0.291	66.119	0.470
C2	0.279	-0.502	61.937	0.420
C3C2C1	0.267	-0.654	59.831	0.396
C3C5C4	0.253	-0.491	57.241	0.369
C5C4H13	0.207	-0.464	48.099	0.285
C5	0.164	-0.578	38.861	0.214
C6C5H14	0.140	-0.644	33.463	0.177
C6H14	0.128	-0.785	30.783	0.160
C2H11	0.123	-0.819	29.623	0.153
C6C5	0.016	-0.906	3.826	0.017

<sup>a</sup> For numbering see Scheme 1. <sup>b</sup> Standard correlation coefficient. <sup>c</sup> Cross-validation regression coefficient. Negative  $r_{cv}$  values close to -1 indicate a reverse relationship between observed and predicted properties. <sup>d</sup> Confidence level defined in eq 10. <sup>e</sup>  $pP = -\log P$ .

Nevertheless, the situation where a given molecular property can be correlated with so many theoretical descriptors raises necessarily the question of the interpretation of these individual correlations, especially with respect to the possibility to extract from them the information about the nature of the active molecular fragment. In the following part the basic idea of such an interpretation will be sketched.

For this purpose, first it is possible to consider the set of all generated QSAR models, and one can evaluate how many times each of the atoms contribute to the examined molecular fragments. Then, in the next step one can do the same but only for the set of fragments, which generated the QSAR models of satisfactory quality. The quality or the statistical significance of the correlations was evaluated using the confidence level (CL) and the cross-validated correlation coefficient  $r_{cv}$  criterion as commented previously. The correlations with  $CL > 99.9\%$ , which, in this case, roughly

**Table 2.** Molecular Fragments and Statistical Parameters of QSAR Models for the Dissociation of Para-Substituted Benzoic Acids Based on Coulomb QS-SM

fragment <sup>a</sup>	r <sup>b</sup>	r <sub>cv</sub> <sup>c</sup>	CL <sup>d</sup>	pP <sup>e</sup>
O8	0.992	0.987	100.000	9.636
O9H10	0.992	0.988	100.000	9.576
C7O8O9H10	0.983	0.977	100.000	8.006
C7O8	0.983	0.976	100.000	7.928
C7O8O9	0.982	0.975	100.000	7.847
C2C1C7	0.980	0.973	100.000	7.665
C1C7O8	0.976	0.962	100.000	7.225
C1C7	0.971	0.955	100.000	6.790
C1C6C7	0.968	0.950	100.000	6.597
C1C7O9	0.954	0.933	100.000	5.818
C2C1H11	0.946	0.930	100.000	5.477
C1C6H14	0.946	0.910	100.000	5.474
C1	0.938	0.914	99.999	5.172
C6C5H13	0.920	0.875	99.998	4.653
O9	0.919	0.888	99.998	4.617
C7O9H10	0.917	0.878	99.997	4.570
C3C2H12	0.909	0.852	99.996	4.384
C5H13	0.899	0.840	99.993	4.152
C3H12	0.895	0.815	99.992	4.073
C1C6	0.875	0.806	99.981	3.715
C2C1	0.868	0.821	99.975	3.599
C1C6C5	0.865	0.779	99.972	3.550
C3C2C1	0.864	0.746	99.971	3.532
C7	0.853	0.784	99.957	3.370
C6C5H14	0.845	0.737	99.946	3.264
C3C2H11	0.834	0.702	99.925	3.123
C6C5	0.764	0.592	99.617	2.417
C5	0.752	0.593	99.525	2.324
C3C2	0.749	0.537	99.491	2.294
C6	0.742	0.633	99.432	2.245
C2	0.732	0.637	99.324	2.170
C3	0.722	0.443	99.203	2.099
C2C1C6	0.674	0.562	98.370	1.788
C4	0.488	0.177	89.271	0.969
C3C2C4	0.442	0.034	84.931	0.822
C3C4	0.434	0.015	84.108	0.799
C6C5C4	0.429	-0.007	83.558	0.784
C5C4	0.412	-0.057	81.620	0.736
C3C4H12	0.359	-0.189	74.835	0.599
C5C4H13	0.335	-0.273	71.217	0.541
C3C5C4	0.334	-0.311	71.075	0.539
C7O9	0.291	-0.261	64.179	0.446
C6H14	0.158	-0.567	37.544	0.204
C2H11	0.110	-0.755	26.560	0.134

<sup>a</sup> For numbering see Scheme 1. <sup>b</sup> Standard correlation coefficient. <sup>c</sup> Cross-validation regression coefficient. Negative  $r_{cv}$  values close to -1 indicate a reverse relationship between observed and predicted properties. <sup>d</sup> Confidence level defined in eq 10. <sup>e</sup>  $pP = -\log P$ .

coincides with the criterion  $r_{cv} > 0.75$ , were considered as having acceptable statistical significance. The results of this analysis are summarized in Table 3.

Looking into Table 3 it is possible to see that some atoms contribute to fragments generating statistically significant correlations more often than others. A typical example in this respect is the carbon atom C1, which contributes to 13 fragments, from the whole considered set of 44. For Coulomb QS-SM, 12 times these fragments are associated with the descriptors yielding the statistically significant QSAR models. Similarly high frequency of participating in fragments, generating the statistically significant QSAR models, can be also observed for the carbon atom C7, oxygen atoms O8 and O9, and hydrogen atom H10. On the other hand, there are other atoms, like C3, C4, and C5, for which the corresponding frequency is much lower.



**Table 3.** Appearance of Individual Atoms in Selected QSAR Models of Dissociation of Substituted Benzoic Acids for Overlap and Coulomb QS-SM

atoms <sup>a</sup>	total ( $N_i$ ) <sup>b</sup>	$r_{cv} > 0.75$ <sup>c</sup>	$N_g/N_i$	CL > 99.9% <sup>d</sup>	$N_g/N_i$
Overlap QS-SM					
C1	13	5	0.385	5	0.385
C2	11	2	0.182	2	0.182
H11	3	0		0	
C3	10	0		0	
H12	3	0		0	
C4	8	0		0	
C5	10	0		0	
H13	3	0		0	
C6	11	2	0.182	2	0.182
H14	3	0		0	
C7	11	4	0.364	4	0.364
O8	5	4	0.800	4	0.800
O9	7	4	0.571	4	0.571
H10	3	2	0.667	2	0.667
Coulomb QS-SM					
C1	13	11	0.846	12	0.923
C2	11	4	0.364	6	0.545
H11	3	1	0.333	2	0.667
C3	10	2	0.200	4	0.400
H12	3	2	0.667	2	0.667
C4	8	0		0	
C5	10	3	0.300	4	0.400
H13	3	2	0.667	2	0.667
C6	11	5	0.455	6	0.545
H14	3	1	0.333	2	0.667
C7	11	10	0.909	10	0.909
O8	5	5	1.000	5	1.000
O9	7	6	0.857	6	0.857
H10	3	3	1.000	3	1.000

<sup>a</sup> For numbering see Scheme 1. <sup>b</sup> Appearance of each atom in all generated molecular fragments ( $N_i$ ). <sup>c</sup> Appearance of a given atom in statistically significant QSAR models satisfying the criterion  $r_{cv} > 0.75$  ( $N_g$ ). <sup>d</sup> Appearance of a given atom in statistically significant QSAR models satisfying the criterion  $CL > 99.9\%$  ( $N_g$ ).

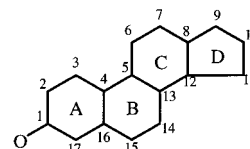
This result is very interesting since the atoms identified by the above frequency analysis exactly coincide with the atoms of COOH group, or its close neighbors, expected to be the active molecular fragment on the basis of classical chemical considerations. Based on this result, we report in the following part the application of the presented methodology to the study of the biological activity of Cramer's steroids.

**(B) Cramer Steroid Data Set.** This molecular set involves a series of 31 steroids whose biological activity is due to their affinity to corticosteroid binding globuline (CBG).<sup>23,25,36,46-63</sup> The structures of these molecules are depicted in Figure 1, and the corresponding biological activities are summarized in Table 4.

In keeping with the above introduced general methodology, the biological activity of this set of steroids was correlated with a series of systematically generated QSAR models based on fragment QS-SM as the corresponding to theoretical descriptors. The fragments considered for the generation of these descriptors were selected using the same protocol as in the previous case of benzoic acids. This involves, in the first step, the generation of the wave function and the density matrix at HF/3-21G\* level for the whole series of molecules. As the geometry optimization of such large molecules is time-consuming, the molecular geometries were optimized at semiempirical AM1 level<sup>69</sup> using AMPAC program.<sup>70</sup> The above optimized geometries were used in

**Table 4.** Cramer Steroids CBG Binding Affinity

Figure 1 no.	CBG (pK <sub>a</sub> )	Figure 1 no.	CBG (pK <sub>a</sub> )
1	-6.279	17	-5.225
2	-5.000	18	-5.000
3	-5.000	19	-7.380
4	-5.763	20	-7.740
5	-5.613	21	-6.724
6	-7.881	22	-7.512
7	-7.881	23	-7.553
8	-6.892	24	-6.779
9	-5.000	25	-7.200
10	-7.653	26	-6.144
11	-7.881	27	-6.247
12	-5.919	28	-7.120
13	-5.000	29	-6.817
14	-5.000	30	-7.688
15	-5.000	31	-5.797
16	-5.225		

**Scheme 2.** Numbering of Atoms for Cramer Steroids**Table 5.** Number of Statistically Significant QSAR Equations Satisfying the  $r_{cv}$  and Confidence Level Criteria for the Cramer Steroids Set

$k$	$\binom{m}{k}$	overlap		Coulomb	
		$r_{cv} > 0.75$	CL > 99.9%	$r_{cv} > 0.75$	CL > 99.9%
2	2485	502	1184	264	1354
3	57155	17820	37080	10587	40307
4	971635	386670	749690	243679	788624
5	13019909	5976443	11161540	3751968	11487055
6	143218999	71355131	130973686	42842979	132716856

the next step for the quantum chemical generation of HF/3-21G\* electron densities using Gaussian 98 software package.<sup>67</sup> Based on these data, the fragments were selected according to the following systematic procedure:

(a) All H atoms were excluded so that only the fragments involving heavy atoms (C, O) of the common molecular skeleton were considered. This skeleton is composed of 18 atoms: 17 are carbons forming the rings A, B, C and D, and the remaining one is the oxygen bonded to the carbon atom C1 of the ring A. The classification of the fragments is based on the numbering shown in Scheme 2.

(b) The following fragments were considered: (b1) all monoatomic fragments corresponding to individual C and O atoms—C1—C17 and O: 18 fragments in total; (b2) all biatomic fragments between the pairs of directly bonded atoms—21 fragments in total; (b3) all triatomic fragments involving directly bonded triads of atoms—31 fragments in total; and (b4) the whole basic skeleton. In this way, the total number of generated active fragments is  $m = 71$ .

The theoretical QSAR fragment models were generated for both overlap-like and Coulomb-like QS-SM as descriptors. These models were constructed in the form of multi-linear regression equations with the number of parameters  $k$  ranging from 2 to 6. The results of this systematic search are summarized in Table 5. As it is evident from the results of Table 5, the number of QSAR models to be considered rapidly increases with the number of parameters  $k$ . Moreover, the number of statistically significant QSAR models satisfy-



**Table 6.** Statistical Significance of Cramer Steroids QSAR Models for Different Number of Parameters  $k$ 

$k^a$	overlap			Coulomb		
	$r_{cv}$	$r$	pP	$r_{cv}$	$r$	pP
2	0.873	0.889	9.475	0.864	0.890	9.554
3	0.924	0.941	12.139	0.897	0.919	10.285
4	<b>0.941</b>	<b>0.958</b>	<b>12.935</b>	0.916	0.941	11.161
5	0.946	0.960	12.339	0.929	0.955	11.613
6	0.945	0.962	11.684	0.909	0.960	11.313
$6^b$	0.947	0.961	11.444	0.938	0.953	10.562

<sup>a</sup> The values in individual rows correspond to the QSAR model with highest  $r$  value for a given number of parameters  $k$ . For  $k = 2-5$ , they also are the QSAR models with highest  $r_{cv}$ . <sup>b</sup> QSAR model with highest  $r_{cv}$  using  $k = 6$  descriptors.

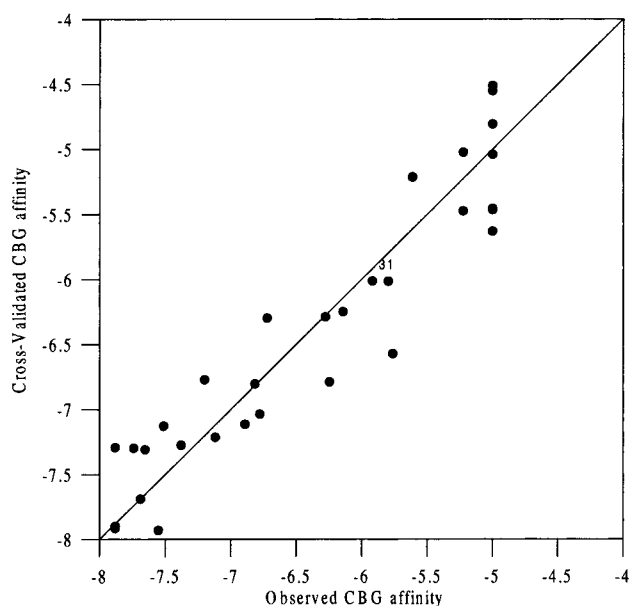
ing the  $r_{cv}$  ( $r_{cv} > 0.75$ ) or  $CL$  ( $CL > 99.9\%$ ) criterion promptly increases with  $k$ . The fact that the number of statistically significant models is bigger for  $CL$  than for  $r_{cv}$  criterion means that the selected confidence level 99.9% is, in this example, a less severe criterion than  $r_{cv} > 0.75$ . To reduce the number of statistically significant QSAR models so as to be comparable with the number resulting from  $r_{cv}$  criterion, it would be necessary to increase the  $CL$  to roughly 99.9999%.

In contrast to the simple form of linear one-parameter correlation equations, describing the dissociation of substituted benzoic acids, the set of statistically significant QSAR models, generated for the studied series of steroids, is not homogeneous and involves equations with the number of parameters  $k$  ranging from 2 to 6. This raises the question of the evaluation of the statistical significance of the observed correlations and, consequently, of the selection of the optimally significant ones. To overcome this problem, two independent methodologies were used. First of them is represented by the recently proposed analytical model,<sup>66</sup> based on the comparison of calculated probabilities of the random generation of the QSAR model with the same number of points  $N$ , parameters  $k$ , and the correlation coefficient  $r$  as the actually observed one. The results of such comparison are summarized in Table 6 from which it is evident, seeing the values of the negative logarithm of probability, pP, that the statistically most important is the four-parameter correlation equation based on overlap-like QS-SM. This theoretical model corresponds to the expression

$$y = -6.592 - 0.762 \times Z(\text{C17}) + 0.511 \times Z(\text{C12C13}) + 0.982 \times Z(\text{C13C14C15}) + 0.438 \times Z(\text{C8C12C11}) \quad (12)$$

A graphical representation of the actual CBG affinities versus the predicted ones for the four-parameter QSAR model (12) is shown in Figure 2, based on a LOO CV analysis. The main characteristic of this representation is that compound **31** becomes not an outlier, a normal feature stated in many QSAR studies.<sup>23,25,36,46-60,62,63</sup>

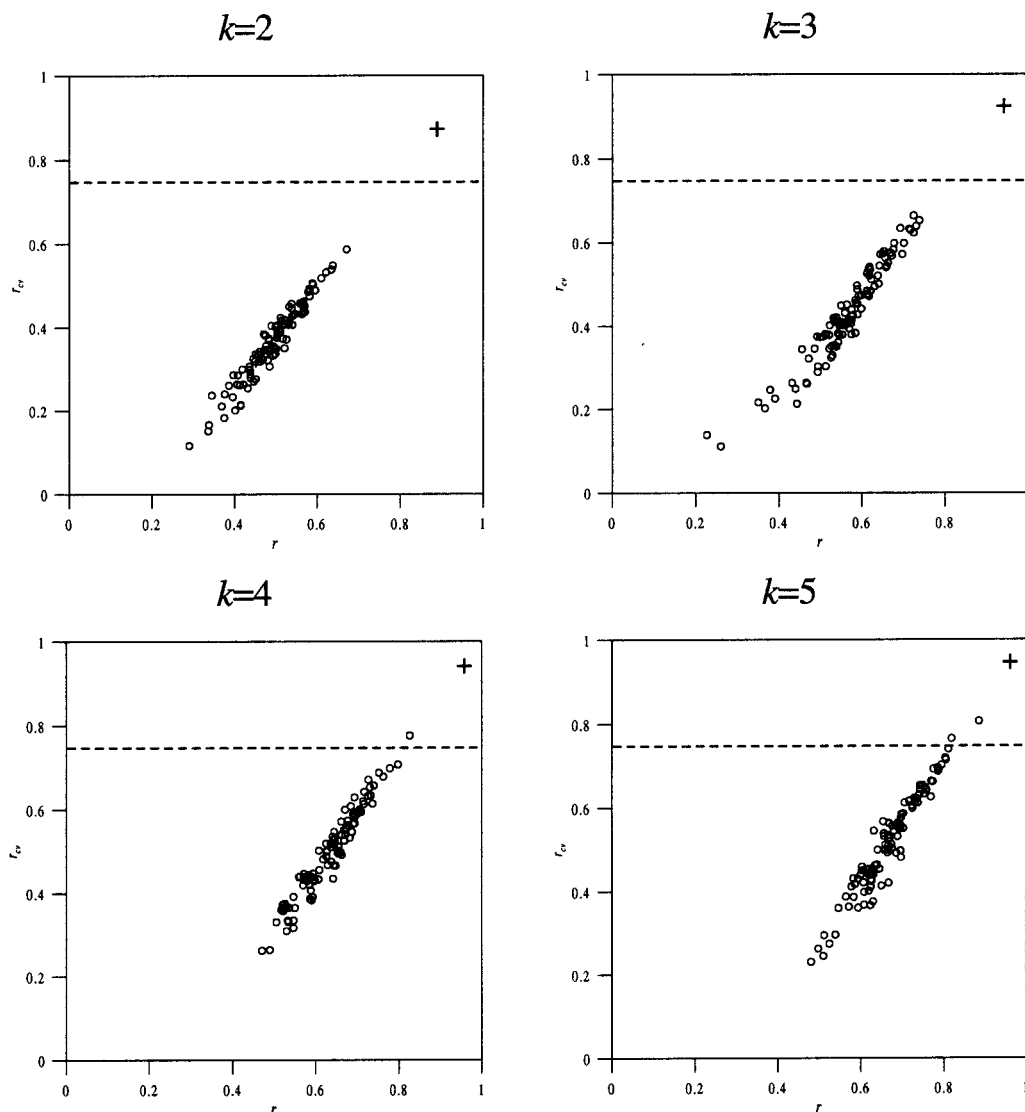
**Randomization Test Analysis.** To verify the conclusions of the analytical model, the evaluation of the statistical significance of the correlations was independently performed by the numerical randomization test. This test consists of randomly reorienting the set of observed CBG activities and in the subsequent determination of the optimal QSAR model based on fragment QS-SM which gives the maximal value

**Figure 2.** Cross-validated versus experimental CBG affinities for the Cramer steroids QSAR model (12).

of cross-validated correlation coefficient  $r_{cv}$  in a LOO analysis. Repeating this process many times, in our case we used 100 random runs, one obtains a set of  $r_{cv}$  values which are plotted against the ordinary correlation coefficient of the optimal QSAR model. The results of this randomization test are summarized in Figure 3 from which it is possible to see that the first possibility of random generation of statistically significant QSAR model ( $r_{cv} > 0.75$ ) is observed for  $k = 5$ .

**Comparison with Previous QSM Results on Steroid Data Set.** In connection with eq 12 it is perhaps worth mentioning that the same series of steroids was also studied in two previous QSM studies.<sup>23,25</sup> Table 7 shows the main results of the corresponding QSAR models. As it is possible to see from this table, the four-parameter correlation eq 12 obtained in this study is clearly superior to any related previously described QSAR model based on QSM. The best previous results were obtained for a tuned QSAR analysis, based on the combination of three quantum similarity matrices and a six-parameter model.<sup>25</sup> For this study, the  $q^{(2)}$  value<sup>71,72</sup> is 0.842, whereas the four-parameter eq 12 yields a value of  $r_{cv}^2 = 0.886$ .

**Identification of the Bioactive Molecular Fragment.** Although the basic philosophy of the localization of the bioactive molecular fragment is exactly the same as in the above analyzed case of substituted benzoic acids, the situation with the steroid data set is obviously much more complex. This is due to the fact that the total number of statistically significant theoretical QSAR models in this case is huge (see Table 5), and the statistical parameters of individual models are often quite close. In this situation, it is very difficult to base the selection of bioactive molecular fragment on only a few of the very best correlation equations, like eq 12, and in order to get reliable predictions, the whole set of statistically significant QSAR models has to be considered. For this purpose a simple universal procedure is proposed. It is based on the construction of histograms, depicting the frequency where each atom of the basic skeleton contributes to statistically significant QSAR models.



**Figure 3.** Numerical randomization test analysis for the Cramer steroids set. Multilinear regressions using overlap QS-SM and two up to five descriptors.

**Table 7.** Comparison of Statistical Significance of the Cramer Steroids QSAR Model (12) with Other Related QSM Models from Previous Studies

	TQSI <sup>a</sup>	MQSM <sup>b</sup>	MQSM <sup>c</sup>	tuned MQSM <sup>d</sup>	fragment QS-SM
$r_{cv}^2 q^{(2) e}$	0.775	0.705	0.759	0.842	0.886
$r^2$	0.837	0.781	0.833	0.903	0.917
pP	9.17	8.29	8.20	10.25	12.93
no. of PCs (k)	4	3	5	6	4

<sup>a</sup> QSAR study using topological quantum similarity indices. From ref 23. <sup>b</sup> QSAR study using simple quantum similarity matrices, PCA technique for variable reduction and no PCs selected. From ref 23. <sup>c</sup> Simple similarity matrices using classical scaling for variable reduction and selection of PCs. From ref 25. <sup>d</sup> Tuned QSAR model using a mixture of three similarity matrices. From ref 25. <sup>e</sup> In previous works<sup>23,25</sup>  $q^{(2)}$  has been used to determine the predictive power of the models.

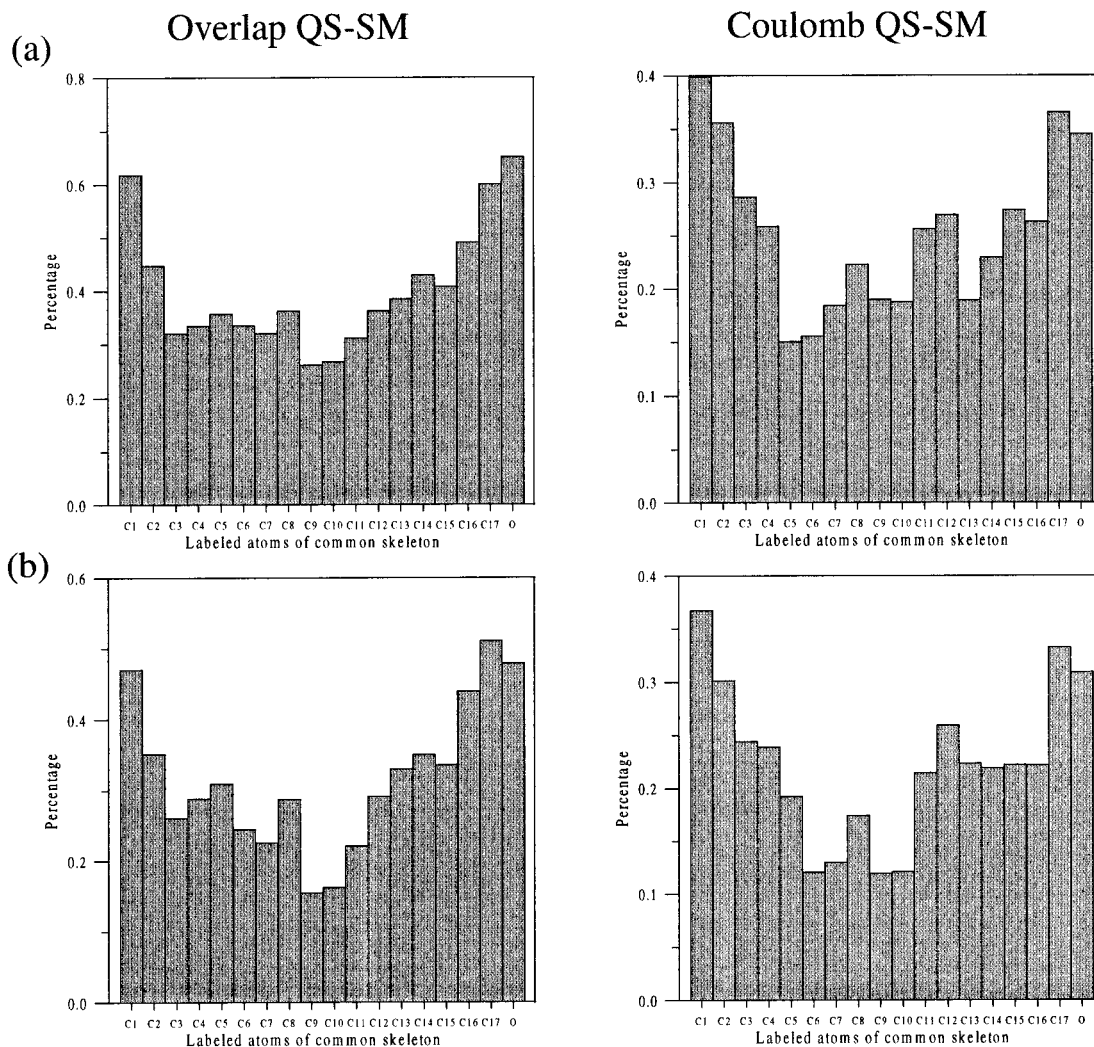
This frequency is for each individual atom defined as the ratio  $N_g/N_i$ , where  $N_g$  is the number of "favorable" cases in which a given atom contributes to statistically significant QSAR models and  $N_i$  is the total number of correlations involving a given atom.

The corresponding histograms based on the set of overlap-like and Coulomb-like fragment QS-SM are depicted in

Figure 4. This figure has been only constructed for QSAR models obtained using four-parameter multilinear equations. The selection of statistically significant QSAR models for the calculation of the corresponding frequencies was based on two criteria: (a)  $r_{cv} > 0.75$  and (b)  $CL > 99.9999\%$ . Then, the selected number of QSAR models which satisfy the above criteria are very similar. For example, 386670 and 307797, respectively, using overlap QS-SM.

As can be deduced from Figure 4, there is no significant difference between the histograms obtained from both criteria. A close parallelism is also observed for the histograms based on overlap-like and Coulomb-like similarity measures. Based on these histograms it is possible to identify the (bio)active part of the molecule with the fragment involving the atoms: C1, C2, C16, C17, O. This suggests that the biological activity of the Cramer set of steroids is very probably due to the presence of a carbonyl group bonded to the ring A of the basic skeleton. It is interesting that the set of most active steroids, the molecules **6**, **7**, **10**, **11**, **19**, **20**, **22**, **23**, **25**, **28**, and **30**, all possess this common structural feature. The main difference between overlap and Coulomb histograms is that for the last one, important contributions





**Figure 4.** Histograms for overlap and Coulomb QS-SM for Cramer steroids. QSAR models using four-parameter regression equations. Selected statistically significant models using criterions: (a)  $r_{cv} > 0.75$  and (b)  $CL > 99.9999\%$ .

can also be expected from the atoms C11 and especially C12 belonging to the five-membered ring D. Very similar conclusions result also from the use of SOMFA methodology,<sup>49</sup> which identifies two important areas in the steroids skeleton: a large area of negative potential around atom C1 and a large area of positive potential around the five-membered ring. To emphasize the proposed spatial regions determining the activity for the Cramer steroids series, two QSAR models have been computed, combining overlap QS-SM of fragment CCO in ring A and Coulomb QS-SM of atom C12:

$$y = -6.384 - 0.668 \times Z^{\text{ove}}(\text{C2C1O}) + 0.416 \times Z^{\text{cou}}(\text{C12})$$

$$n = 31, r = 0.887, r_{cv} = 0.858, pP = 9.375 \quad (13)$$

$$y = -6.384 - 0.668 \times Z^{\text{ove}}(\text{C17C1O}) + 0.416 \times Z^{\text{cou}}(\text{C12})$$

$$n = 31, r = 0.884, r_{cv} = 0.856, pP = 9.372 \quad (14)$$

Sound correlations have been obtained, comparable with the

optimally significant model presented in Table 6 for  $k = 2$  parameters. These elucidated fragments could be considered to be representative of the regions where the differences in the electron density preferentially influence the binding affinity of the steroids.

**Predictive Ability of the QSAR Models Associated to the Fragments Favoring Activity.** The information extracted from the above-mentioned histograms can be used for designing new inhibitors with unknown activity. But in this case, it could be more convenient to compare the predictive capacity of proposed models (13) and (14) with the conclusions of other related studies reported for the same series of steroids in the literature.<sup>25,36,46-63</sup> However, most of these studies do not use the whole set of 31 steroids, but, instead, the set of the first 21 molecules (**1-21**) is considered as the training set and the quality of the models is then assessed by comparison of their predictions for the test set of last molecules (**22-31**). The standard deviation of errors of prediction (SDEP) is employed as a coefficient to estimate the quality of the model, which is a root-mean-square error of the predictions:  $[\sum(y_{\text{pred}} - y_{\text{obs}})^2/n]^{1/2}$ . To make possible direct comparison with Cramer's set earlier studies, QSAR models (13) and (14) have been recalculated for the training

**Table 8.** Predicted Values of Cramer Steroids Test Set (22–31) for Different Approaches

steroid	actual activity	CoMFA (FFD) <sup>a</sup>	compass <sup>a</sup>	MS-WHIM <sup>a</sup>	SOMFA <sup>a</sup>	TQSAR <sup>b</sup>	fragment QS-SM
22	-7.512	-7.883	-7.062	-7.300	-7.279	-7.237	-7.036
23	-7.553	-7.430	-7.729	-8.332	-7.034	-7.879	-7.221
24	-6.779	-6.642	-6.462	-6.821	-6.925	-6.648	-7.023
25	-7.200	-7.705	-7.466	-7.445	-7.232	-7.809	-7.307
26	-6.144	-6.495	-5.994	-6.121	-5.744	-6.832	-6.345
27	-6.247	-6.962	-6.383	-6.901	-6.800	-7.318	-7.322
28	-7.120	-6.848	-6.625	-6.532	-6.603	-7.363	-7.536
29	-6.817	-6.816	-7.403	-6.838	-6.692	-7.540	-7.296
30	-7.688	-7.767	-7.741	-7.860	-7.345	-7.628	-7.264
31	-5.797	-7.793	-7.779	-7.491	-7.283	-7.537	-6.675
	SDEP	0.716	0.705	0.662	0.584	0.762	0.544
	SDEP <sup>c</sup>	0.356	0.339	0.411	0.367	0.555	0.493

<sup>a</sup> See ref 49, Table 4. <sup>b</sup> Reference 25. <sup>c</sup> Excludes steroid 31.

set of 21 steroids. The results of the corresponding analysis are

$$y = -6.454 - 0.726 \times Z^{\text{ove}}(\text{C1C2O}) + 0.401 \times Z^{\text{cou}}(\text{C12})$$

$$n = 21, r = 0.909, r_{cv} = 0.865, pP = 6.844, \text{SDEP} = 0.544 \quad (15)$$

$$y = -6.449 - 0.719 \times Z^{\text{ove}}(\text{C1C17O}) + 0.407 \times Z^{\text{cou}}(\text{C12})$$

$$n = 21, r = 0.908, r_{cv} = 0.865, pP = 6.817, \text{SDEP} = 0.550 \quad (16)$$

As can be seen, the coefficients  $r$  and  $r_{cv}$  increase with respect those of eqs 13 and 14, but the negative logarithm of the probability decreases because the number of molecules is has been reduced. A comparison of the predictive powers of fragment QS-SM and other QSAR approaches is given in Table 8.

It would be interesting to note that some involuntary mistake was performed in a previous study<sup>25</sup> on the same steroid Cramer set. Erroneous SDEP values were given for tuned MQSM, which now have been corrected in Table 8.

## CONCLUSIONS

In this study we report a new systematical procedure for the detection and localization of molecular fragments, responsible for the observed activity in a series of structurally related molecules. The approach, based on the use of QS-SM as new set of theoretical molecular descriptors, was applied to description of the CBG activity in the series of 31 Cramer's steroids. The results of our approach are equivalent, even better, in comparison to those obtained using other alternative approaches. In addition, a new methodology to measure the number of statistically significant multilinear regression parameters is proposed.

## APPENDIX

**Direct Computation of  $r_{cv}^2$  Coefficient in Multiple Linear LOO Procedures.** Given a  $n \times k$  descriptors matrix,  $\mathbf{X} = \{x_{ij}\}$ , and the vector containing the dependent parameters,  $\mathbf{y} = (y_1, y_2, \dots, y_n)^T$ , the vector collecting the

coefficients of the attached multilinear regression is

$$\mathbf{c} = [\mathbf{X}^T \mathbf{X}]^{-1} \mathbf{X}^T \mathbf{y} \quad (17)$$

Defining the prediction matrix

$$\mathbf{H} = \{h_{ij}\} = \mathbf{X}[\mathbf{X}^T \mathbf{X}]^{-1} \mathbf{X}^T \quad (18)$$

the dependent values fitted by the model are obtained by means of the product

$$\mathbf{y}' = (y'_1, y'_2, \dots, y'_n)^T = \mathbf{Xc} = \mathbf{Hy}$$

In this context it is defined as the coefficient of multiple determination:

$$r^2 = 1 - \frac{\sum_{p=1}^n (y_p - y'_p)^2}{\sum_{p=1}^n (y_p - \bar{y})^2} \quad (19)$$

Here,  $\bar{y}$  is the mean value of the observed variables. This term coincides with the correlation coefficient between the  $\mathbf{y}$  variables and the ones fitted by the model ( $\mathbf{y}'$ ).

Considering a standard process of cross-validation and collecting the cross-validated property values in the vector  $\hat{\mathbf{y}} = (\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n)^T$ . The hat over the variables indicates that they must be obtained for each molecule, say  $p$ , from a linear model fit constructed without considering the  $p$ th observation, that is, in a LOO procedure. It is customary to represent, in a bidimensional plot, the dependent cross-validated values ( $\hat{\mathbf{y}}$ ) against the experimental ones ( $\mathbf{y}$ ). By analogy with expression 19, an *estimation* of the correlation coefficient for the cross-validation procedure is

$$q^{(2)} = 1 - \frac{\sum_{p=1}^n (y_p - \hat{y}_p)^2}{\sum_{p=1}^n (y_p - \bar{y})^2} = 1 - \frac{PRESS}{S_{yy}}$$

This constitutes the usual definition for the  $q^{(2)}$  parameter. The two summations are identified with the prediction error of the sum of squares (*PRESS*) statistic<sup>73</sup> and the sum of quadratic errors from the mean value,  $S_{yy}$ .

The  $q^{(2)}$  parameter can be *negative*, as it is discussed in ref 74. That is the reason in this paper is preferred to use the  $q^{(2)}$  notation instead of the standard  $Q^2$  or  $q^2$  ones.

Nevertheless, it is straightforward to compute the correlation coefficient attached to the linear LOO procedure. In standard textbooks,<sup>73</sup> a demonstration is carried out allowing the computation of *PRESS* variable once the matrix of predictions has been obtained. The demonstration in ref 73 is exact, but it is focused into obtaining the  $\mathbf{y}_p - \hat{\mathbf{y}}_p$  differences. Here, a very similar algebraic procedure will be followed, but this time only the  $\hat{\mathbf{y}}_p$  values will be computed.

Defining the column vector  $\mathbf{x}_p = (x_{p1}, x_{p2}, \dots, x_{pm})^T$  as the one collecting the original independent descriptors for the molecule number  $p$  coming from the  $p$ th row of the  $\mathbf{X}$  matrix. Then, if the data of molecule  $p$  are eliminated, that is, the value  $y_p$  and the vector  $\mathbf{x}_p$  are set to zero, a new properties vector  $\mathbf{y}_{(p)}$  and descriptors matrix  $\mathbf{X}_{(p)}$  are being defined. Following a similar notation as in (17), the coefficients of the linear model are

$$\mathbf{c}_{(p)} = [\mathbf{X}_{(p)}^T \mathbf{X}_{(p)}]^{-1} \mathbf{X}_{(p)}^T \mathbf{y}_{(p)}$$

This vector allows the computation of the cross-validated property value for the molecule number  $p$ :

$$\hat{y}_p = \mathbf{x}_p^T \mathbf{c}_{(p)} = \mathbf{x}_p^T [\mathbf{X}_{(p)}^T \mathbf{X}_{(p)}]^{-1} \mathbf{X}_{(p)}^T \mathbf{y}_{(p)}$$

On the other hand, from (18), the prediction matrix elements are defined as

$$h_{ij} = \mathbf{x}_i^T [\mathbf{X}^T \mathbf{X}]^{-1} \mathbf{x}_j \quad (20)$$

and it is straightforward to demonstrate that the following relationship

$$[\mathbf{X}_{(p)}^T \mathbf{X}_{(p)}]^{-1} = [\mathbf{X}^T \mathbf{X}]^{-1} + \frac{[\mathbf{X}^T \mathbf{X}]^{-1} \mathbf{x}_p \mathbf{x}_p^T [\mathbf{X}^T \mathbf{X}]^{-1}}{1 - h_{pp}}$$

holds.<sup>73</sup> In this way, one is able to write

$$\hat{y}_p = \mathbf{x}_p^T \left\{ [\mathbf{X}^T \mathbf{X}]^{-1} + \frac{[\mathbf{X}^T \mathbf{X}]^{-1} \mathbf{x}_p \mathbf{x}_p^T [\mathbf{X}^T \mathbf{X}]^{-1}}{1 - h_{pp}} \right\} \mathbf{X}_{(p)}^T \mathbf{y}_{(p)} = \frac{\mathbf{x}_p^T [\mathbf{X}^T \mathbf{X}]^{-1} \mathbf{X}_{(p)}^T \mathbf{y}_{(p)}}{1 - h_{pp}}$$

Since  $\mathbf{X}^T \mathbf{y} = \mathbf{X}_{(p)}^T \mathbf{y}_{(p)} + \mathbf{x}_p \mathbf{y}_p$ , then

$$\hat{y}_p = \frac{\mathbf{x}_p^T [\mathbf{X}^T \mathbf{X}]^{-1} [\mathbf{X}^T \mathbf{y} - \mathbf{x}_p \mathbf{y}_p]}{1 - h_{pp}} = \frac{\mathbf{x}_p^T [\mathbf{X}^T \mathbf{X}]^{-1} \mathbf{X}^T \mathbf{y} - \mathbf{x}_p^T [\mathbf{X}^T \mathbf{X}]^{-1} \mathbf{x}_p \mathbf{y}_p}{1 - h_{pp}}$$

and from (17) and (20) it is easily obtained

$$\hat{y}_p = \frac{1}{1 - h_{pp}} \sum_{i=1}^n h_{pi} y_i$$

The correlation coefficient between the elements contained in vectors  $\mathbf{y}$  and  $\hat{\mathbf{y}} = (\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n)^T$  gives the value of  $r_{cv}$ .

The methodology described here can be also applied in leave-two- or leave-many-out procedures.<sup>74</sup>

#### ACKNOWLEDGMENT

The present work was supported in part by the *Fundació Maria Francisca de Roviralta* as well as the European Commission contract #ENV4-CT97-0508, a UdG grant #3/00, and the CICYT project #SAF2000-223. This research has been carried out using the CEsCA and CEPBA resources, coordinated by C<sup>4</sup>. One of us (R. Ponc) acknowledges a CEPBA grant and also the support from the Czech Ministry of Education grant No. D09.20. The authors also thank the referees for their constructive criticism, which improved several aspects of this work.

#### REFERENCES AND NOTES

- (1) Carbó, R.; Leyda, L.; 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) Bowen-Jenkins, P. E.; Richards, W. G. Ab initio computations of molecular similarity. *J. Phys. Chem.* **1985**, *89*, 2195–2197.
- (3) Carbó, R.; Domingo, L. L. LCAO-MO Similarity Measures and Taxonomy. *Int. J. Quantum Chem.* **1987**, *23*, 517–545.
- (4) Hodgkin, E. E.; Richards, W. G. Molecular similarity based on electrostatic potential and electric field. *Int. J. Quantum Chem. Biol. Symp.* **1987**, *14*, 105–110.
- (5) Ponc, R. Topological aspects of chemical reactivity. On the similarity of molecular structures. *Collect. Czech. Chem. Commun.* **1987**, *52*, 555–561.
- (6) *Concepts and Applications of Molecular Similarity*; Johnson, M. A., Maggiora, G., Eds.; John Wiley & Sons: New York, 1990.
- (7) Cooper, D. L.; Allan, N. L. A novel approach to molecular similarity. *J. Comput.-Aided Mol. Design* **1989**, *3*, 253–259.
- (8) Cioslowski, J.; Fleischmann, E. D. Assessing molecular similarity from results of ab initio electronic structure calculations. *J. Am. Chem. Soc.* **1991**, *113*, 64–67.
- (9) Allan, N. L.; Cooper, D. L. A momentum space approach to molecular similarity. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 587–590.
- (10) *Shape in chemistry: and introduction to molecular shape and topology*; Mezey, P. G., Eds.; VCH: New York, 1993.
- (11) Carbó, R.; Calabuig, B.; Vera, L.; Besalú, E. Molecular quantum similarity: theoretical framework, ordering principles, and visualization techniques. *Adv. Quantum Chem.* **1994**, *25*, 253–313.
- (12) Solà, M.; Mestres, J.; Carbó, R.; Duran, M. Use of ab initio quantum molecular similarities as an interpretative tool for the study of chemical reactions. *J. Am. Chem. Soc.* **1994**, *116*, 5909–5915.
- (13) *Molecular Similarity and Reactivity: From Quantum Chemical to Phenomenological Approaches*; Carbó, R., Ed.; Kluwer Academic: Amsterdam, 1995.
- (14) Molecular Similarity I. In *Topics in Current Chemistry*; Sean, K. D., Ed.; Springer-Verlag: Berlin, 1995; Vol. 173.
- (15) Molecular Similarity II. In: *Topics in Current Chemistry*; Sean, K. D., Ed.; Springer-Verlag: Berlin, 1995; Vol. 174.
- (16) *Advances in Molecular Similarity*; Carbó-Dorca, R., Mezey, P. G., Eds.; JAI Press: Greenwich, CT, 1996; Vol. 1. 1998; Vol. 2.
- (17) Constans, 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.
- (18) 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 (QSPR). *J. Math. Chem.* **1995**, *18*, 237–246.
- (19) Carbó-Dorca, R. Tagged sets, convex sets and quantum similarity measures. *J. Math. Chem.* **1998**, *23*, 353–364.
- (20) Carbó-Dorca, R.; Besalú, E. A general survey of molecular quantum similarity. *J. Mol. Struct. (THEOCHEM)* **1998**, *451*, 11–23.
- (21) 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.
- (22) Carbó-Dorca, R. Stochastic transformation of quantum similarity matrices an their fuse in quantum QSAR (QQSAR) models. *Int. J. Quantum Chem.* **2000**, *79*, 163–177.
- (23) Lobato, M.; Amat, L.; Besalú, E.; 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.



- (24) 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.
- (25) 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.
- (26) Robert, D.; Gironés, X.; Carbó-Dorca, R. Facet diagrams for quantum similarity data. *J. Comput. Aided Mol. Design* **1999**, *13*, 597–610.
- (27) Robert, D.; Carbó-Dorca, R. Aromatic compounds aquatic toxicity QSAR using quantum similarity measures. *SAR QSAR Environ. Res.* **1999**, *10*, 401–422.
- (28) Robert, D.; Amat, L.; Carbó-Dorca, R. Quantum similarity QSAR: Study of inhibitors binding to thrombin, trypsin, and factor Xa, including a comparison with CoMFA and CoMSIA methods. *Int. J. Quantum Chem.* **2000**, *80*, 265–282.
- (29) Robert, D.; Gironés, X.; Carbó-Dorca, R. Quantification of the influence of single-point mutations on Haloalkane Dehalogenase activity: a molecular quantum study. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 839–846.
- (30) Amat, L.; Carbó-Dorca, R.; Ponec, R. Molecular quantum similarity measures as an alternative to log *P* values in QSAR studies. *J. Comput. Chem.* **1998**, *19*, 1575–1583.
- (31) Ponec, R.; Amat, L.; Carbó-Dorca, R. Molecular basis of quantitative structure-properties relationships (QSPR): a quantum similarity approach. *J. Comput. Aided Mol. Design* **1999**, *13*, 259–270.
- (32) Ponec, R.; Amat, L.; Carbó-Dorca, R. Quantum similarity approach to LFER: substituent and solvent effects on the acidities of carboxylic acids. *J. Phys. Org. Chem.* **1999**, *12*, 447–454.
- (33) Amat, L.; Carbó-Dorca, R.; Ponec, R. Simple linear QSAR models based on quantum similarity measures. *J. Med. Chem.* **1999**, *42*, 5169–5180.
- (34) Carbó-Dorca, R.; Robert, D.; Amat, L.; Gironés, X.; Besalú, E. Molecular quantum similarity in QSAR and drug design. In *Lecture Notes in Chemistry*; Springer: Berlin, 2000; Vol. 73.
- (35) Good, A. C.; Hodgkin, E. E.; Richards, W. G. Similarity screening of molecular data sets. *J. Comput.-Aided Mol. Design* **1992**, *6*, 513–520.
- (36) Good, A. C.; So, S.-S.; Richards, W. G. Structure–activity relationships from molecular similarity matrices. *J. Med. Chem.* **1993**, *36*, 433–438.
- (37) 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.
- (38) Lee, C.; Smithline, S. An approach to molecular similarity using density functional theory. *J. Phys. Chem.* **1994**, *98*, 1135–1138.
- (39) Measures, P. T.; Mort, K. A.; Allan, N. L.; Cooper, D. L. Applications of momentum-space similarity. *J. Comput.-Aided Mol. Design* **1995**, *9*, 331–340.
- (40) Benigni, R.; Cotta-Ramusino, M.; Giorgi, F.; Gallo, G. Molecular similarity matrices and quantitative structure–activity relationships: a case study with methodological implications. *J. Med. Chem.* **1995**, *38*, 629–635.
- (41) Mestres, J.; Rohrer, D. C.; Maggiora, G. M. A molecular field-based similarity approach to pharmacophoric pattern recognition. *J. Mol. Graphics Modelling* **1997**, *15*, 114–121.
- (42) Measures, P. T.; Mort, K. A.; Cooper, D. L.; Allan, N. L. A quantum molecular similarity approach to anti-HIV activity. *J. Mol. Struct. (THEOCHEM)* **1998**, *423*, 113–123.
- (43) Popelier, P. L. A. Quantum molecular similarity. 1. BCP space. *J. Phys. Chem. A* **1999**, *103*, 2883–2890.
- (44) Mestres, J.; Rohrer, D. C.; Maggiora, G. M. A molecular-field-based similarity study of nonnucleoside HIV-1 reverse transcriptase inhibitors. *J. Comput. Aided-Mol. Des.* **1999**, *13*, 79–93.
- (45) Goodford, P. J. A Computational procedure for determining energetically favorable binding sites on biologically important molecules. *J. Med. Chem.* **1985**, *28*, 849–857.
- (46) 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.
- (47) 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.
- (48) Silverman, B. D.; Platt, D. E. Comparative Molecular Moment Analysis (CoMMA): 3D-QSAR without Molecular Superposition. *J. Med. Chem.* **1996**, *39*, 2129–2140.
- (49) Robinson, D. D.; Winn, P. J.; Lyne, P. D.; Richards, W. G. Self-Organizing Molecular Field analysis: A Tool for Structure–Activity Studies. *J. Med. Chem.* **1999**, *42*, 573–583.
- (50) Oprea, T. I.; Ciubotariu, D.; Sulea, T. L.; Simon, Z. Comparison of the minimal Steric Difference (MTD) and Comparative Molecular Field Analysis (CoMFA) Methods for Analysis of Binding of Steroids to Carrier Proteins. *Quant. Struct.-Act. Relat.* **1993**, *12*, 21–26.
- (51) 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.
- (52) Hahn, M.; Rogers, D. Receptor Surface Models. 2. Application to Quantitative Structure–Activity Relationships Studies. *J. Med. Chem.* **1995**, *38*, 2091–2102.
- (53) Wagener, M.; Sadowski, J.; Gasteiger, J. Autocorrelation of Molecular Surface Properties for Modeling Corticosteroid Binding Globulin and Cytosolic Ah Receptor Activity by Neural Networks. *J. Am. Chem. Soc.* **1995**, *117*, 7769–7775.
- (54) Kellogg, G. E.; Kier, L. B.; Gaillard, P.; Hall, L. H. E-state Fields: Applications to 3D QSAR. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 513–520.
- (55) Anzali, S.; Barnickel, G.; Krug, M.; Sadowski, J.; Wagener, M.; Gasteiger, J.; Polanski, J. The comparison of geometric and electronic properties of molecular surfaces by neural networks: Application to the analysis of corticosteroid-binding globulin activity of steroids. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 521–534.
- (56) Norinder, U. 3D-QSAR Investigation of the Tripos Benchmark Steroids and some Protein-Tyrosine Kinase Inhibitors of Styrene Type using the TDQ Approach. *J. Chemom.* **1996**, *10*, 533–545.
- (57) Schnitker, J.; Gopalaswamy, R.; Crippen, G. M. Objective models for steroid binding sites of human globulins. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 93–110.
- (58) Bravi, G.; Gancia, E.; Mascagni, P.; Pegna, M.; Todeschini, R.; 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.* **1997**, *11*, 79–92.
- (59) Turner, D. B.; Willett, P.; Ferguson, A. M.; Heritage, T. Evaluation of a novel infrared range vibration-based descriptor (EVA) for QSAR studies. 1. General application. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 409–422.
- (60) Parretti, M. F.; Kroemer, R. T.; Rothman, J. H.; Richards, W. G. Alignment of Molecules by the Monte Carlo Optimization of Molecular Similarity Indices. *J. Comput. Chem.* **1997**, *18*, 1344–1353.
- (61) So, S.-S.; Karplus, M. Three-Dimensional Quantitative Structure–Activity Relationships from Molecular Similarity Matrixes an Genetic Neural Networks. 1. Method and Validations. *J. Med. Chem.* **1997**, *40*, 4347–4359.
- (62) Tominaga, Y.; Fujiwara, I. Prediction-Weighted Partial Least-Squares Regression Method (PWPLS) 2: application to CoMFA. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1152–1157.
- (63) Chen, H.; Zhou, J.; Xie, G. PARM: A Genetic Evolved Algorithm To Predict Bioactivity. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 243–250.
- (64) Carbó, R.; Besalú, E. Definition, mathematical examples an quantum chemical applications of nested summation symbols and logical Kronecker deltas. *Computers Chem.* **1994**, *18*, 117–126.
- (65) Carbó, R.; Besalú, E. Definition and quantum chemical applications of nested summations symbols and logical functions: Pedagogical artificial intelligence devices for formulae writing, sequential programming and automatic parallel implementation. *J. Math. Chem.* **1995**, *18*, 37–72.
- (66) Pecka, J.; Ponec, R. Simple analytical method for evaluation of statistical importance of correlations in QSAR studies. *J. Math. Chem.* **2000**, *23*, 13–22.
- (67) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Zakrzewski, V. G.; Montgomery, J. A., Jr.; Stratmann, R. E.; Burant, J. C.; Dapprich, S.; Millam, J. M.; Daniels, A. 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, T.; Al-Laham, M. 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.
- (68) Mezey, P. G. The Holographic Electron Density Theorem and Quantum Similarity Measures. *Mol. Phys.* **1999**, *96*, 169–178.
- (69) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. AM1: A New General Purpose Quantum Mechanical Molecular Model. *J. Am. Chem. Soc.* **1985**, *107*, 3902–3909.
- (70) AMPAC 6.01; Semichem, Inc.: 7128 Summit, Shawnee, KS 66216. D.A.



- (71) Wold, S. Cross-validatory estimation of the number of components in factor and principal components models. *Technometrics* **1978**, *20*, 397–405.
- (72) Wold, S. Validation of QSARs. *Quant. Struct.-Act. Relat.* **1991**, *10*, 191–193.
- (73) Montgomery, D. C.; Peck, E. A. *Introduction to linear regression analysis*; Wiley: New York, 1992.
- (74) Besalú, E. Fast computation of cross-validated properties in full linear leave-many-out procedures. IT-IQC-00-36; Institute of Computational Chemistry: *J. Math. Chem.* (in press).

CI000160U

#### 7.5.4 *Influència de factors estructurals en l'aproximació QS-SM de fragments*

D'igual manera com s'ha fet en el capítol 6 amb el mètode fonamentat en les matrius de semblança ( $n \times n$ ), en aquest apartat es vol avaluar l'efecte que tenen determinats factors en els coeficients estadístics resultants de l'aproximació *QS-SM* de fragments. El principal avantatge de l'aproximació de fragments respecte al mètode de matrius de semblança és la no dependència dels resultats finals envers la superposició molecular. Però hi ha altres factors que poden influir en els resultats finals, com són la densitat electrònica i la geometria de les molècules que s'escull per fer els càlculs de les *QS-SM* de fragments. Per mostrar la repercussió dels diferents components en els models *QSAR* s'ha escollit l'exemple dels derivats de l'àcid benzoic *para*-substituït. En la *Table 1* i la *Table 2* de l'article 7.3 es llisten els coeficients estadístics de les rectes de regressió lineal entre les *QS-SM* de 44 fragments definits sobre l'estructura comuna dels derivats de l'àcid benzoic i la constant  $\sigma$  de Hammett. Els resultats s'han obtingut per una sèrie de 12 compostos, amb geometries optimitzades amb el mètode de HF i el conjunt de funcions de base 3-21G\*, emprant mesures de semblança de tipus solapament i Coulomb. En l'estudi que es presenta tot seguit s'han optimitzat les geometries dels 12 derivats de l'àcid benzoic en el nivell de càlcul HF/6-31G\*, i sobre aquestes estructures de mínima energia s'han avaluat les *QS-SM* dels 44 fragments definits en l'article 7.3 emprant les densitats electròniques HF/3-21G\* i HF/6-311G. La intenció és, en primer lloc, veure si hi ha diferències entre els càlculs HF/3-21G\*//HF/3-21G\* i HF/6-31G\*//HF/3-21G\*, on únicament varia la geometria molecular. I en segon terme comprovar quina és la influència de la densitat electrònica en l'aproximació *QS-SM* de fragments a través d'observar les diferències en els models *QSAR* de les mesures HF/6-31G\*//HF/3-21G\* i HF/6-31G\*//HF/6-311G.

En la taula 7.4 s'exposen els coeficients  $r$  i  $r_{cv}$  per cadascuna de les *QS-SM* de fragment de tipus solapament calculades en els nivells HF/6-31G\*//HF/3-21G\* i HF/6-31G\*//HF/6-311G, mentre que en la taula 7.5 es llisten els resultats obtinguts per les mesures de Coulomb. La numeració dels àtoms de l'esquelet comú dels derivats de l'àcid benzoic que serveix per identificar els fragments moleculars és la mateixa que la descrita en l'article 7.3.

Mesures de solapament Fragment	HF/3-21G*		HF/6-311G		
	<i>r</i>	<i>r<sub>cv</sub></i>	<i>r</i>	<i>r<sub>cv</sub></i>	
O9H10	0.978	0.968	O8	0.982	0.963
O9	0.969	0.952	C7O8	0.970	0.941
C7O8	0.960	0.931	C7O8O9	0.966	0.938
C7O8O9	0.952	0.924	C1C7O8	0.962	0.949
C1	0.937	0.920	O9H10	0.943	0.915
C2C1C6	0.926	0.902	C7O8O9H10	0.941	0.900
C7O8O9H10	0.921	0.883	C1	0.938	0.920
C1C7O8	0.899	0.868	C1C6C5	0.906	0.874
C7	0.868	0.814	C1C7O9	0.883	0.843
C2C1	0.863	0.793	C3C2C1	0.873	0.836
C1C6	0.837	0.788	C1C7	0.862	0.807
C7O9	0.835	0.779	C1C6	0.859	0.804
O8	0.809	0.733	C2H11	0.836	0.765
C2C1H11	0.808	0.714	C6H14	0.805	0.726
C2C1C7	0.740	0.622	C2C1	0.796	0.689
C1C6H14	0.731	0.601	O9	0.794	0.697
C1C7	0.676	0.495	C5	0.782	0.703
C3C2H12	0.661	0.541	C7O9	0.768	0.690
C7O9H10	0.636	0.421	C2	0.764	0.676
C1C7O9	0.605	0.360	C1C6H14	0.761	0.637
C3H12	0.587	0.405	C6C5	0.758	0.641
C1C6C7	0.571	0.221	C6	0.741	0.639
C3C2H11	0.539	0.356	C2C1H11	0.718	0.584
C1C6C5	0.515	0.115	C3	0.687	0.561
C5H13	0.464	0.196	C6C5H14	0.679	0.526
C3C2	0.445	0.167	C2C1C6	0.646	0.474
C2	0.438	-0.034	C2C1C7	0.637	0.478
C6	0.413	-0.042	C3C5C4	0.633	0.388
C3	0.364	-0.015	C1C6C7	0.629	0.376
C3C2C4	0.351	-0.087	C5H13	0.604	0.417
C3C4	0.342	-0.152	C3C4	0.493	0.155
C4	0.320	-0.022	C5C4	0.469	0.027
C6C5H13	0.315	-0.098	C3C4H12	0.443	0.051
C6C5C4	0.307	-0.192	C6C5H13	0.434	0.152
C3C2C1	0.298	-0.726	C3C2C4	0.433	0.009
C3C4H12	0.261	-0.356	C6C5C4	0.430	-0.040
C5C4	0.236	-0.428	C3H12	0.411	-0.224
C6H14	0.165	-0.640	C5C4H13	0.402	-0.130
C3C5C4	0.158	-0.570	C3C2	0.327	-0.596
C5	0.150	-0.644	C4	0.290	-0.233
C5C4H13	0.126	-0.650	C3C2H12	0.189	-0.586
C2H11	0.103	-0.816	C7O9H10	0.180	-0.723
C6C5	0.081	-0.790	C3C2H11	0.137	-0.889
C6C5H14	0.073	-0.875	C7	0.044	-0.876

**Taula 7.4** Fragments moleculars i paràmetres estadístics de les equacions de regressió lineal pels derivats de l'àcid benzoic emprant *QS-SM ab initio* de tipus solapament en els nivells HF/3-21G\* i HF/6-311G.

Mesures de Coulomb Fragment	HF/3-21G*		HF/6-311G		
	<i>r</i>	<i>r<sub>cv</sub></i>	<i>r</i>	<i>r<sub>cv</sub></i>	
O8	0.994	0.990	O8	0.990	0.984
O9H10	0.992	0.989	O9H10	0.987	0.981
C7O8O9H10	0.983	0.977	C1C7O8	0.964	0.948
C7O8	0.983	0.976	O9	0.953	0.930
C7O8O9	0.980	0.972	C7O8O9H10	0.948	0.927
C2C1C7	0.977	0.970	C1C7	0.945	0.921
C1C7O8	0.975	0.962	C1C7O9	0.933	0.903
C1C7	0.971	0.956	C7O8O9	0.925	0.895
C1C6C7	0.964	0.942	C7O8	0.921	0.888
C1C7O9	0.956	0.938	C3C2C1	0.891	0.822
O9	0.955	0.937	C1C6C5	0.890	0.822
C1C6H14	0.938	0.900	C5H13	0.884	0.820
C1	0.930	0.905	C1	0.880	0.821
C2C1H11	0.929	0.906	C1C6C7	0.879	0.810
C6C5H13	0.915	0.867	C3H12	0.879	0.816
C7O9H10	0.913	0.874	C6C5H13	0.862	0.788
C3C2H12	0.901	0.838	C3C2H12	0.861	0.786
C5H13	0.889	0.822	C2	0.845	0.773
C3H12	0.880	0.782	C6C5H14	0.840	0.746
C1C6	0.857	0.785	C3C2H11	0.840	0.746
C1C6C5	0.852	0.756	C3	0.835	0.718
C3C2C1	0.849	0.713	C5	0.832	0.716
C6C5H14	0.835	0.713	C2C1C7	0.832	0.776
C3C2H11	0.822	0.680	C3C2	0.807	0.696
C7	0.819	0.736	C6C5	0.805	0.688
C2C1	0.818	0.746	C1C6H14	0.797	0.676
C6	0.753	0.647	C6	0.740	0.597
C6C5	0.748	0.552	C2H11	0.687	0.523
C2	0.741	0.648	C1C6	0.674	0.542
C3C2	0.734	0.511	C7O9	0.632	0.419
C5	0.733	0.557	C2C1H11	0.595	0.341
C3	0.696	0.392	C6H14	0.526	0.216
C2C1C6	0.647	0.543	C7O9H10	0.451	0.214
C7O9	0.587	0.402	C2C1C6	0.424	0.108
C4	0.480	0.162	C4	0.409	0.050
C3C2C4	0.438	0.030	C3C2C4	0.350	-0.124
C3C4	0.430	0.012	C6C5C4	0.332	-0.157
C6C5C4	0.426	-0.007	C3C4	0.296	-0.297
C5C4	0.407	-0.062	C5C4	0.273	-0.343
C3C4H12	0.357	-0.184	C7	0.260	-0.355
C5C4H13	0.332	-0.269	C3C4H12	0.231	-0.507
C3C5C4	0.331	-0.304	C5C4H13	0.213	-0.540
C6H14	0.190	-0.510	C3C5C4	0.109	-0.875
C2H11	0.120	-0.743	C2C1	0.102	-0.688

**Taula 7.5** Fragments moleculars i paràmetres estadístics de les equacions de regressió lineal pels derivats de l'àcid benzoic emprant *QS-SM ab initio* de tipus Coulomb en els nivells HF/3-21G\* i HF/6-311G.

Si es comparen els resultats HF/3-21G\*//HF/3-21G\* de les *Tables 1 i 2* de l'article 7.3 amb els valors HF/6-31G\*//HF/3-21G\* de les tres primeres columnes de les taules 7.4 i 7.5, s'aprecien petites variacions tant en els coeficients estadístics com en l'ordenació dels diferents fragments moleculars. Això indica que les *QS-SM* de fragment són sensibles a petits canvis en l'estructura molecular. Per exemple, la *QS-SM* de solapament del fragment COOH té un coeficient de correlació de 0.893 en el càlcul HF/3-21G\*//HF/3-21G\* presentat en l'article 7.3, mentre que en el nivell HF/6-31G\*//HF/3-21G\* és de 0.921.

El segon factor que es pot analitzar amb els resultats presentats en les taules 7.4 i 7.5 és l'efecte de la densitat electrònica en els càlculs *QS-SM*. Si es comparen els resultats HF/6-31G\*//HF/3-21G\* i HF/6-31G\*//HF/6-311G per les mesures de solapament i de Coulomb s'aprecien uns canvis més ostensibles que els produïts per la modificació de la geometria HF/3-21G\* a HF/6-31G\*. El tipus de densitat electrònica escollida en els càlculs de *QS-SM* té molta influència en els valors finals dels coeficients  $r$  i  $r_{cv}$ . Això també s'evidencia en els càlculs *PASA* de fragments, els quals depenen en gran mesura de les càrregues atòmiques escollides per modificar els coeficients atòmics.

Però en tots els estudis presentats, les *QS-SM* que millor descriuen la constant  $\sigma$  de Hammett són les generades mitjançant fragments moleculars construïts amb àtoms del grup COOH. Tant les mesures de solapament com les de Coulomb determinades amb les densitats electròniques HF/3-21G\* i HF/6-311G donen uns resultats qualitius similars, dels quals s'infereix que els fragments que millor descriuen els efectes electrònics dels substituents en els derivats de l'àcid benzoic són els formats amb àtoms del grup COOH.

## Discussió

La metodologia presentada en aquest capítol es fonamenta bàsicament en la utilització de *QS-SM* de fragments en qualitat de descriptors moleculars en la generació de models *QSAR*. En els apartats inicials s'ha demostrat que la *QS-SM* de tota la molècula o d'un determinat fragment es pot utilitzar com a alternativa a alguns descriptors clàssics de caire físico-químic, tal com  $\log P$  i la constant  $\sigma$  de Hammett. Un dels primers estudis ha estat la caracterització dels efectes electrònics dels substituents en la constant de dissociació dels àcids carboxílics per mitjà del càlcul de la *QS-SM* del grup COOH. S'han obtingut unes bones correlacions entre la constant  $\sigma$  de Hammett i les *QS-SM* del fragment COOH. Aquests resultats s'han vist confirmats amb les rectes de regressió entre els valors esperats de l'espai del moment i la constant  $\sigma$ . Posteriorment, i amb la finalitat de corroborar que la *QS-SM* del fragment COOH és la mesura de semblança que millor descriu els efectes electrònics que quantifica la constant  $\sigma$  de Hammett, s'ha analitzat una gran varietat de fragments moleculars definits sobre l'estructura comuna dels derivats de l'àcid benzoic *para*-substituït, i s'ha comprovat que els coeficients de regressió òptims corresponen a *QS-SM* de fragments moleculars generats amb els àtoms que pertanyen al grup COOH.

Quant a la hidrofobicitat, i concretament al terme  $\log P$ , el càlcul proposat inicialment de la mesura de semblança entre dues funcions densitat de la mateixa molècula però calculades en solvents diferents, aigua i octanol, és molt costós respecte del temps de computació perquè requereix el càlcul de densitats electròniques a nivell *ab initio*. Posteriorment s'ha observat que en els estudis *QSAR* es pot emprar la *QS-SM* de la funció densitat de la molècula entera en fase gas, i així s'eviten els càlculs teòrics en solvent. El principal desavantatge d'aquest tipus de descriptors teòrics és la seva limitació a sèries de compostos homogenis.

La següent fase ha estat la caracterització d'alguns models de *QSAR* clàssica emprant descriptors mecanicoquàntics en lloc dels paràmetres empírics. A la pràctica l'estudi s'ha centrat en la recerca de les *QS-SM* apropiades que poden reemplaçar els descriptors  $\log P$  i  $\sigma$  en les equacions de regressió multilineal clàssiques. S'ha comprovat que les propietats electròniques provocades per l'efecte dels substituents es poden caracteritzar emprant la *QS-SM* de regions moleculars locals, corresponents al

grup funcional que s'identifica amb la part molecular que participa en el procés (re)actiu. Tanmateix, en la majoria de problemes *QSAR* relacionats amb l'activat biològica o la toxicitat, els llocs o cavitats moleculars sensibles a l'activitat no són evidents. En atenció a això s'ha proposat una metodologia general que sigui capaç d'identificar els fragments moleculars que millor descriuen la propietat molecular, sense imposar cap restricció o especificació a priori. Bàsicament, el procés permet la detecció d'aquelles regions moleculars, comunes en tota la sèrie molecular, les quals són responsables d'una alta resposta biològica. Això permet obtenir un patró amb les regions actives que és d'evident interès en el propòsit del disseny de fàrmacs.

De manera simplificada, la tècnica s'aplica en sèries de compostos químics derivats d'una estructura comuna, en els quals la seva activitat biològica és funció de la presència o absència de diversos substituents. Els substituents provoquen variacions en la densitat electrònica dels àtoms de l'esquelet comú que es veuen reflectides en les *QS-SM* de fragment. Llavors, mitjançant anàlisis *QSAR*, es determinen les *QS-SM* dels fragments moleculars que influeixen des del punt de vista de la semblança molecular quàntica en la interacció de la molècula considerada amb el receptor. La representació de la freqüència que apareixen els diferents àtoms de l'esquelet comú en les *QS-SM* seleccionades permet dissenyar un patró genèric de l'activitat de la sèrie considerada. El proper objectiu en l'aproximació *QS-SM* de fragments és tractar de millorar la combinació de substituents per assolir l'òptim d'activitat a partir del patró dissenyat.

En l'aproximació *QS-SM* de fragment únicament es calculen els elements de la diagonal de la matriu de semblança, i per tant no és necessari efectuar optimitzacions de la superposició molecular. El nombre de mesures d'autosemblança a realitzar és igual al nombre de molècules que conté la sèrie analitzada. Això permet calcular les densitats electròniques a nivell *ab initio* sempre que la dimensió de les molècules no sigui massa gran. D'altra banda s'ha observat que les mesures *QS-SM* de fragment són molt sensibles a les variacions de l'estructura molecular i a petits canvis en la densitat electrònica. És per aquest motiu que l'aproximació *QS-SM* de fragments s'ha d'entendre com una anàlisi qualitativa més que quantitativa, que permet identificar les regions moleculars importants per a una determinada activitat des d'un punt de vista de la semblança molecular quàntica.

## Referències

1. A. Crum-Brown, T. R. Fraser. On the connection between chemical constitution and physiological action. Part 1. On the physiological action of the ammonium bases, derived from strychnia, brucia, thebaia, codeia, morphia and nicotia. *Trans. Royal Soc. Edinburgh* **1868**, 25, 257–274.
2. H. Meyer. Theorie der alkoholnarkose, welche eigenschaft die anästhetica bedingt ihre narkotische wirkung. *Arch. Explt. Pathol. Pharmacol.* **1899**, 42, 109–118.
3. E. Overton. Studien über die Narkose. Gustav Fisher, Jena, 1901.
4. J. Ferguson. The use of chemical potentials as indicators of toxicity. *Proc. Roy. Soc. (London)* **1939**, 127B, 387.
5. J. C. McGowan. *J. Appl. Chem. (London)* **1951**, 1, 120.
6. J. C. McGowan. *J. Appl. Chem. (London)* **1954**, 4, 41.
7. S. M. Free, J. W. Wilson. A mathematical contribution to structure-activity studies. *J. Med. Chem.* **1964**, 7, 395–399.
8. L. P. Hammett. The effect of structures upon the reactions of organic compounds. Benzene derivatives. *J. Am. Chem. Soc.* **1937**, 59, 96–103.
9. L. P. Hammett. Physical organic chemistry. McGraw-Hill, New York, 1940.
10. R. W. Taft. Polar and steric substituent constants for aliphatic and *o*-benzoate groups from rates of esterification and hydrolysis of esters. *J. Am. Chem. Soc.* **1952**, 74, 3120–3128.
11. R. W. Taft. Steric effects in organic chemistry. M. S. Newman (ed.). Wiley, New York, pg. 556–675, 1956.
12. C. Hansch, T. Fujita.  $\rho$ - $\sigma$ - $\pi$  analysis. A method for the correlation of biological activity and chemical structure. *J. Am. Chem. Soc.* **1964**, 86, 1616–1626.
13. T. Fujita, J. Iwasa, C. Hansch. A new substituent constant,  $\pi$ , derived from partition coefficients. *J. Am. Chem. Soc.* **1964**, 86, 5175–5180.
14. R. Collander. The partition of organic compounds between higher alcohols and water. *Acta Chem. Scand.* **1951**, 5, 774–780.
15. C. Hansch, A. Leo. Exploring QSAR. Fundamentals and applications in chemistry and biology. ACS professional reference book, American Chemical Society, Washington, DC 1995.
16. C. Hansch, D. Kim, A. J. Leo, E. Novellino, C. Silipo, A. Vittoria. Toward a quantitative comparative toxicology of organic compounds. *CRC Crit. Rev. Toxicol.* **1989**, 19, 185–226.
17. K. H. Kim, C. Hansch, J. Y. Fukunaga, E. E. Steller, P. Y. C. Jow, P. N. Craig, J. Page. Quantitative structure-activity relationships in 1-aryl-2-(alkylamino)ethanol antimalarials. *J. Med. Chem.* **1979**, 22, 366–391.
18. C. Hansch, R. N. Smith, R. Engle, H. Wood. Quantitative structure activity relationships of antineoplastic drugs: nitrosoureas and triazenoimidozales. *Cancer Chemother. Rep.* **1972**, 56, 443–456.
19. S. P. Gupta. QSAR studies on drugs acting at the central nervous system. *Chem. Rev.* **1989**, 89, 1765–1800.



20. F. Helmer, K. Kiehs, C. Hansch. The linear free-energy relationship between partition coefficients and the binding and conformational perturbation of macromolecules by small organic compounds. *Biochemistry* **1968**, *7*, 2858–2863.
21. M. Matsumura, M. W. Bechtel, B. W. Mathews. Hydrophobic stabilization in T4 lysozyme determined directly by multiple substitution of Ile 3. *Nature* **1988**, *334*, 406–410.
22. R. F. Rekker. The hydrophobic fragmental constant. Its derivation and applications. A means of characterizing membrane systems. Elsevier, Amsterdam, 1977.
23. G. C. Nys, R. F. Rekker. Statistical analysis of a series of partition coefficients with special reference to the predictability of folding of drug molecules. The introduction of hydrophobic fragmental constants ( $f$  Values). *Chim. Ther.* **1973**, *8*, 521–535.
24. R. F. Rekker, H. M. De Kort. The hydrophobic fragmental constant, an extension to a 1000 datapoint set. *Eur. J. Med. Chem.* **1979**, *14*, 479–488.
25. C. Hansch, A. Leo. Substituent constants for correlation analysis in chemistry and biology. Wiley, New York, 1979.
26. A. Ghose, G. Crippen. Atomic physicochemical parameters for three-dimensional structure-directed quantitative structure-activity relationships. I. Partition coefficients as a measure of hydrophobicity. *J. Comput. Chem.* **1986**, *7*, 565–577.
27. G. Klopman, L. Iroff. Calculation of partition coefficients by the charge density method. *J. Comput. Chem.* **1981**, *2*, 157–160.
28. G. Klopman, K. Nambodiri, M. Schochet. Simple method of computing the partition coefficient. *J. Comput. Chem.* **1985**, *6*, 28–38.
29. Ll. Amat, R. Carbó-Dorca, R. Ponc. Molecular quantum similarity measures as an alternative to log P values in QSAR studies. *J. Comput. Chem.* **1998**, *19*, 1575–1583.
30. S. Miertus, E. Scrocco, J. Tomasi. Electrostatic interaction of a solute with a continuum. A direct utilization of *ab initio* molecular potentials for the prevision of solvent effects. *Chem. Phys.* **1981**, *55*, 117–129.
31. S. Miertus, J. Tomasi. Approximate evaluations of the electrostatic free energy and internal energy changes in solution processes. *Chem. Phys.* **1982**, *65*, 239–245.
32. GAUSSIAN 98, Revision A.6. M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, V. G. Zakrzewski, J. A. Montgomery Jr., R. E. Stratmann, J. C. Burant, S. Dapprich, J. M. Millam, A. D. Daniels, K. N. Kudin, M. C. Strain, O. Farkas, J. Tomasi, V. Barone, M. Cossi, R. Cammi, B. Mennucci, C. Pomelli, C. Adamo, S. Clifford, J. Ochterski, G. A. Petersson, P. Y. Ayala, Q. Cui, K. Morokuma, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. Cioslowski, J. V. Ortiz, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. Gomperts, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, C. Gonzalez, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, J. L. Andres, C. Gonzalez, M. Head-Gordon, E. S. Replogle, J. A. Pople. Gaussian, Inc.: Pittsburgh, PA, 1998.
33. X. Gironés, Ll. Amat, D. Robert, R. Carbó-Dorca. Use of electron-electron repulsion energy as a molecular descriptor in QSAR and QSPR studies. *J. Comput.-Aided Mol. Design* **2000**, *14*, 477–485.
34. X. Gironés, Ll. Amat, R. Carbó-Dorca. Using molecular quantum similarity measures as descriptors in quantitative structure-toxicity relationships. *SAR QSAR Environ. Res.* **1999**, *10*, 545–556.
35. O. Exner. Correlations Analysis of Chemical Data. Plenum Press, New York, 1988.

36. M. J. S. Dewar, E. G. Zoebisch, E. F. Healy, J. J. P. Stewart. AM1: a new general purpose quantum mechanical molecular model. *J. Am. Chem. Soc.* **1985**, *107*, 3902–3909.
37. J. J. P. Stewart, MOPAC 6.0 QCPE 455, Indiana University, Bloomington, IN, 1993.
38. D. Vlachová, L. Drobnička. Some relationships between biological activity and physico-chemical properties of monosubstituted phenylisothiocyanates. *Collect. Czech. Chem. Commun.* **1966**, *31*, 997–1008.
39. Ll. Amat, R. Carbó-Dorca, D. L. Cooper, N. L. Allan, R. Ponec. Structure-property relationships and momentum-space quantities: Hammett  $\sigma$  constants. *Mol. Phys.* **2003**, *en premsa*.
40. P. G. Mezey. The holographic electron density theorem and quantum similarity measures. *Mol. Phys.* **1999**, *96*, 169–178.
41. AMPAC 6.01, Semicem, Inc., 7128 Summit, Shawnee, KS 66216. D.A.
42. R. Carbó, E. Besalú. Definition, mathematical examples and quantum chemical applications of nested summation symbols and logical Kronecker deltas. *Comput. Chem.* **1994**, *18*, 117–126.
43. R. Carbó, E. Besalú. Definition and quantum chemical applications of nested summations symbols and logical functions: pedagogical artificial intelligence devices for formulae writing sequential programming and automatic parallel implementation. *J. Math. Chem.* **1995**, *18*, 37–72.
44. C. Hansch, J. McClarin, T. Klein, R. Langridge. A quantitative structure-activity relationship and molecular graphics study of carbonic anhydrase inhibitors. *Molec. Pharmac.* **1985**, *27*, 493–498.
45. F. Markwart, H. Landmann, P. Walsmann. Comparative studies on the inhibition of trypsin, plasmin, and thrombin by derivatives of benzylamine and benzamidine. *Eur. J. Biochem.* **1968**, *6*, 502–506.
46. D. Hadjipavlou-Litina, C. Hansch. Quantitative structure-activity relationships of the benzodiazepines. A review and reevaluation. *Chem. Rev.* **1994**, *94*, 1483–1505.
47. W. Zhang, K. F. Koehler, P. Zhang, J. M. Cook. Development of a comprehensive pharmacophore model for the benzodiazepine receptor. *Drug Des. Discovery* **1995**, *12*, 193–248.
48. A. Da Settimo, G. Primofiore, F. Da Settimo, A. M. Marini, E. Novellino, G. Greco, C. Martini, G. Giannaccini, A. Lucacchini. Synthesis, structure-activity relationships, and molecular modeling studies of *N*-(Indole-3-ylglyoxylyl)benzylamine derivatives acting at the benzodiazepine receptor. *J. Med. Chem.* **1996**, *39*, 5083–5091.
49. R. D. Cramer III, D. E. Patterson, J. D. Bunce. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959–5967.
50. M. Lobato, Ll. Amat, E. Besalú, R. Carbó-Dorca. Structure-activity relationships of a steroid family using quantum similarity measures and topological quantum similarity indices. *Quant. Struct.-Act. Relat.* **1997**, *16*, 465–472.
51. D. Robert, Ll. Amat, R. Carbó-Dorca. 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.



## Conclusions

---

Al llarg d'aquesta tesi s'han tractat diferents aspectes relacionats amb el càlcul de les mesures de semblança quàntica, així com la seva aplicació en la racionalització i predicció de l'activitat de fàrmacs. A ressenyar els avenços produïts en la descripció de les molècules mitjançant les funcions densitat aproximades *PASA* i la ideació de tècniques de superposició molecular específiques de les mesures de semblança quàntica. El desenvolupament d'aquests nous procediments i algorismes matemàtics associats a les *MQSM* ha estat essencial per poder progressar en diferents àmbits de la recerca, sobretot els relacionats amb les anàlisis *QSAR*. Precisament en el domini de les relacions estructura-activitat s'han presentat dues aproximacions fonamentades en la semblança molecular quàntica que s'originen a partir de dues representacions diferents de les molècules. La primera descripció considera la densitat electrònica global de les molècules i és important, entre altres, la disposició dels objectes comparats en l'espai i la seva conformació tridimensional. El resultat és una matriu de semblança amb les mesures de semblança de tots els parells de compostos que formen el conjunt estudiat. La segona descripció es fonamenta en la partició de la densitat global de les molècules en fragments. S'utilitzen mesures d'autosemblança per analitzar els requeriments bàsics d'una determinada activitat des del punt de vista de la semblança quàntica. El principal factor que influeix en les mesures de fragments és el tipus de densitat electrònica escollida per realitzar els càlculs.

A part de les discussions exposades en els capítols *Funcions densitat ASA*, *Superposició molecular*, *Matrius de semblança en anàlisis QSAR* i *Aproximació QS-SM de fragments*, es poden extreure les següents conclusions del treball presentat:

- I. S'ha demostrat que es pot obtenir una bona descripció de la densitat electrònica d'una molècula mitjançant una definició pomolecular, consistent en sumar les densitats individuals dels àtoms que la formen. Les densitats *PASA* són molt simples però suficientment acurades per possibilitar l'avaluació pràctica de les

*MQSM*. La seva disponibilitat ha permès dur a terme un gran nombre d'aplicacions en el camp de les anàlisis *QSAR*.

- II. S'ha proposat un nou mètode d'ajust de la densitat electrònica dels àtoms amb dos trets distintius. La primera peculiaritat és el desenvolupament d'un algorisme fonamentat en la tècnica de rotacions de Jacobi per ajustar els coeficients de l'expansió lineal. La segona innovació és l'adaptació dels exponents de les capes atòmiques mitjançant un mètode de Newton. El resultat són unes funcions densitat aproximades que reproduïxen molt acuradament les distribucions de densitat de càrrega electrònica *ab initio* amb el mínim nombre de capes.
- III. S'ha mostrat un exemple d'aplicació de les densitats electròniques *PASA* en el càlcul d'energies electròniques. En concret les funcions *PASA* s'han utilitzat per disminuir el nombre de cicles *SCF* en el càlcul de l'energia HF. Es defineix un Hamiltonià inicial igual a la suma de la contribució monoelectrònica calculada a nivell *ab initio* més una estimació de la repulsió bielectrònica avaluada combinant les funcions *PASA* amb el conjunt de funcions de base *ab initio*. Els resultats indiquen que el nombre de cicles *SCF* es redueix sobretot en complexos de metalls de transició si es compara amb algunes aproximacions dels programes comercials.
- IV. El procés de superposició molecular constitueix un dels eixos centrals de qualsevol estudi de semblança molecular fonamentat en descriptors tridimensionals. S'ha dissenyat un algorisme de sobreposició molecular basat en el màxim de semblança i deduït a partir d'una solució límit. A més, els diferents algorismes simplificats de la maximització global de la funció de semblança són procediments ràpids i molt eficaços en l'obtenció del màxim de semblança entre molècules.
- V. L'anàlisi de la influència de determinats factors en els models *QSAR* generats a partir de matrius de semblança quàntica ha demostrat que l'ús de les densitats *PASA* no altera els resultats estadístics finals si es comparen amb els valors *ab initio*. En canvi, la superposició molecular i la conformació de les molècules estudiades són dos factors a tenir en compte.

- VI. S'ha proposat una nova aproximació *QSAR* fonamentada en *QS-SM* de fragments moleculars. En els primers estudis s'ha comprovat que les propietats electròniques degudes a l'efecte dels substituents poden ser caracteritzades mitjançant la *QS-SM* de regions moleculars locals, corresponents al grup funcional que es pot identificar amb la part molecular que participa en el procés (re)actiu. Posteriorment s'ha proposat una metodologia general, capaç d'identificar els fragments moleculars que millor descriuen la propietat molecular sense imposar cap restricció o especificació a priori. Bàsicament, el procés permet la detecció d'aquelles regions moleculars, comunes en tota la sèrie química, les quals són responsables d'una alta resposta biològica. Això permet obtenir un patró amb les regions actives, que és d'evident interès per als propòsits del disseny de fàrmacs.



## Llistat de publicacions

- E. Besalú, **Ll. Amat**, X. Fradera, R. Carbó. An application of the molecular quantum similarity: ordering of some properties of the hexanes. Publicat en el llibre: QSAR and molecular modeling: concepts, computational tools and biological applications. F. Sanz, J. Giraldo, F. Manaut (Eds.). Prous Science Publishers, pàgines 396-399, 1995.
- **Ll. Amat**, E. Besalú, R. Carbó, X. Fradera. Practical applications of quantum molecular similarity measures (QMSM): Programs and examples. *SCIENTIA gerundensis* **1995**, 21, 127–143.
- R. Carbó, E. Besalú, **Ll. Amat**, X. Fradera. Quantum molecular similarity measures (QMSM) as a natural way leading towards a theoretical foundation of quantitative structure-properties relationships (QSPR). *J. Math. Chem.* **1995**, 18, 237–246.
- **Ll. Amat**, R. Carbó, P. Constans. Algorisme d'optimització global de les mesures de semblança quàntica molecular. *SCIENTIA gerundensis* **1996**, 22, 109–121.
- **Ll. Amat**, X. Fradera, R. Carbó. Sobre els mapes de semblança quàntica molecular. *SCIENTIA gerundensis* **1996**, 22, 97–107.
- R. Carbó, E. Besalú, **Ll. Amat**, X. Fradera. On quantum molecular similarity measures (QMSM) and indices (QMSI). *J. Math. Chem.* **1996**, 19, 47–56.
- R. Carbó-Dorca, E. Besalú, **Ll. Amat**, X. Fradera. Quantum molecular similarity measures: concepts, definitions, and applications to quantitative structure-property relationships. Publicat en el llibre: Advances in molecular similarity. R. Carbó-Dorca, P. G. Mezey (Eds.). JAI Press, London, volum 1, pàgines 1–41, 1996.
- P. Constans, **Ll. Amat**, X. Fradera, R. Carbó-Dorca. Quantum molecular similarity measures (QMSM) and the atomic shell approximation (ASA). Publicat en el llibre: Advances in molecular similarity. R. Carbó-Dorca, P. G. Mezey (Eds.). JAI Press, London, volum 1, pàgines 187–211, 1996.
- X. Fradera, **Ll. Amat**, M. Torrent, J. Mestres, P. Constans, E. Besalú, J. Martí, S. Simon, M. Lobato, J. M. Oliva, J. M. Luis, J. L. Andrés, M. Solà, R. Carbó, M. Duran. Analysis of the changes on the potential energy surface of Menshutkin reactions induced by external perturbations. *J. Mol. Struct. (Theochem)* **1996**, 371, 171–183.
- P. Constans, **Ll. Amat**, R. Carbó-Dorca. Toward a global maximization of the molecular similarity function: superposition of two molecules. *J. Comput. Chem.* **1997**, 18, 826–846.
- **Ll. Amat**, R. Carbó-Dorca. Quantum similarity measures under atomic shell approximation: first order density fitting using elementary Jacobi rotations. *J. Comput. Chem.* **1997**, 18, 2023–2039.



- X. Fradera, **Ll. Amat**, E. Besalú, R. Carbó-Dorca. Application of molecular quantum similarity to *QSAR*. *Quant. Struct.-Act. Relat.* **1997**, *16*, 25–32.
- M. Lobato, **Ll. Amat**, E. Besalú, R. Carbó-Dorca. Structure-activity relationships of a steroid family using quantum similarity measures and topological quantum similarity indices. *Quant. Struct.-Act. Relat.* **1997**, *16*, 465–472.
- M. Lobato, **Ll. Amat**, E. Besalú, R. Carbó-Dorca. Estudi d'una família de quinolones utilitzant índexs de semblança i índexs topològics de semblança. *SCIENTIA gerundensis* **1997**, *23*, 17–27.
- **Ll. Amat**, R. Carbó-Dorca, R. Ponec. Molecular quantum similarity measures as an alternative to log P values in *QSAR* studies. *J. Comput. Chem.* **1998**, *19*, 1575–1583.
- R. Carbó-Dorca, **Ll. Amat**, E. Besalú, M. Lobato. Quantum similarity. Publicat en el llibre: Advances in molecular similarity. R. Carbó-Dorca, P. G. Mezey (Eds.). JAI Press, London, volum 2, pàgines 1–41, 1998.
- **Ll. Amat**, D. Robert, E. Besalú, R. Carbó-Dorca. Molecular quantum similarity measures tuned 3D *QSAR*: an antitumoral family validation study. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 624–631.
- X. Gironés, **Ll. Amat**, R. Carbó-Dorca. A comparative study of isodensity surfaces using *ab initio* and ASA density functions. *J. Mol. Graph. Model.* **1998**, *16*, 190–196.
- **Ll. Amat**, R. Carbó-Dorca. Fitted electronic density functions from H to Rn for use in quantum similarity measures: cis-diammine-dichloroplatinum(II) complex as an application example. *J. Comput. Chem.* **1999**, *20*, 911–920.
- R. Ponec, **Ll. Amat**, R. Carbó-Dorca. Molecular basis of quantitative structure-properties relationships (*QSPR*): a quantum similarity approach. *J. Comput.-Aided Mol. Design* **1999**, *13*, 259–270.
- R. Ponec, **Ll. Amat**, R. Carbó-Dorca. Quantum similarity approach to LFER: substituent and solvent effects on the acidities of carboxylic acids. *J. Phys. Org. Chem.* **1999**, *12*, 447–454.
- **Ll. Amat**, R. Carbó-Dorca, R. Ponec. Simple linear *QSAR* models based on quantum similarity measures. *J. Med. Chem.* **1999**, *42*, 5169–5180.
- P. G. Mezey, R. Ponec, **Ll. Amat**, R. Carbó-Dorca. Quantum similarity approach to the characterization of molecular chirality. *Enantiomer* **1999**, *4*, 371–378.
- X. Gironés, **Ll. Amat**, R. Carbó-Dorca. Using molecular quantum similarity measures as descriptors in quantitative structure-toxicity relationships. *SAR QSAR Environ. Res.* **1999**, *10*, 545–556.

- X. Gironés, **Ll. Amat**, R. Carbó-Dorca. Descripció de propietats moleculars i activitats biològiques emprant l'energia de repulsió electró-electró. *SCIENTIA gerundensis* **1999**, *24*, 197–208.
- D. Robert, **Ll. Amat**, R. Carbó-Dorca. 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.
- X. Gironés, **Ll. Amat**, D. Robert, R. Carbó-Dorca. Use of electron-electron repulsion energy as a molecular descriptor in *QSAR* and *QSPR* studies. *J. Comput.-Aided Mol. Design* **2000**, *14*, 477–485.
- R. Carbó-Dorca, **Ll. Amat**, E. Besalú, X. Gironés, D. Robert. Quantum mechanical origin of *QSAR*: theory and applications. *J. Mol. Struct. (Theochem)* **2000**, *504*, 181–228.
- R. Carbó-Dorca, D. Robert, **Ll. Amat**, X. Gironés, E. Besalú, Molecular quantum similarity in *QSAR* and drug design. *Lecture Notes in Chemistry*, *73*, Springer Verlag, Berlin, 2000.
- D. Robert, **Ll. Amat**, R. Carbó-Dorca. Quantum similarity *QSAR*: study of inhibitors binding to Trombin, Trypsin, and Factor Xa, including a comparison with CoMFA and CoMSIA methods. *Int. J. Quantum Chem.* **2000**, *80*, 265–282.
- **Ll. Amat**, R. Carbó-Dorca. Molecular electronic density fitting using elementary Jacobi rotations under atomic shell approximation. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1188–1198.
- A. Bach, **Ll. Amat**, E. Besalú, R. Carbó-Dorca, R. Ponec. Quantum chemistry, Sobolev spaces and SCF. *J. Math. Chem.* **2000**, *28*, 59–70.
- **Ll. Amat**, E. Besalú, R. Carbó-Dorca, R. Ponec. Identification of active molecular sites using quantum-self-similarity measures. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 978–991.
- R. Carbó-Dorca, **Ll. Amat**, E. Besalú, X. Gironés, D. Robert. Quantum molecular similarity measures: theory and applications to the evaluation of molecular properties, biological activities and toxicity. Publicat en el llibre: Fundamentals of molecular similarity. R. Carbó-Dorca, X. Gironés, P. G. Mezey (Eds.). Kluwer Academic/Plenum Press, New York, 2001.
- **Ll. Amat**, R. Carbó-Dorca. Use of promolecular ASA density functions as a general algorithm to obtain starting MO in SCF calculations. *Int. J. Quantum Chem.* **2002**, *87*, 59–67.
- X. Gironés, **Ll. Amat**, R. Carbó-Dorca. Modeling large macromolecular structures using promolecular densities. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 847–852.

- E. Besalú, X. Gironés, **Ll. Amat**, R. Carbó-Dorca. Molecular quantum similarity and the fundamentals of QSAR. *Acc. Chem. Res.* **2002**, *35*, 289–295.
- **Ll. Amat**, R. Carbó-Dorca, D. L. Cooper, N. L. Allan. Classification of reaction pathways via momentum-space and quantum molecular similarity measures. *Chem. Phys. Lett.* **2003**, *367* 207–213.
- **Ll. Amat**, R. Carbó-Dorca, D. L. Cooper, N. L. Allan, R. Ponec. Structure-property relationships and momentum-space quantities: Hammett  $\sigma$  constants. *Mol. Phys.* **2003**, *en premsa*.
- D. L. Cooper, N. L. Allan, **Ll. Amat**, R. Carbó-Dorca. Transition states and linear free-energy relationships explored by momentum-space and quantum similarity concepts. Proceedings of the 5th Girona Seminar on Molecular Similarity.
- G. Espinosa, A. Arenas, F. Giralt, **Ll. Amat**, X. Gironés, R. Carbó-Dorca. QSAR for TD<sub>50</sub> of aromatic compound by using an integrated SOM-Fuzzy Artmap based neural system with topological and quantum molecular similarity descriptors. *SAR QSAR Environ. Res.*, *en revisió*. Dins del monogràfic dedicat a la tercera sessió del 5th Girona Seminar on Molecular Similarity.