SCIENTIA gerundensis, 24: 175-181 (1999)

ANÀLISI DE PROCRUSTES I ALINEAMENT MOLECULAR

D. Robert i R. Carbó-Dorca

Institut de Química Computacional
Universitat de Girona
Campus Montilivi, 17071 Girona

RESUM

L'algorisme de McLachlan per a l'alineament de dos conjunts de coordenades atòmiques és interpretat sota l'òptica de l'Anàlisi Multivariant, que posa de manifest que el plantejament d'aquest problema és equivalent al de l'anàlisi de Procrustes i que la solució proposada per Kabsch és anàloga a la de Sibson, desenvolupada independentment.

RESUMEN

El algoritmo de McLachlan para el alineamiento de dos conjuntos de coordenadas atómicas se interpreta bajo la óptica del Análisis Multivariante, poniendo de manifiesto que el planteamiento de este problema es equivalente al del análisis de Procrustes y que la solución propuesta por Kabsch es análoga a la de Sibson, desarrollada independientemente.

ABSTRACT

McLachlan's algorithm for the alignment of two sets of atomic coordinates is interpreted under the Multivariate Analysis point of view, putting into evidence that it is equivalent to Procrustes analysis, and that the solution proposed by Kabsch is analogous to Sibson's, independently developed.

Keywords: Atomic coordinates, Molecular alignment, Procrustes analysis

INTRODUCCIÓ

Un problema obert en el camp de la química teòrica és l'establiment d'un algorisme eficient per a alinear en l'espai compostos pertanyents a una mateixa família. La superposició d'estructures moleculars té aplicació en la cerca automatitzada en bases de dades tridimensionals (1), en els models 3D-QSAR construïts en càlculs de camps en punts d'una xarxa rectangular tridimensional (2) i en totes aquelles tècniques comparatives que utilitzin estructures moleculars tridimensionals, com ara la Semblança Molecular Quàntica (3). Entre les possibles solucions que s'han donat a aquesta qüestió, hi figura la minimització de distàncies atòmiques intramoleculars d'aquells fragments que formen l'esquelet comú dels compostos (4).

MATERIALS I MÈTODES

L'anàlisi de Procrustes

L'anàlisi de Procrustes (5) és una eina d'anàlisi multivariant dissenyada per comparar configuracions derivades amb tècniques d'escalat multidimensional (6), entenent-se per *configuració* o *representació* el conjunt de coordenades de n objectes en un espai Euclidià p-dimensional, expressades en forma de matriu (pxn). L'anàlisi de Procrustes cerca l'escalat isotròpic i les translacions i rotacions rígides que millor facin coincidir ambdues configuracions. Per a aquest problema han estat proposades diverses solucions (7-9). En l'àmbit de la química ha estat utilitzat per a comparar representacions generades per diferents operadors i índexs de Semblança Quàntica (10), per a comparar diferents conjunts de descriptors moleculars (11) i com a tècnica de selecció de variables en models QSAR (12).

L'anàlisi de Procrustes es pot descriure de la manera següent: siguin **A** i **B** dues configuracions de n objectes en un espai p-dimensional. S'assumeix que existeix una relació 1:1 entre cada objecte d'un espai i l'altre. La suma dels quadrats de les distàncies euclidianes entre els punts de les configuracions **A** i **B** val:

$$M^2 = \sum_{i=1}^{n} \left\| \mathbf{b}_i - \mathbf{a}_i \right\|^2 = \sum_{i=1}^{n} \left(\mathbf{b}_i - \mathbf{a}_i\right)^{\mathrm{T}} \left(\mathbf{b}_i - \mathbf{a}_i\right) \tag{1}$$

Per comparar les configuracions, es permetrà als punts de la representació **A** ser traslladats, rotats i dilatats o contrets fins a obtenir una nova representació, denotada per **A'**, que coincideixi el millor possible amb la configuració **B**. Les coordenades de l'espai **A** es transformen seguint l'expressió:

$$\mathbf{a}'_i = \delta\, \mathbf{R}^{\mathrm{T}} \mathbf{a}_i + \mathbf{t} \tag{2}$$

En l'equació anterior, δ és un factor d'escala (dilatació/contracció), **R** és una matriu (pxp) de rotació i **t** és un vector de translació. El vector $\mathbf{a}_i$, al contrari, està format per les p coordenades de l'objecte i-èsim. Substituint l'expressió (2) dins l'equació (1):

$$M^2 = \sum_{i=1}^{n} \left(\mathbf{b}_i - \delta \mathbf{R}^{\mathrm{T}} \mathbf{a}_i - \mathbf{t}\right)^{\mathrm{T}} \left(\mathbf{b}_i - \delta \mathbf{R}^{\mathrm{T}} \mathbf{a}_i - \mathbf{t}\right) \tag{3}$$

El criteri comparatiu és clar: que dues configuracions s'assemblin voldrà dir que la suma dels quadrats de les distàncies entre llurs punts serà el més petita possible. La derivació de les expressions de δ, **R** i **t** perquè la funció M^2 sigui mínima es pot trobar al treball de Krzanowski i Marriott (13), i no serà reproduïda aquí. Manipulacions algebraiques de l'expressió anterior duen a considerar la translació òptima com aquella que situa els centroids d'ambdues configuracions al mateix punt. Sibson (14) va deduir la fórmula de la matriu de rotacions òptima amb una elegant

demostració que no requeria calcular la derivada de M^2. Existeix una demostració alternativa proporcionada per Mardia *et al.* (15). La forma explícita d'aquesta rotació és:

$$\mathbf{R} = \left(\mathbf{A}^{\mathrm{T}}\mathbf{B}\mathbf{B}^{\mathrm{T}}\mathbf{A}\right)^{1/2}\left(\mathbf{B}^{\mathrm{T}}\mathbf{A}\right)^{-1} \tag{4}$$

El factor d'escala òptim δ depèn de la rotació òptima i segueix la següent expressió:

$$\delta = \frac{\mathrm{tr}\left(\mathbf{A}\mathbf{R}\mathbf{B}^{\mathrm{T}}\right)}{\mathrm{tr}\left(\mathbf{A}\mathbf{A}^{\mathrm{T}}\right)} = \frac{\mathrm{tr}\left(\mathbf{R}\mathbf{B}^{\mathrm{T}}\mathbf{A}\right)}{\mathrm{tr}\left(\mathbf{A}\mathbf{A}^{\mathrm{T}}\right)} = \frac{\mathrm{tr}\left(\mathbf{A}^{\mathrm{T}}\mathbf{B}\mathbf{B}^{\mathrm{T}}\mathbf{A}\right)^{1/2}}{\mathrm{tr}\left(\mathbf{A}\mathbf{A}^{\mathrm{T}}\right)} \tag{5}$$

on a la darrera igualtat s'ha substituït el valor de **R** pel donat a l'equació (4).

La mesura de semblança entre les dues configuracions pot dur-se a terme mitjançant el valor minimitzat de M^2, M_0^2, del qual ara ja es coneixen els termes que el formen:

$$M_0^2 = \mathrm{tr}\left(\mathbf{B}\mathbf{B}^{\mathrm{T}}\right) - \frac{\left\{\mathrm{tr}\left(\mathbf{A}^{\mathrm{T}}\mathbf{B}\mathbf{B}^{\mathrm{T}}\mathbf{A}\right)^{1/2}\right\}^2}{\mathrm{tr}\left(\mathbf{A}^{\mathrm{T}}\mathbf{A}\right)} \tag{6}$$

Aquesta fórmula no és simètrica respecte de **A** i **B**, de manera que donarà valors diferents si és la configuració **A** o la **B** la que es pren de referència. Per solucionar aquest problema, s'escala aquesta expressió dividint l'equació (6) per $\mathrm{tr}\left(\mathbf{B}^{\mathrm{T}}\mathbf{B}\right)$:

$$M_0^2 = 1 - \frac{\left\{\mathrm{tr}\left(\mathbf{A}^{\mathrm{T}}\mathbf{B}\mathbf{B}^{\mathrm{T}}\mathbf{A}\right)^{1/2}\right\}^2}{\mathrm{tr}\left(\mathbf{A}^{\mathrm{T}}\mathbf{A}\right)\mathrm{tr}\left(\mathbf{B}^{\mathrm{T}}\mathbf{B}\right)} \tag{7}$$

Aquesta nova equació és simètrica i rep el nom d'*estadística de Procrustes*. És un índex de dissemblança comprès dins l'interval [0,1], i és més proper a zero com més semblants són les configuracions.

RESULTATS I DISCUSSIÓ

Connexió entre l'anàlisi Procrustes i l'alineament molecular

Siguin *A* i *B* dues molècules de m_A i m_B àtoms, respectivament, que comparteixen *n* àtoms de la seva estructura. Les coordenades dels àtoms d'aquestes molècules poden expressar-se de forma matricial amb les matrius $\mathbf{A}_0$ i $\mathbf{B}_0$, de dimensions $(3\times m_A)$ i $(3\times m_B)$. L'origen d'aquestes coordenades pot ser un càlcul d'optimització de geometria amb algun mètode quimicoquàntic, o bé un experiment de difracció de raigs X. S'assumeix que l'estructura dels cossos és rígida, i s'elimina en conseqüència la possibilitat que els àtoms puguin patir deformacions anisotròpiques en comparar-se. Com que les estructures són conegudes, s'escullen els àtoms de l'esquelet

comú i es creen dues noves matrius de coordenades **A** i **B**, que seran les que es compararan. La dimensió d'aquestes matrius és (3x*n*):

$$\mathbf{A} = \begin{pmatrix} \mathbf{a}_1 & \mathbf{a}_2 & \dots & \mathbf{a}_n \end{pmatrix} \wedge \mathbf{a}_i = \begin{pmatrix} a_x^{(i)} \\ a_y^{(i)} \\ a_z^{(i)} \end{pmatrix}$$

$$\mathbf{B} = \begin{pmatrix} \mathbf{b}_1 & \mathbf{b}_2 & \dots & \mathbf{b}_n \end{pmatrix} \wedge \mathbf{b}_i = \begin{pmatrix} b_x^{(i)} \\ b_y^{(i)} \\ b_z^{(i)} \end{pmatrix}$$

Els dos conjunts de coordenades atòmiques comunes poden ser comparats amb l'anàlisi de Procrustes seguint el desenvolupament teòric exposat anteriorment. Així, es considerarà la molècula *B* com a fixa, i es permetrà a la molècula *A* traslladar-se, rotar i dilatar-se o contreure's a fi de minimitzar la funció M^2.

Primer de tot, aquests conjunts de coordenades poden ser traslladats trivialment mitjançant una senzilla operació aritmètica. Aquesta translació situarà l'origen de coordenades al centroid de les dades satisfent la translació òptima **t** de l'anàlisi de Procrustes.

En principi, el terme δ de contracció/dilatació de l'equació de Procrustes no hauria de tenir cap rellevància aquí, atès que els conjunts de coordenades se suposen d'igual escala. L'únic paper possible apareix quan s'utilitzen unitats diferents per a cadascun dels dos sistemes; per exemple, àngströms i unitats atòmiques de distància. En aquest cas, el coeficient δ equivaldrà al factor de conversió entre ambdues unitats.

Molt més interessants que la translació i l'escalat de les coordenades moleculars resulta la rotació òptima que les orientarà adequadament. Aquest terme és el que proporciona la transformació més significativa per a l'alineament entre les dues molècules. Un cop centrats i escalats apropiadament ambdós conjunts de coordenades, la matriu **R** s'aplica sobre **A** per sobreposar les coordenades. El més rellevant d'aquest tractament és que, encara que la matriu de rotació **R**, de dimensions (3x3), ha estat construïda utilitzant el subconjunt format pels àtoms de l'esquelet comú, pot ésser aplicada al conjunt *complet* de coordenades atòmiques de la molècula transformada, $\mathbf{A}_0$:

$$\mathbf{A}_0' = \mathbf{R}^{\mathrm{T}}\mathbf{A}_0 = \left[\left(\mathbf{A}^{\mathrm{T}}\mathbf{B}\mathbf{B}^{\mathrm{T}}\mathbf{A}\right)^{1/2}\left(\mathbf{B}^{\mathrm{T}}\mathbf{A}\right)^{-1}\right]^{\mathrm{T}}\mathbf{A}_0 \qquad (8)$$

on s'ha assumit que la translació i l'escalat ja havien estat duts a terme prèviament. Aquesta darrera equació dóna l'alineament entre les dues molècules, i no només entre les subestructures comunes.

Un desenvolupament molt similar a l'anterior va ser utilitzat per McLachlan com a base per al seu algorisme d'alineament molecular (4). Així, l'autor permetia a una de les dues molècules rotar i traslladar-se fins a minimitzar una funció

d'error. El residual a minimitzar era una extensió de l'equació (1), on s'introduïen uns coeficients de pes $\frac{1}{2} w_i$ dins de cada terme del sumatori. McLachlan obtenia la mateixa solució per a la translació òptima (superposició de centroids), i el factor d'escala era ignorat.

No existeix una solució única per a resoldre el problema de la rotació òptima. Així, McLachlan (16) proposà una matriu que era derivada a partir d'un procediment iteratiu de minimització pel mètode del gradient conjugat. Diamond (17) suggerí una factorització de la solució en dues matrius: una de simètrica i una d'ortogonal. La millor solució, però, la va proporcionar Kabsch (18,19), que va derivar una expressió analítica per a la matriu de rotacions a partir d'un ajustament per mínims quadrats. El més interessant és que la solució al problema de Procrustes aportada per Sibson el 1978 (14) coincideix amb el mètode de Kabsch, publicat dos anys abans amb un enfocament diferent. Com s'ha esmentat, entre ambdós algorismes hi ha lleugeres diferències: mentre que l'anàlisi de Procrustes permet la dilatació o contracció del conjunt de dades, l'enfocament de McLachlan inclou uns coeficients de pes a les coordenades. En essència, però, tracten el mateix problema i coincideixen també en la solució.

Reinterpretació de l'estadística Procrustes

Un dels efectes habitualment detectables en l'estructura molecular de compostos d'una mateixa família és la distorsió de l'esquelet comú deguda a la interacció amb els substituents no comuns. Aquesta interacció pot donar lloc a variacions en les distàncies d'enllaç interatòmiques, en angles, etc., que poden evidenciar-se experimentalment mitjançant tècniques espectroscòpiques.

Com s'ha esmentat, el valor minimitzat de M^2 dóna una mesura de la semblança entre les dues configuracions. Seguint el raonament exposat, l'estadística de Procrustes, quan es calcula per a coordenades atòmiques de l'esquelet comú de dues molècules, donarà una mesura quantitativa dels efectes de deformació d'aquest esquelet produïda per la resta de substituents. Com més proper a zero sigui l'índex, més similars seran els dos conjunts de coordenades i, en conseqüència, menor serà la distorsió de llur estructura.

A més, l'expressió $1-M_0^2$ (vegeu equació (7)) pot interpretar-se com el quadrat d'un pseudoíndex de Carbó, ja que el denominador inclou el producte escalar dels descriptors de les molècules A i B per separat, mentre que el numerador té la forma d'un producte creuat entre A i B:

$$C_{AB}^2 = \frac{\left|\langle A|B\rangle\right|^2}{\langle A|A\rangle\langle B|B\rangle} \qquad (9)$$

Aquest índex de semblança està comprès en l'interval [0,1], més proper a 1 com més similars siguin els dos conjunts de dades comparats.

Extensió a diferents conjunts de coordenades atòmiques

L'objectiu, tant del mètode de McLachlan com el de l'anàlisi de Procrustes, era superposar dos conjunts de coordenades entre les quals existia una relació 1:1. Tanmateix, res impedeix que el problema es generalitzi a un nombre arbitrari d'objectes. Sota aquestes consideracions, Gerber i Müller van extendre el treball de McLachlan al cas de voler comparar més de dos conjunts de coordenades atòmiques (20). Tot i que no s'incidirà en aquesta generalització, resulta interessant assenyalar que paral·lelament també ha estat descrita una extensió dins la teoria de l'anàlisi de Procrustes, que ha dut a l'anomenat *anàlisi de Procrustes generalitzat* (GPA), introduït inicialment per Kristof i Wingersky (21) i desenvolupat per Gower (22) i Ten Berge (23).

CONCLUSIONS

En aquest treball s'ha mostrat com l'enfocament de McLachlan al problema de la superposició molecular coincideix amb una metodologia ben coneguda al camp de l'Anàlisi Multivariant: l'anomenat anàlisi de Procrustes. La solució per a la matriu de rotació proposada per Sibson coincideix amb la de Kabsch, desenvolupada independentment.

Val a dir que l'algorisme, encara que extremadament ràpid i eficient, ha esdevingut inoperant a causa d'una dificultat insalvable: la intervenció externa de l'investigador, que ha de seleccionar els àtoms pertanyents a l'esquelet comú i recopilar les seves coordenades en submatrius. Molts altres mètodes han estat proposats des d'aleshores, basats en altres principis i capaços d'orientar grans sèries de compostos d'una forma ràpida, efectiva i completament transparent per a l'usuari. La metodologia discutida roman útil únicament en sistemes molt grans, com ara proteïnes, on el temps de càlcul requerit per a executar els algorismes alternatius els fa inaplicables.

AGRAÏMENTS

Aquest treball ha estat finançat parcialment amb el contracte de la Comissió Europea #ENV4-CT97-0508. Els autors volen expressar el seu agraïment a Wolfgang Kabsch (Institut Max Planck, Heidelberg) per la seva amabilitat i les seves interessants observacions.

Bibliografia

1. WILLETT, P.; WINTERMAN, V.; BAWDEN, D. Implementation of nearest-neighbour searching in an online chemical structure search system. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 36-41
2. CRAMER III, 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

3. CARBÓ, R.; ARNAU, J.; LEYDA, L. How similar is a molecule to another? An electron density measure of similarity between two molecular structures. *Int. J. Quantum Chem.* **1980**, *17*, 1185-1189
4. McLACHLAN, A.D. A mathematical procedure for superimposing atomic coordinates of proteins. *Acta Cryst.* **1972**, *A28*, 656
5. Veure, per exemple: COX, T.F.; COX, M.A.A. *Multidimensional Scaling*. Chapman & Hall, London, 1994
6. BORG, I.; GROENEN, P. *Modern Multidimensional Scaling*. Springer, New York, 1997
7. GREEN, B.F. The orthogonal approximation of an oblique structure in factor analysis. *Psychometrika*, **1952**, *17*, 429-440.
8. SCHÖNEMANN, P.H. A generalized solution of the orthogonal Procrustes problem. *Psychometrika*, **1966**, *31*, 1-10.
9. SCHÖNEMANN, P.H.; CARROLL, R.M. Fitting one matrix to another under choice of central dilation and a rigid motion. *Psychometrika*, **1970**, *35*, 246-256
10. ROBERT, D.; CARBÓ-DORCA, R. A formal comparison between Molecular Quantum Similarity Measures and Indices. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 469-475
11. ROSE, V.S.; RAHR, E.; HUDSON, B.D. The Use of Procrustes Analysis to Compare Different Property Sets for the Characterization of a Diverse Set of Compounds. *Quant. Struct.-Act. Relat.* **1994**, *13*, 152-158
12. GREENWOOD, R. Use of generalised Procrustes Analysis in QSAR studies with several sets of descriptors. *UK QSAR Discussion Group Meeting*, 1998
13. KRZANOWSKI, W.J.; MARRIOTT, F.H.C. *Multivariate Analysis. I. Distributions, Ordination and Inference*. Edward Arnold, London, 1994; p. 134-137
14. SIBSON, R. Studies in the robustness of multidimensional scaling: Procrustes statistics. *J. R. Statist. Soc.* **1978**, *B40*, 234-238
15. MARDIA, K.V.; KENT, J.T.; BIBBY, J.M. *Multivariate Analysis* Academic Press. London, 1978
16. McLACHLAN, A.D. Rapid comparison of protein structures. *Acta Cryst.* **1982**, *A38*, 871-873
17. DIAMOND, R. On the comparison of conformations using linear and quadratic transformations. *Acta Cryst.* **1976**, *A32*, 1-10
18. KABSCH, W. A solution for the best rotation to relate two sets of vectors. *Acta Cryst.* **1976**, *A32*, 922-923
19. KABSCH, W. A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Cryst.* **1978**, *A34*, 827-828
20. GERBER, P.R.; MÜLLER, K. Superimposing several sets of Molecules. *Acta Cryst.* **1987**, *A43*, 426-428
21. KRISTOF, W.; WINGERSKY, B. Generalization of the orthogonal Procrustes rotation procedure to more than two matrices. A *Proceedings, 79th Annual Convention of the American Psychological Association*, p. 81-90
22. GOWER, J.C. Generalized Procrustes Analysis. *Psychometrika*, **1975**, *40*, 33-51
23. TEN BERGE, J.M.F. Orthogonal Procrustes rotation for two or more matrices. *Psychometrika*, **1977**, *42*, 267-276