SCIENTIA gerundensis, 20: 87-93 (1994)

UN ALGORISME PER AL RECONEIXEMENT AUTOMÀTIC DE GRAFOS DICOTOMITZATS

E. Besalú

Institut de Química Computacional
Universitat de Girona
17071 Girona, Catalunya (Espanya)

RESUM

Es descriu un algorisme computacional que permet el reconeixement de grafos que presenten característiques dicotòmiques. Alhora que s'exposa l'aplicació purament geomètrica, es descriu conjuntament l'aplicació pràctica que això té en el camp de la Semblança Molecular. Es mostra que és una eina molt valuosa per tal d'inferir les relacions que hi ha entre propietats físico-químiques o farmacològiques en un grup arbitrari de molècules.

RESUMEN

Se describe un algoritmo computacional que permite el reconocimiento de grafos que presentan características dicotómicas. Conjuntamente, se exponen la aplicación puramente geométrica y la aplicación práctica del algoritmo en el campo de la Semejanza Molecular, mostrando así que constituye una herramienta muy valiosa para inferir las relaciones entre propiedades físico-químicas o farmacológicas de un grupo arbitrario de moléculas.

ABSTRACT

A computational algorithm which allows the recognition of graphs presenting dicotomic characteristics is described. Its geometric application and also the practic implementation in the Molecular Similarity field is exposed. It is also shown how the algorithm can be used as a useful tool to deduce intrinsic relationships between the physicochemical properties of a molecular group.

Keywords: Graphs, Pattern Recognition, Quantum Similarity, Molecular Quantum Similarity.

INTRODUCCIÓ

Al nostre Laboratori s'ha fet un desenvolupament progressiu en el camp de la Semblança Molecular que ha quedat palès en diferents treballs (1). S'ha volgut construir una definició rigorosa de les Mesures de Semblança Quàntica i proveir les bases del seu significat mecano-quàntic.

Hi ha moltes maneres d'enfrontar-se al problema de la Semblança Molecular, i això ha donat lloc a la formació i al creixement de nombrosos grups que treballen sobre el mateix tema (2). La nostra visió consisteix a seguir una metodologia ben definida, que ha demostrat que dóna molt bons resultats (1,3). Alhora, aquest *modus operandi* està sòlidament lligat a la nostra idea de concebre mètodes que siguin interpretatius tant des del punt de vista conceptual químic com des del punt de vista mecano-quàntic.

MESURES DE SEMBLANÇA MOLECULAR QUÀNTICA

Es pot concebre un Estudi de Semblança Molecular Quàntica (ESMQ) d'un conjunt de molècules de la manera següent. Imaginem que d'algunes d'aquestes molècules, se'n coneix el valor d'una determinada propietat fisico-química o un índex d'activitat bioquímica o farmacològica. Les molècules es poden agrupar en subconjunts que tinguin com a elements les que presentin el valor de la propietat dins un determinat rang. És possible que no es conegui el valor de la propietat per a un subconjunt molecular no catalogat del conjunt total sotmès a estudi. El que es pretén és predir els valors de les propietats d'aquestes molècules.

Fonamentalment, el que proveeix l'ESMQ aplicat sobre el conjunt molecular en estudi és una matriu anomenada Matriu d'Índexs de Semblança (MIS). La MIS conté tota la informació relacionada amb la possible ordenació del conjunt molecular respecte a la propietat investigada. Per extreure'n aquesta informació hi ha molts mètodes, alguns relacionats amb el reconeixement de formes o anàlisi taxonòmica (4). Al nostre laboratori, usualment fem una anàlisi de la MIS amb un suport gràfic: es considera cada molècula representada per un vector corresponent a una fila o columna de la MIS, llavors es diagonalitza aquesta matriu i es projecten els vectors coordenades de cada molècula en els plans definits per cada parella d'eixos principals de la matriu. La tasca consisteix ara a iterar per a tots els plans possibles i, per a cada pla, rotar per a tots els angles possibles. Per a cada iteració s'aconsegueix una representació del conjunt molecular o imatge bidimensional formada per punts disseminats en un pla. Cada punt representa una determinada molècula. El que es pretén és detectar quina gràfica mostra les molècules amb propietats conegudes agrupades correctament. L'anomenat *Principi de Mendeleev* (3) assegura que, en el moment en què es troba aquesta figura, la posició que ocupin les molècules amb propietats desconegudes dóna una pista qualitativa de quin valor ha de tenir l'esmentada propietat. Aquest és l'objectiu d'un ESMQ: predir valors de propietats de molècules no catalogades o desconegudes.

És en aquesta darrera fase de l'estudi de Semblança Molecular on, fins ara, es requeria molt de treball manual per tal d'anar iterant sobre tots els plans i rotar per tots els angles i analitzar visualment totes les figures que es generen. El nombre d'imatges per analitzar és molt gran. Per exemple, per a un conjunt de 10 molècules, hi ha 45 plans per estudiar i, si per cada pla es rota sobre un eix amb increments de 5 graus, es tenen un total de 25.920 figures per estudiar. Per a un conjunt de 15 molècules, aquesta xifra puja fins a un total de 98.280. D'altra banda, l'anàlisi manual no assegura una reproductibilitat total del mètode.

Ja s'ha dit que el que es pretén és triar, d'entre totes les figures possibles, la que representa els grups de molècules amb propietats conegudes el màxim de classificats. S'entén com a figura que dóna compte d'una bona classificació aquella que mostra les molècules separades per subconjunts que agrupin molècules amb propietats iguals o semblants, alhora que aquests subconjunts presentin un solapament mínim entre ells sobre el pla on es representen. Com a exemple d'això es pot veure a la **Figura 1** la representació del conjunt de 9 cloro-fluoro-metans. La propietat que s'estudia és el punt d'ebullició i les molècules que estan representades per la mateixa figura tenen un valor comparable d'aquesta propietat (vegeu **Taula 1**). Així, es tenen 2 subconjunts moleculars. A la **Figura 2** es pot veure el mateix con-

junt, però projectat en un altre pla i després d'efectuar una determinada rotació. Es pot apreciar com aquest punt de vista classifica bé els subconjunts segons el criteri que s'ha exposat més amunt.

El que aquí es vol exposar és l'algorisme que recentment ens ha permès triar la millor projecció d'una manera totalment automàtica i reproduïble. Aquest algorisme és capaç de seleccionar de forma automàtica a la **Figura 2** com a preferent respecte a la **Figura 1**. És el que hem anomenat l'algorisme del reconeixement de grafos dicotomitzats. A continuació s'exposa el seu fonament.

L'ALGORISME DE RECONEIXEMENT DE GRAFOS DICOTOMITZATS

Es defineixen dos subconjunts de dos elements cadascun. A les **Figures 3a** i **3b** es mostra aquest conjunt representat de dues maneres diferents. Tenim 4 cercles i definim cada subconjunt com el que agrupa cercles del mateix color. A més a més, els elements del mateix subconjunt es poden considerar units per un segment imaginari que també s'ha representa, formant així un parell de grafos en cada cas. Anomenem l'estructura **3a** com un parell de grafos no dicotomitzats o estructura del Tipus I. A la **Figura 3b** es mostra el mateix conjunt però amb una estructura dicotomitzada. Aquesta estructura l'anomenem estructura del Tipus II. L'estructura del Tipus II és la que ens interessa detectar mitjançant el nostre algorisme. És clar que la **Figura 1** és una superposició de moltes estructures dels Tipus I i II. A tall d'exemple, hi ha marcats dos segments imaginaris que mostren un cas d'estructura del Tipus I. A la **Figura 2** només s'aprecien estructures del Tipus II, com la que hi ha assenyalada. L'algorisme que es defineix més avall detecta les estructures dels Tipus I i II representades a les **Figures 3a** i **3b**.

Definició de l'algorisme

L'algorisme de reconeixement de grafos dicotòmics es basa en el principi

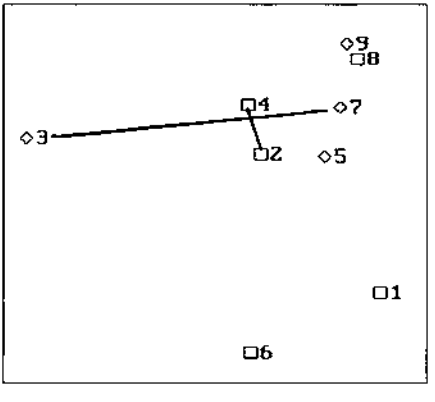

Figura 1: Representació gràfica dels 9 cloro-fluoro-metans.

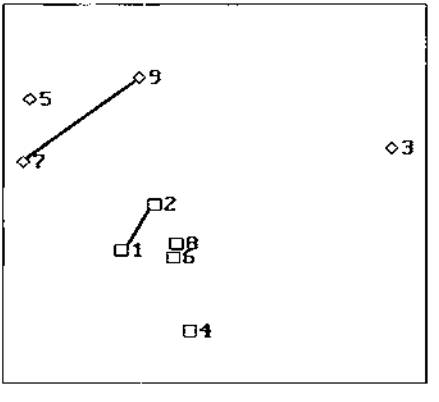

Figura 2: Una ordenació més perfecta de les molècules sobre el pla.

següent: una parella de grafos presenta l'estructura dicotomitzada si i només si es creuen els segments que uneixen els elements o vèrtexs d'un mateix graf. Per tant, es tracta d'un algorisme ideat per detectar quan hi ha aquest tipus de creuaments entre aquests segments.

La manera de codificar això en forma d'algorisme es pot concebre de la manera següent (vegeu **Figura 4**):

1. A partir de les coordenades de cada element en el pla (**A**,**B**,**C**,**D**), es generen les rectes r:$Ax+By+C=0$ i s:$A'x+B'y+C'=0$ que contenen els segments que uneixen els dos vèrtexs d'un mateix graf. Així, s és la recta que passa pels punts **A** i **B** i r és la que passa per **C** i **D**.
2. S'avalua la quantitat $S=(Ax_a+By_a+C)(Ax_b+By_b+C)$.
3. Si S>0, es conclou que no hi ha creuament i el conjunt és dicotòmic. En cas contrari,
4. si S≤0, s'avalua la quantitat $T=(A'x_c+B'y_c+C')(A'x_d+B'y_d+C')$.
5. Si T>0, es conclou que no hi ha creuament i el conjunt és dicotòmic.
6. Si no, es conclou que hi ha creuament i el conjunt no és dicotòmic.

El fonament matemàtic de l'algorisme es pot trobar més avall a l'Apèndix. Aquest fonament ha estat implementat al programa ND-CLOUD de representació gràfica de conjunts moleculars.

Aplicació pràctica

Per aplicar l'algorisme en un cas real com el de la **Figura 1** o la **Figura 2**, cal generar totes les parelles de grafos del Tipus I o II continguts a la figura. Així, cal generar totes les possibles parelles d'elements de cada subconjunt molecular. Cada parella genera un graf i cal considerar totes les parelles possibles de grafos pertanyents a diferents grups moleculars. Llavors s'aplica l'algorisme entre aquestes parelles de grafos. El que es fa, per tant, és generar tots els possibles sistemes com els de les **Figures 3a** i **3b** que es troben a les **Figures 1** o **2**. Cada creuament trobat es computa i s'acumula en un comptador. Al final, a cada figura per analitzar se li pot associar una etiqueta que indica quants creuaments del Tipus I conté. Es tracta de seleccionar la figura amb el menor nombre de creuaments, que és la que té un major grau de dicotomització. Aquest procediment es pot generalitzar

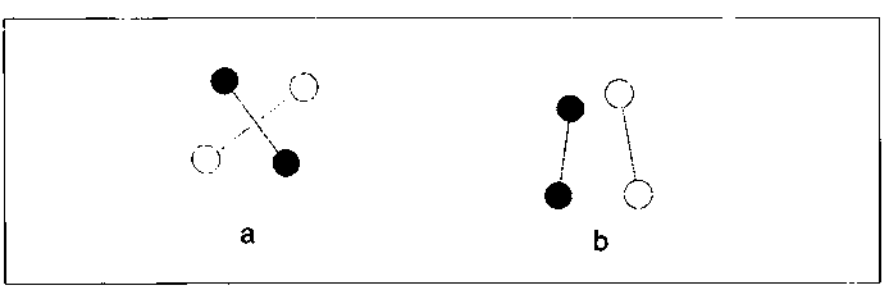

Figura 3: a) Estructura no dicotomitzada o del Tipus I. b) Estructura dicotomitzada o del Tipus II.

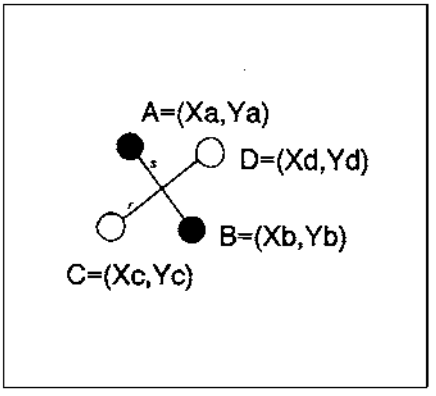

Figura 4: Definició de les rectes *r* i *s*. Vegeu el text per al seu significat.

Molècula	Punt d'ebullició	Número
CF_4	-129.0	8
CH_4	-164.0	1
CHF_3	-82.2	6
CH_3F	-78.4	2
CH_2F_2	-51.6	4
CH_3Cl	-24.2	3
CH_2Cl_2	40.0	5
$CHCl_3$	61.7	7
CCl_4	76.5	9

Taula 1: Definició dels grups moleculars respecte al seu punt d'ebullició (°C).

fàcilment quan el conjunt molecular està subdividit en més de dos subconjunts.

L'algorisme és molt ràpid a pesar del nombre de subsistemes del Tipus I o II que es generen per cada figura. Per exemple, en el cas de la **Figura 1** cal aplicar l'algorisme entre 60 parelles de grafos. D'altra banda, per a moltes figures, no cal generar tots els subsistemes de grafos que les formen. Això és així perquè es busca un nombre mínim de creuaments. Per tant, quan es van analitzant les diferents figures, sempre es té guardada la que presenta el mínim nombre de creuaments aconseguit fins llavors. Moltes figures, una vegada es veu que el seu comptador de creuaments ja sobrepassa el mínim, ja es poden descartar sense acabar de generar tots els subsistemes de grafos que les formen.

Usualment, quan ja s'ha iterat per a totes les figures, el que es troba és que n'hi ha moltes que són topològicament equivalents, en el sentit que presenten el mateix nombre de creuaments mínim. En aquest cas, cal aplicar un altre criteri per escollir un sol grafo d'aquest conjunt de figures. Actualment tenim implementats en el programa ND-CLOUD tres criteris que, a la pràctica, no presenten diferències notables entre ells:

1. Escollir la figura que presenti una dispersió més gran dels punts que la formen.
2. Escollir la que presenta una dispersió més gran dels centres de massa de cada subconjunt de punts.
3. Escollir la que presenti les distàncies més grans entre els elements més propers entre subconjunts diferents.

CONCLUSIONS

S'ha definit un algorisme computacional fàcilment programable. Aquest algorisme permet l'anàlisi i selecció automàtica d'un tipus de gràfics que són útils per treure conclusions en un Estudi de Mesures de Semblança Molecular Quàntica. El procediment descrit permet estalviar moltes hores de treball en l'anàlisi dels resultats de Mesures de Semblança.

APÈNDIX: FONAMENT MATEMÀTIC DE L'ALGORISME

El fonament matemàtic de l'algorisme que s'acaba d'exposar és ben senzill i conegut: Si tenim una recta definida en un pla cartesià, r:$Ax+By+C=0$. Si tenim també dos punts $\mathbf{A}=(x_a,y_a)$ i $\mathbf{B}=(x_b,y_b)$. Si els dos punts estan situats a banda i banda de la recta, es compleix sempre que $(Ax_a+By_a+C)(Ax_b+By_b+C)<0$.

AGRAÏMENTS

Aquest treball ha estat finançat per la CICYT-CIRIT, el Programa de Química Fina de la Generalitat de Catalunya amb la beca núm. QFN91-4206. Els càlculs han estat fets a l'Institut de Química Computacional de la Universitat de Girona i s'han utilitzat els programes QMOLSIM i ND-CLOUD. L'autor és beneficiari d'una beca de Formació d'Investigadors del Departament d'Ensenyament de la Generalitat de Catalunya i agraeix els suggeriments donats pel Dr. M. Solà, que li han permès redactar aquest treball d'una manera molt més estructurada i comprensible.

Bibliografia

1. a) CARBÓ R., ARNAU M. and LEYDA L., Int.J.Quantum Chem. *1980* **17** 1185.
   b) CARBÓ R. i ARNAU C., "Molecular Engineering: a general approach to QSAR", dins: de las Heras F.G. and Vega S. (ed.) *Medicinal Chemistry Advances*, Pergamon Press, Oxford, 85, *1981*.
   c) CARBÓ R., DOMINGO L., Int.J.Quantum.Chem. *1987* **23** 517.
   d) CARBÓ R., CALABUIG B., Comp. Phys. Commun. *1989* **55** 117.
   e) CARBÓ R., CALABUIG B., "Molecular Similarity and Quantum Chemistry", capítol 6, pàg. 147, a la referència (2k).
   f) CARBÓ R., CALABUIG B., Proceedings del XIX Congresso Internazionale dei Chimici Teorici dei Paesi di Espressione Latina, Roma, Itàlia, Setembre 10-14, 1990. J.Mol.Struct. (Teochem), *1992* **254** 517.
   g) CARBÓ R. i CALABUIG B., J.Chem.Inf.Comput.Sci.*1992* **32** 600.
   h) CARBÓ R. i CALABUIG B., "Quantum Similarity", in Fraga S. (ed.), *Structure, Interactions and Reactivity*, Elsevier Pub., Amsterdam, *1992*.
   i) CARBÓ R. i CALABUIG B., Int.J.Quantum Chem. *1992* **42** 1681.
   j) CARBÓ R. i CALABUIG B., Int.J.Quantum Chem. *1992* **42** 1695.
   k) CARBÓ R., CALABUIG B., BESALÚ E. and MARTÍNEZ A., Molecular Engineering. *1992* **2** 43.
   l) CARBÓ R., BESALÚ, E., CALABUIG, B. and VERA L.; Adv.Quan.Chem. (en premsa)

2. a) COOPER D.L. i ALLAN N.L., J.Chem.Soc., Faraday Trans. *1987* **83** 449.
   b) COOPER D.L. i ALLAN N.L., J.Computer-Aided Mol.Design. *1989* **3** 253.
   c) COOPER D.L. i ALLAN N.L., J.Am.Chem.Soc. *1992* **114** 4773.
   d) CIOSLOWSKI J. i FLEISCHMANN E.D., J.Am.Chem.Soc., *1991* **113** 64.
   e) CIOSLOWSKI J. i CHALLACOMBE M., Int.J.Quant.Chem., *1991* **25** 81.
   f) ORTIZ J.V. i CIOSLOWSKI J., Chem.Phys.Lett., *1991* **185** 270.
   g) CIOSLOWSKI J. i SURJÁN P.R., Jour.Mol.Struct.(Theochem), *1992* **255** 9.

h) PONEC R. i STRNAD M.; Collect.Czech.Chem.Commun. *1990* **55** 2583.
i) PONEC R. i STRNAD M.; J.Phys.Org.Chem. *1991* **4** 701.
j) PONEC R. i STRNAD M.; Int.J.Quantum Chem. *1992* **42** 501.
k) JOHNSON M.A. i MAGGIORA G. (ed.), *Concepts and Applications of Molecular Similarity*, John Wiley & Sons Inc, New York, *1990*.
l) HODGKIN E.E. i RICHARDS W.G., Int.J.Quant.Chem. *1987* **14** 105.
m) GOOD A.C., HODGKIN E.E. i RICHARDS W.G., J.Chem.Inf.Comput.Sci. *1992* **32** 188.
n) MARTIN M., SANZ F., CAMPILLO M., PARDO L., PÉREZ J. i TURMO J., Int.J.Quant.Chem. *1983* **23** 1627.
o) MARTIN M., SANZ F., CAMPILLO M., PARDO L., PÉREZ J., TURMO J. i AULLO J.M., Int.J.Quant.Chem. *1983* **23** 1643.
p) SANZ F., MARTIN M., PÉREZ J., TURMO J., MITJANA A. i MORENO V., in: Dearden J.C. (ed.), *Quantitative Approaches to Drug Design*, Elsevier, Amsterdam, *1983*.
q) SANZ F., MARTIN M., LAPEÑA F. i MANAUT F., Quant.Struct.-Act.Relat. *1986* **5** 54.
r) SANZ F., MANAUT F., JOSÉ J., SEGURA J. , CARBO M. i DE LA TORRE R., Jour.Mol.Struct. (Theochem), *1988* **170** 171.
s) LUQUE F.J., SANZ F., ILLAS F., POUPLANA R. i SMEYERS Y.G., Eur.J.Med.Chem, *1988* **23** 7.

3. CARBÓ R. i BESALÚ E. Proceedings of the first Girona Seminar on Molecular Similarity (en premsa).

4. a) TOU J.T. i GONZÁLEZ R.C., *Pattern Recognition Principles*. Addison-Wesley Reading, MA *1974*.
b) HARDIGAN J.A., *Clustering Algorithms*. Wiley, New York, *1975*.
c) SNEATH P.H.A. i SOKAL R.R., *Numerical Taxonomy*, W.H.Freeman, San Francisco, *1973*.
d) JARDINE N. i SIBSON R., *Mathematical Taxonomy*, Wiley, New York, *1977*.