ALGORITMO DE CLASSIFICAÇÃO HIERÁRQUICA APLICADO AO PLANEJAMENTO DE FÁRMACOS

Maria da Paz N. Moreno, Nereide S. S. Magalhães, Sócrates C. H. Cavalcanti e Antonio J. Alves*
Universidade Federal de Pernambuco - Departamento de Farmácia - Av. Prof. Artur Sá - S/N - Cid. Universitária - 51740-520 - Recife - PE

Recebido em 5/9/95; aceito em 19/7/96

HIERARQUICAL CLUSTER ANALYSIS APPLIED TO DRUG DESIGN. An algorithm is introduced which allows a substituents choice for the original compound of a chemical series in order to achieve a high biological activity. This procedure is based upon a relationship between the biological activity and physicochemical parameters. Substituents are classified by cluster analysis, building a cluster tree for each site. Active compounds are obtained by a choice of paths in these trees. This framework is particularly suitable in chemical series having two or more sites, dealing with a large number of possible substituents per site. It has a great flexibility and easy interpretation. Application to a series of 2-aryl-1,3-indandiones illustrates that approach.

Keywords: cluster analysis; biological activity; indandiones derivatives; Q. S. A. R.

INTRODUÇÃO

O grande desafio no planejamento e síntese de novos fármacos é a escolha criteriosa de substituintes para modificação de um composto original gerando uma série química de derivados que apresentem atividade biológica. Há um grande número de possíveis compostos que podem ser sintetizados a partir de uma simples substituição em cada sítio da molécula protótipo (crescimento exponencial com o número de sítios de substituição). Torna-se, portanto, imperativo a indicação de derivados com potencial atividade biológica para que a síntese de novos fármacos seja um procedimento racional.

Em 1972, Topliss[1] descreveu um modelo de árvore de decisão (árvore de agrupamentos) como auxílio na escolha racional de substituintes. A proposta é a seleção de uma série inicial de substituintes, onde são evitados substituintes isometricamente bioisósteros, visto que eles seriam providos de informações semelhantes e contribuiriam provavelmente da mesma forma na atividade biológica.

Na literatura, são mencionados procedimentos numéricos para agrupar substituintes em subgrupos objetivamente diferentes. Um destes métodos, conhecido como classificação hierárquica[2], é um tipo de reconhecimento de padrões utilizado na solução de problemas químicos por Kowalski e Bender[3,4].

Dada uma série química, existem subgrupos de substituintes mais ou menos homogêneos com relação a vários parâmetros físico-químicos relevantes em uma relação quantitativa entre a estrutura e a atividade (Q.S.A.R.). Estes substituintes constituem um dendrograma (diagrama de árvore) quando tratados pelo método de classificação hierárquica. Os agrupamentos são formados a partir de um procedimento objetivo, através da definição de uma métrica, que permite avaliar a similaridade entre diferentes substituintes. Cada substituinte é associado a um vetor multidimensional no espaço Euclidiano cujas coordenadas indicam os valores dos parâmetros físico-químicos. O exemplo usual considera a distância Euclidiana entre os pontos no espaço neste espaço de parâmetros como uma métrica apropriada. No planejamento de novos derivados químicos, considera-se uma substituição em cada sítio ativo da molécula original empregando substituintes com parâmetros físico-químicos qualitativamente e quantitativamente diferentes. Este procedimento pode auxiliar a encontrar, mais rapidamente, compostos da série química com potencial atividade biológica. Além disso, deve-se considerar que a série de derivados planejada apresente síntese viável em laboratório.

Em 1987, Giacca e colaboradores[5] utilizaram um algoritmo de Classificação Hierárquica conhecido como *average linkage clustering*, no qual os indivíduos ou grupos formados fundem-se quando suas similaridades são maiores. O método é fundamentado em uma matriz de similaridades que permite encontrar indivíduos de mais alta verossimilhança, excluindo comparações entre os indivíduos iguais. O programa utilizado pelos autores foi escrito em *Microsoft Basic* e representa uma classificação hierárquica aglomerativa.

Os resultados foram apresentados em um dendrograma onde o eixo vertical representa a função de semelhança avaliada.

A classificação hierárquica pode ser aplicada em pratica-mente todas as áreas do conhecimento científico, sendo muito útil em economia[6], reconhecimento de padrões[7], controle de infecção hospitalar[8], medicina[9-13], planejamento de fármacos[14], etc.

Atualmente, a classificação hierárquica é utilizada em planejamento de fármacos no estágio precedente à análise de Q. S. A. R., agrupando dois ou mais substituintes em cada passo e obtendo uma árvore de agrupamento para posterior escolha dos substituintes[15].

O modelo de classificação descrito por Cavalcanti[16], tem como base um método estatístico de classificação hierárquica (*cluster analysis*). A técnica agrupa substituintes em subgrupos homogêneos com base em suas similaridades. Neste método é possível agrupar dois ou mais parâmetros em um só passo, simplificando a análise hierárquica.

O presente trabalho descreve um Método de Classificação Hierárquica aplicável ao planejamento de fármacos. Este modelo é implementado por um aplicativo conversacional (NEWCLUS) que executa os passos necessários ao agrupamento, fornecendo diferentes níveis de agrupamento. Esta versão amigável do *software* foi desenvolvida visando implementar maior praticidade na sua utilização.

O modelo prevê os tipos de substituintes que podem ser introduzidos numa molécula orgânica visando propiciar uma maior atividade biológica dos derivados. Na primeira etapa, os substituintes a serem introduzidos numa molécula protótipo são classificados usando o método de classificação hierár-quica. Uma árvore de agrupamento é então construída para cada sítio de substituição da molécula. Os diferentes com-postos são obtidos através da escolha de caminhos nas árvores de agrupamento.

A classificação hierárquica realizada pelo programa NEWCLUS é ilustrada através da aplicação em uma série de derivados da 2-aril-1,3-indanodionas com atividade anti-inflamatória[17].

METODOLOGIA

No método de classificação hierárquica, considera-se um número K de parâmetros, e um número de N possíveis substituintes por sítio de substituição da molécula. Os dados dos parâmetros físico-químicos de todos os substituintes podem ser descritos por uma matriz [X(i, j)], onde i varia de 1 a N e j varia de 1 a K. A distância entre dois substituintes é medida pela distância Euclidiana entre os pontos representativos (vetores de parâmetros) no espaço Euclidiano $\Re^K$.

Os parâmetros considerados não necessariamente estão na mesma escala. No entanto, o processo de *Classificação* exige normalização dos valores desses parâmetros. A equação de normalização determina o desvio de cada parâmetro em relação ao seu valor médio e expressa estes valores em unidades de desvio padrão (escore reduzido).

No agrupamento hierárquico, os pontos mais próximos são agrupados segundo suas distâncias, definindo um subgrupo ou *pseudo*ponto. O procedimento continua até que um único grupo seja formado.

Em qualquer nível de agrupamento, os pontos são objetivamente mais similares dentro de um mesmo grupo e diferentes em relação aos outros fora deste grupo. A diferença entre elementos internos dos grupos é minimizada, enquanto que a diferença entre os grupos é maximizada.

1. Normalização

Os dados [X(i,j)] introduzidos no NEWCLUS sob forma matricial são inicialmente normalizados. Esta etapa envolve o cálculo das médias e desvios dos valores dos parâmetros físico-químicos.

- Cálculo das médias e desvios padrões dos valores de X por coluna, ou seja, para cada parâmetro determinado por j:

$$\overline{X}_j = \frac{1}{N}\sum_{i=1}^{N} X(i,j) \quad \text{e} \quad d.p.(j) = \sqrt{\frac{\sum_{i=1}^{N}\left[X(i,j) - \overline{X}_j\right]^2}{N-1}} \qquad (2.1)$$

- Cálculo dos valores normalizados para cada X(i, j):

$$N(i,j) = \frac{X(i,j) - \overline{X}_j}{d.p.(j)} \qquad (2.2)$$

onde, N é o número de substituintes, X (i, j) é o valor a ser normalizado, $\overline{X}_j$ é o valor da média entre os valores para cada j, *N*(i, j) é o valor normalizado e d.p.(j) é o desvio padrão para os valores, onde j (índice do parâmetro físico-químico) é fixo.

Calculados os valores normalizados, obtêm-se uma nova matriz [N(i,j)] (N x K) com N objetos (substituintes) e K variáveis (parâmetros).

A matriz normalizada facilita a comparação dos dados que a constituem. Na comparação de duas colunas distintas é examinada a relação entre as variáveis (parâmetros físico-químicos dos substituintes) e na comparação entre duas linhas é examinada a relação entre dois objetos (substituintes químicos).

2. Matriz de distâncias

A matriz de distâncias armazena todas as distâncias entre os substituintes, em função de suas similaridades físico-químicas. Trata-se de uma matriz quadrada N x N, onde N é o número de substituintes. A i-ésima linha e i'-ésima coluna contém o valor da distância entre os substituintes de índices i e i'.

Calcula-se as distâncias entre todos os pares de substituintes. Obviamente, a distância entre um substituinte e ele próprio é zero, produzindo uma diagonal nula na matriz de distâncias. Devido à comutatividade da métrica utilizada, a distância relativa à dois determinados substituintes i e i' é idêntica àquela entre os substituintes i' e i. Assim, a matriz de distâncias apresenta duas características básicas: (i) possui diagonal principal nula; e (ii) é uma matriz simétrica, i.e., d(i,i') = 0 ($\forall$ i = i') e d(i,i') = d(i',i), ($\forall$ i, i').

As distâncias entre dois substituintes são calculadas com base na distância Euclidiana[2], e são obtidas pela fórmula:

$$d(i,i') = \sqrt{\sum_{j=1}^{K}\left[N(i,j) - N(i',j)\right]^2}, \qquad (2.3)$$

onde $\{N(i,j)\}_{j=1}^{K}$ e $\{N(i',j)\}_{j=1}^{K}$ são os pontos no espaço de parâmetros $\Re^K$ para os quais são calculadas as distâncias. O cálculo da distância considera os valores normalizados dos parâmetros N(i ,j) ao invés de X(i , j). A distância entre dois substituintes de rótulos i e i' é dada por d(i,i') com i e i' variando de 1 a N, construindo uma matriz:

$$[d(i,i')] = \begin{bmatrix} d(1,1) & d(1,2) & d(1,3) & \ldots & d(1,N) \\ d(2,1) & d(2,2) & d(2,3) & \ldots & d(2,N) \\ d(3,1) & d(3,2) & d(3,3) & \ldots & d(3,N) \\ . & . & . & . & \\ . & . & . & . & \\ . & . & . & . & \\ d(N,1) & d(N,2) & d(N,3) & \ldots & d(N,N) \end{bmatrix}$$

3. Distância admissível

Esta etapa descreve o critério de agrupamento dos substituintes. A matriz de distâncias é analisada em busca de um limiar chamado de *distância admissível*. Esta análise envolve um método para obtenção da distância limite onde os substituintes são considerados "próximos" e tidos como *vizinhos*. Trata-se de uma maneira racional de escolher um limiar para o agrupamento de substituintes.

Um substituinte Si_1 é dito vizinho de um outro substituinte Si_2 , se e somente se a distância entre eles for menor ou igual a distância admissível. Esta distância é definida como uma quota limite abaixo da qual todos os substituintes podem ser agrupados. Uma maneira formal de escolha para este nível consiste em considerar os menores valores das distâncias de S_i relativas aos outros substituintes, obtendo-se um valor para cada substituinte, e dentre estes, escolher o maior valor como sendo a distância admissível.

CONSTRUÇÃO DE ÁRVORES DE AGRUPAMENTOS

A partir da matriz de distâncias, os valores de distância de um dado substituinte com relação aos demais são ordenados de forma crescente. Os substituintes que possuem distância menor ou igual à distância mínima admissível são considerados vizinhos. Assim, podem ser considerados vizinhos aqueles substituintes i e i' para os quais d (i,i') $\leq d_{mín}$. Denota-se por Γ_i a vizinhança do substituinte químico i. Estas vizinhanças são representadas sob a seguinte forma:

$$\Gamma_1 = \{1, v_k \quad k=2,\ldots,n_1\}$$
$$\Gamma_2 = \{2, v_k \quad k=2,\ldots,n_2\}$$
$$\vdots$$
$$\Gamma_n = \{n, v_k \quad k=2,\ldots,n_n\}$$

sendo n_i o número de vizinhos do i-ésimo substituinte. Os vizinhos v_i em uma lista Γ_i estão dispostos em ordem crescente de proximidade ao substituinte de rótulo i. Naturalmente, v_1 (o primeiro vizinho) é o próprio substituinte, isto porque o seu vizinho mais próximo é ele próprio.

Após o cálculo das vizinhanças, um dendrograma é construído. Esta árvore de agrupamentos possui forma gráfica, onde o eixo horizontal representa os substituintes e o eixo vertical representa o nível de redução da árvore. Chama-se uma redução ao conjunto de agrupamentos de elementos de uma vizinhança. A cada etapa, o número de linhas da matriz é reduzido através do agrupamento de um ou mais substituintes. Para cada sítio de substituição é construída uma árvore na qual os agrupamentos representam os ramos.

A construção da árvore de agrupamentos segue a seguinte metodologia: i) Após obtenção de todas as vizinhanças correspondentes aos substituintes de um determinado sítio de substituição, observa-se quantos e quais substituintes agruparam em uma mesma vizinhança (redução); ii) A medida que se obtém um novo agrupamento, os integrantes deste grupo são deslocados para a posição de menor índice; iii) Terminada a redução corrente, cada um dos grupos de *pseudo*substituintes, juntamente com os substituintes não agrupados, formam os ramos neste nível de agregação. Estes ramos são descritos a uma profundidade a partir da qual inicia-se a redução seguinte. Reitera-se este procedimento até que sejam obtidos apenas dois ramos, que correspondem ao topo da árvore de agrupamentos; iv) No final do processo é estabelecido um caminho em cada árvore que conduz aos substituintes potencialmente mais ativos a serem sintetizados.

UM ALGORITMO DE CLASSIFICAÇÃO HIERÁRQUICA

O algoritmo de classificação proposto é aglomerativo e considera agrupamentos sucessivos de substituintes em uma série química. Ele está fundamentado na Classificação Hierárquica e é implementado pelo programa denominado NEWCLUS, escrito em linguagem Turbo Pascal 7.0. Além de fornecer os resultados necessários à construção da árvore, o programa é interativo (conversacional), sendo apresentado de forma didática. Ele exibe uma janela de ajuda, que possibilita ao usuário o acesso às informações básicas sobre o programa, além de informações emitidas durante sua execução.

Este algoritmo realiza a redução total da árvore, agrupando os substituintes mais próximos, formando *pseudo*substituintes. Cada *pseudo*substituinte representa o conjunto de substituintes agrupados e tem como valores para os K parâmetros físico-químicos, a média dos valores dos parâmetros dos substituinte do grupo. Esta operação é repetida até que seja obtido um único grupo. Após efetuar todos os agrupamentos, o algoritmo produz o dendrograma.

ETAPAS DO ALGORITMO:

P1 - Dimensionar a entrada de dados e exibir o número limite de i e j , dada uma matriz de dados [X (i , j)];
P2 - Entrar os dados em formato matricial, via teclado ou através de arquivo externo;
P3 - Armazenar a matriz de dados em arquivo externo, utilizando o formato ASCII;
P4 - Obter os valores normalizados da matriz original de dados, que contém os valores dos parâmetros físico-químicos para cada substituinte. Os parâmetros utilizados são os mesmos da equação de Q. S. A. R. escolhida na literatura.
P5 - Obter a matriz de distâncias, calculando as distâncias relativas entre todos os pares de substituintes.
P6 - Obter as vizinhanças, classificando em ordem decrescente de proximidade todos os substituintes.
P7 - Armazenar o conjunto de vizinhos em arquivo externo (formato ASCII);
P8 - Realizar agrupamentos ordenados, calculando as médias entre os vizinhos absolutamente próximos.
P9 - Continuar agrupando os substituintes absolutamente próximos até que se encerre a redução (conjunto de agrupamentos de substituintes ou *pseudo*substituintes).
P10 - Armazenar em arquivo externo a matriz de dados resultante dos agrupamentos de uma redução.
P11 - Iniciar nova redução, executando os passos de 4 a 8;
P12 - Finalizar ao ser obtida uma matriz bidimensional de dados.
P13 - Construir e exibir na tela o dendrograma do processo de classificação hierárquica dos substituintes.

APLICAÇÃO DO ALGORITMO A UMA SÉRIE DE INDADIONAS

O algoritmo de classificação foi aplicado a uma série de derivados de 2-aril-1,3-indanodionas, com atividade anti-inflamatória. As tabelas 1 e 2 apresentam os valores dos parâmetros físico-químicos e a contribuição à atividade biológica dos substituintes nas posições *orto* e *meta/para*[17].

Tabela 1. Dados dos substituintes em *orto*.

N	Substituintes	π_o	E^o_s	σ	Log $1/C_{50}$
1	CH_3	0,56	1,24	0,29	3,46
2	CH_2CH_3	1,02	1,17	0,41	3,65
3	*i*-propil	1,53	0,77	0,56	3,92
4	*t*-butil	1,98	-0,30	0,69	3,66
5	CF_3	0,88	0,08	1,61	3,08
6	F	0,14	2,02	0,93	3,25
7	Cl	0,71	1,51	1,28	3,28
8	Br	0,86	1,32	1,35	3,31

π_o, E^o_s e σ são os parâmetros hidrofóbicos de Hansch para substituintes em *orto*, o parâmetro estérico de Taft e parâmetro eletrônico de Hammett, respectivamente, e Log$1/C_{50}$ representa a atividade biológica.

Tabela 2. Dados dos substituintes em *meta*.

N	Substituintes em *meta*	π_{mp}	σ	Log $1/C_{50}$
1	CH_3	0,56	-0,07	4,01
2	CH_2 CH_3	1,02	-0,07	4,42
3	*i*-propil	1,53	-0,05	4,69
4	*t*-butil	1,98	-0,10	4,95
5	OCH_3	-0,02	0,12	3,62
6	CF_3	0,88	0,43	3,99
7	Cl	0,71	0,37	4,02

π_{mp} é o parâmetro hidrofóbico de Hansch para substituintes nas posições *meta* ou *para*.

A tabela 3 ilustra os resultados dos agrupamentos dos substituintes nas posições *orto* e *meta/para*, obedecendo a ordem das tabelas 1 e 2 quanto a numeração. Pode ser observado que para cada redução (conjunto de agrupamentos), tem-se uma distância admissível correspondente.

As matrizes de distância dos substituintes em *orto* e *meta* são mostradas nas tabelas 4 e 5, respectivamente.

O cálculo da distância admissível é ilustrado considerando a primeira matriz, que corresponde ao primeiro grupo de vizinhanças em *orto* (Tabela 3). Os menores valores não nulos de distâncias de cada uma das linhas são: 0,8486; 0,8486;

Tabela 3. Distâncias admissíveis dos agrupamentos em *orto e meta.*

Sítio de substituição	Grupos	Nº de grupos	Distância Admissível
orto	(1,2) (3) (4) (5) (6) (7,8)	6	du = 1,7131
	(1, 2, 3) (4) (5) (6, 7, 8)	4	du = 1,6703
	(1, 2, 3, 6, 7, 8) (4) (5)	3	du = 2,1424
	(1, 2, 3, 6, 7, 8, 5) (4)	2	du = 2,5147
meta	(1, 2) (3, 4) (5) (6, 7)	4	du = 1,2275
	(1, 2, 3, 4) (5) (6, 7)	3	du = 1,6825
	(1, 2, 3, 4) (5, 6, 7)	2	du = 2,1186

1,0819; 1,6315; **1,7131**; 1,4045; 0,3902; 0,3902. Dentre estes, a maior distância é considerada como sendo a distância admissível deste nível de agrupamento. Nesta etapa, os substituintes que possuem entre eles uma distância inferior a "1,7131", formam um grupo. Os menores valores de cada linha estão em **negrito** e o maior dentre eles em **negrito sublinhado.**

As vizinhanças são: Γ_1=(1,2) Γ_2=(2,1,3) Γ_3=(3,2,4) Γ_4=(4,3) Γ_5=(5,8) Γ_6=(6,7) Γ_7=(7,8,6) e Γ_8=(8,7,5). No primeiro nível de agrupamento, são agrupados os substituintes {1,2} e {7,8}, gerando dois *pseudo*substituintes. Estes últimos, juntamente com os substituintes não agrupados {3}, {4}, {5} e {6}, constituem a primeira redução.

Uma nova matriz de dados [X(i,j)] é obtida pela redução da matriz anterior considerando os parâmetros médios dos substituintes que foram agrupados (eq.2.1, calculada somente entre os elementos do grupo). Por exemplo, o *pseudo*substituinte {1,2} apresenta parâmetros π_0=0,79; E_s^0=1,205 e σ=0.35 (vide tabela 1). A distância admissível para o próximo nível de agrupamentos (**1,6703**) é recalculada a partir da matriz de distâncias entre os "novos" substituintes, incluindo os *pseudo*-substituintes gerados. Neste nível, as vizinhanças são Γ_1=(1,2) Γ_2=(2,1,3) Γ_3=(3,2) Γ_4=(4,6) Γ_5=(5,6) e Γ_6=(6,5,4), gerando os *pseudo*substituintes {1,2} e {5,6}. As vizinhanças da próxima etapa são obtidas a partir da 3ª matriz de distâncias como sendo Γ_1=(1,4,2) Γ_2=(2,1) Γ_3=(3,4) Γ_4=(4,1,3). Os substituintes agrupados são apenas {1,4}, gerando uma nova redução. A última iteração resulta em vizinhanças Γ_1=(1,3) Γ_2=(2,3) e Γ_3=(3,1,2), determinando o agrupamento {1,3}. Todo o processo é resumido em um dendrograma por sítio de substituição.

A figura 1 apresenta as árvores de agrupamentos para a aplicação da série de 2-aril-1,3-indadionas: (a) substituição na posição *orto* e (b) substituição em *meta*. Os níveis de agregação são indicados no eixo vertical, mostrando os *pseudo*substituintes gerados em cada passo da iteração. A atividade dos penúltimos *pseudo*substituintes formados na árvore é calculada substituindo-se os valores dos parâmetros dos *pseudo*substituintes na equação de Q.S.A.R. Os ramos da árvore que correspondem ao *pseudo*composto mais ativo são selecionados. No próximo nível de agregação, repetem-se os cálculos apenas com os *pseudo*substituintes sobreviventes. Este procedimento é iterado até que se obtenha o substituinte que resulta em maior atividade (prevista) na série. No final deste processo, é estabelecido um caminho em cada árvore que conduz aos substituintes mais ativos. O caminho em negrito no dendrograma representa a seleção do substituinte que produz a maior atividade biológica quando substituído na molécula protótipo.

DISCUSSÃO DOS RESULTADOS

O algoritmo para planejamento de fármacos via classificação hierárquica proposto neste trabalho foi aplicado à diversas

Tabela 4. Matrizes de distâncias dos substituintes em *orto.*

	1	2	3	4	5	6	7	8
1	0,000	**0,8486**	1,8927	3,3098	3,1845	1,8312	2,1027	2,2657
2	**0,8486**	0,000	1,0819	2,6230	2,8831	2,1874	1,9386	1,9820
3	1,8927	**1,0819**	0,000	1,6315	2,6211	3,0342	2,2893	2,1423
4	3,3098	2,6230	**1,6315**	0,000	2,7592	4,4598	3,4759	3,1998
5	3,1845	2,8831	2,6211	2,7592	0,000	3,1846	2,0179	**1,7131**
6	1,8312	2,1874	3,0342	4,4598	3,1846	0,000	**1,4045**	1,7872
7	2,1027	1,9386	2,2893	3,4759	2,0179	1,4045	0,000	**0,3902**
8	2,2657	1,9820	2,1423	3,1998	1,7131	1,7872	**0,3902**	0,000

	1	2	3	4	5	6
1	0,000	**1,327**	2,634	2,9468	1,8385	2,0365
2	**1,327**	0,000	1,4467	2,5478	2,7072	2,0955
3	2,634	**1,4467**	0,000	2,6101	3,9475	3,0132
4	2,9468	2,5478	2,6101	0,000	2,8953	**1,6703**
5	1,8385	2,7072	3,9475	2,8953	0,000	**1,4625**
6	2,0365	2,0955	3,0132	1,6703	**1,4625**	0,000

	1	2	3	4
1	0,000	1,9645	2,5184	**1,8847**
2	**1,9645**	0,000	2,5285	3,3560
3	2,5184	2,5285	0,000	**2,1424**
4	**1,8847**	3,3560	2,1424	0,000

	1	2	3
1	0,000	2,6191	**2,1947**
2	2,6191	0,000	**2,5147**
3	**2,1947**	2,5147	0,000

Tabela 5. Matrizes de distâncias dos substituintes em *meta*.

	1	2	3	4	5	6	7
1	0,000	**0,7045**	1,4883	2,1790	1,2275	2,2826	1,9753
2	**0,7045**	0,000	0,7862	1,4764	1,8042	2,2397	2,0185
3	1,4883	0,7862	0,000	**0,7244**	2,4921	2,3604	2,2548
4	2,1790	1,4764	**0,7244**	0,000	3,2165	2,9022	2,8592
5	**1,2275**	1,8042	2,4921	0,2165	0,000	1,9521	1,5788
6	2,2826	2,2397	2,3604	2,9022	1,9521	0,000	**0,3733**
7	1,9753	2,0185	2,2548	2,8592	1,5788	**0,3733**	0,000

	1	2	3	4
1	0,000	**1,3295**	1,4029	2,1035
2	**1,3295**	0,000	2,5962	2,5037
3	**1,4029**	2,5962	0,000	1,6825
4	2,1035	2,5037	**1,6825**	0,000

	1	2	3
1	0,000	2,1372	**2,1186**
2	2,1372	0,000	**1,7158**
3	2,1186	**1,7158**	0,000

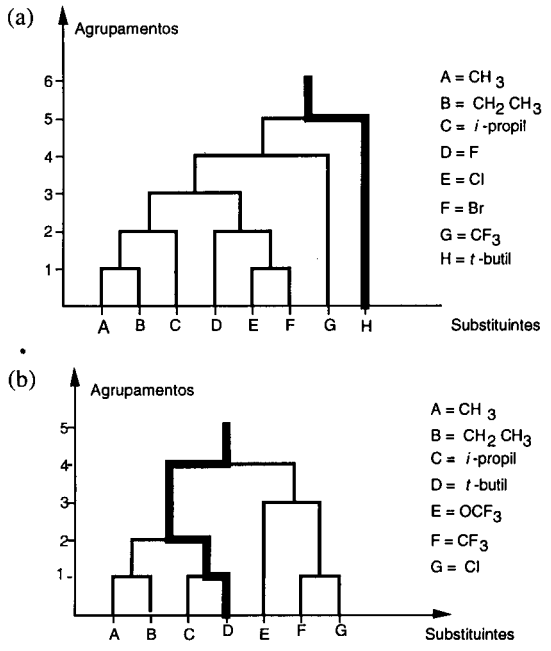

Figura 1. *Árvore de agrupamentos na posição (a)* orto *e (b)* meta.

séries químicas: 2-aril-1,3-indanodionas de ação anti-inflamatória[17], derivados do ácido propiônico 2,4-(tiazol-2-il)fenil inibidores da cicloxigenase[18], fenetilaminas[19], ácidos *trans*-3-benzoilacrílicos[20] e trifluorometanos-sulfonanilidas[21], N^2-fenilguaninas[22] antivirais e guanidinotiazolcarboxamidas[23] anticancerígenos. Os resultados obtidos com a aplicação do algoritmo a todas as séries acima mencionadas foram coerentes com os resultados experimentais da literatura[24].

O composto potencialmente mais ativo (substituição indicada na Figura 1) corresponde ao derivado 3,5-di-*t*-butil, cuja atividade prevista é $pC_{50} = \log 1/C_{50} = 5.85$. Ele apresenta atividade biológica de destaque em relação aos outros compostos da série em questão[17]. Obviamente, existem riscos de extrapolação da validade da equação de Q. S. A. R. Nestes casos, os resultados experimentais podem diferir substancialmente daqueles calculados. O exemplo ilustrado neste trabalho é um pouco acadêmico por apresentar um número reduzido de substituintes por sítio. Todavia, o potencial e atratividade deste procedimento torna-se explícito em séries químicas que apresentam um elevado número de sítios de substituição, envolvendo centenas de substituintes possíveis por sítio, já que serão sintetizados poucos compostos envolvendo cada agrupamento de substituintes.

O algoritmo proposto orienta o pesquisador na escolha de derivados originais a serem sintetizados. No caso de uma série química previamente sintetizada onde novos compostos devam ser obtidos, a equação de Q. S. A. R. pode ser utilizada para predizer a atividade biológica dos *pseudo*compostos obtidos usando *pseudo*substituintes. Este cálculo deve ser interpretado de forma qualitativa em relação aos valores experimentais.

Esta análise mostra que o algoritmo de classificação hierárquica desenvolvido é satisfatório na seleção de derivados a serem sintetizados. Adicionalmente, o programa vem sendo usado no ensino de pós-graduação, proporcionando aos estudantes de planejamento e síntese racional de fármacos, melhor compreensão da utilização de algoritmos como ferramentas de auxílio à pesquisa.

O NEWCLUS pode ser ainda mais prático se reescrito em linguagem Delphi®, utilizando janelas e menus. A nova proposta é modernizar o modelo utilizando-se outras métricas (e.g. entropia, conectividade) na obtenção dos agrupamentos. Os resultados ora obtidos foram derivados a partir de séries químicas estudadas. Seria interessante escolher uma série original, classificar os substituintes, determinar compostos potencialmente ativos, e verificar experimentalmente os resultados.

AGRADECIMENTOS

O primeiro autor agradece o suporte da *Coordenação de Apoio ao Pessoal de Nível Superior* - CAPES. Este trabalho também contou com o apoio parcial do *Conselho Nacional de Desenvolvimento Científico e Tecnológico* - CNPq. Os autores agradecem a colaboração dos alunos de Iniciação Científica, Adriano Antunes de Souza Araújo, Leonides Justino Júnior e Irwin Rose Alencar de Menezes, nas simulações computacionais e os comentários de revisores anônimos.

REFERÊNCIAS

1. Topliss, J. G.; *J. Med. Chem.* **1972**, *15*, 1006.
2. Farrar, D. E. and Glauber, R. R.; *Rev. Econ. Stat.* **1967**, *49*, 92.
3. Kowalski, B. R. and Bender; *J. Amer. Chem. Soc.* **1972**, *94*, 5632.
4. Kowalski, B. R. and Bender; *J. Amer. Chem. Soc.* **1973**, *95*, 686.

5. Giacca, M.; Menzo, S.; Trojan, S.; Monti-Bragadin, C.; *Eur. J. Epidemiol.* **1987**, *3*, 155.
6. Bouroche, J. M., Saporta, G.; *L'analyse des données*; Presses Universitaires de France; Paris; France, 1980.
7. Diday, E.; Simon, J. C.; *Clustering analysis, In: Communication and Cybernetics 10-digital pattern recognition* ; Heidelberg:Spring-Verlag, Ed.; New York, 1976; p 47.
8. Culasso F.; Lenzi, A.; Favilil, S.; Dondero, F.; *Arch. Androl* **1991**, *26*, 163.
9. Allen, M. T.; Boquet, A. J.; Shelley, K. S.; *Psychosom. Med.* **1991**, *53*, 272.
10. Cayuela, D. A.; Lacalle, R. Jr.; Gili, M.; *Gac. Sanit.* **1990**, *4*, 227.
11. Pekas, J. C.; Wray, J. E.; *J. Nutric.* **1991**, *121*, 231.
12. Maes, M.; Cosyns, P.; Maes, L.; D'Hondt, P.; Sinotte, C.; *Psychiatriy Res.* **1990**, *34*, 29.
13. Craig, P. N.; *J. Med. Chem.* **1971**, *14*, 680.
14. Wooton, R.; Cranfield, R.; Sheppey, G. C.; Goodford, P. J.; *J. Med. Chem.* **1975**, *18*, 607.
15. Hansch, C.; Leo, A.; *Substituent constants for correlation analysis in Chemistry and Biology*; John Wiley, Ed.; New York, 1980.
16. Cavalcanti, S. C. H. *Dissertação de Mestrado* - Universidade Federal de Pernambuco - Brasil. 1994.
17. Berg, G. V. D.; Bultsma, T.; Rekker, R. F.; Tomas, W. *Eur. J. Med. Chem.* - Chimica Therapeutica. **1975**, *3*, 242.
18. Naito, Y.; Yamaura, Y.; Inoue, Y.; Fukaya, C.; Yokoyama, K.; Nakagawa, Y.; Fujita, T. *Eur. J. Med. Chem.* **1992**, *27*, 645.
19. Unger, S. H.; Hansch, C.; *J. Med. Chem.* **1973**, *16*, 745.
20. Bowden, K.; Henry, M. P.; *Structure activity relations. II. Antibacterial activity of trans-3-benzoylacrylic acids and esters, In: Biological correlations-the Hansch approach.* American Chemical Society; Washington, 1972, 130.
21. Yapel, A. F. Jr.; *Structure-activity correlations for meta and para substituted trifluoromethane sulfonanilide pre-emergence herbicides, In: Biological correlations-the Hansch approach*; American Chemical Society; Washington, 1972, 183.
22. Gambino, J.; Focher, F.; Hildebrand, C.; Maga, G.; Noonan, T.; Spadari, S.; and Wright, G.; *J. Med. Chem.* **1992**, *35*, 2979.
23. Schnur, R.C.; Gallaschun, R. J.; H Singleton, D.; Grissom, M.; Sloan, D. E.; Goodwin, P.; Mc Niff, P. A.; Fliri, A. F. J.; Mangano, F. M.; Olson, T. H. and Pollack, V. A. *J. Med. Chem.* **1991**, *34*, 1975.
24. Moreno, M. P. N. Dissertação de Mestrado. Universidade Federal de Pernambuco. Brasil. 1995.