Quim. Nova, Vol. 25, No. 3, 439-448, 2002.

Divulgação

SELEÇÃO DE VARIÁVEIS EM QSAR

Márcia Miguel Castro Ferreira
Instituto de Química, Universidade Estadual de Campinas, CP 6154, 13083-970 Campinas - SP
Carlos Alberto Montanari
Departamento de Química, Universidade Federal de Minas Gerais, Campus da Pampulha, 31270-901 Belo Horizonte - MG
Anderson Coser Gaudio*
Departamento de Física, Centro de Ciências Exatas, Universidade Federal do Espírito Santo, Campus de Goiabeiras, 29060-900 Vitória - ES

Recebido em 11/12/00; aceito em 4/2/02

VARIABLE SELECTION IN QSAR. The process of building mathematical models in quantitative structure-activity relationship (QSAR) studies is generally limited by the size of the dataset used to select variables from. For huge datasets, the task of selecting a given number of variables that produces the best linear model can be enormous, if not unfeasible. In this case, some methods can be used to separate good parameter combinations from the bad ones. In this paper three methodologies are analyzed: systematic search, genetic algorithm and chemometric methods. These methods have been exposed and discussed through practical examples.

Keywords: systematic search; genetic algorithm; chemometric methods.

INTRODUÇÃO

As pesquisas na área de QSAR (*Quantitative Structure-Activity Relationships*) têm como principal objetivo a construção de modelos matemáticos que relacionem a estrutura química e a atividade biológica de uma série de compostos análogos. Em geral, esses compostos diferem entre si pela presença de um ou mais grupos substituintes em posições definidas da estrutura química comum da série[1-4]. A construção dos modelos requer a elaboração de conjunto ou matriz de dados contendo a medida quantitativa da atividade biológica e os parâmetros físico-químicos e estruturais capazes de descrever as propriedades dos compostos. Em resumo, o conjunto de dados contém os valores da atividade biológica *Y* e das *m* variáveis descritivas *X* referentes aos *n* compostos (Quadro 1). O conjunto de dados é a matéria prima para a construção dos modelos matemáticos, que em geral são lineares e multidimensionais, representados genericamente pela eq 1.

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_k X_k \quad (1)$$

Nessa equação, $\hat{Y}$ representa os valores previstos da resposta biológica; X_1, X_2, …, X_k são as propriedades de caráter lipofílico, eletrônico, estéreo e polar[3]; e b_0, b_1, …, b_k são coeficientes de ajuste. Segundo o método de Hansch-Fujita[1-4], esses coeficientes são obtidos através de regressão linear múltipla (RLM)[5-7]. A qualidade do ajuste do modelo aos valores observados da atividade biológica pode ser avaliada através do cálculo do coeficiente de correlação (*R*), do desvio-padrão (*s*) e do teste de Fischer (*F*). Em termos simplificados, um modelo bem ajustado deverá apresentar valor de *R* próximo à unidade, *s* pequeno e *F* grande.

Deve-se notar que, apesar do conjunto de dados conter um total de *m* variáveis, apenas um subconjunto *k* será utilizado na construção de cada modelo. Existe limite para o valor de *k*, no caso de equações de regressões lineares, para que a mesma tenha solução única[8].

Quadro 1. Representação genérica do conjunto de dados utilizado em QSAR, que contém os valores da atividade biológica *Y* e os valores das *m* variáveis descritivas *X* referentes aos *n* compostos da série

Y	X_1	X_2	…	X_m
Y_1	$X_{1,1}$	$X_{1,2}$	…	$X_{1,m}$
Y_2	$X_{2,1}$	$X_{2,2}$	…	$X_{2,m}$
…	…	…	…	…
Y_n	$X_{n,1}$	$X_{n,2}$	…	$X_{n,m}$

Do ponto de vista matemático, o valor máximo de *k* é igual a *n* - 1. Assim, um modelo linear que inclui dezesseis compostos ($n = 16$) pode acomodar no máximo quinze variáveis ($k = 15$). Porém, à medida que *k* se aproxima de *n* ocorre *overfitting*, que pode ser traduzido como *ajuste forçado*. O *overfitting* consiste na obtenção de valor elevado do coeficiente de correlação decorrente do número excessivo de variáveis incluídas no modelo e não de seu ajuste natural aos valores observados da atividade. A utilização de *k* elevado limita o número de graus de liberdade que o modelo oferece para os desvios entre *Y* e $\hat{Y}$, de forma que esses desvios vão sendo artificialmente reduzidos. Isso produz a ilusão de que o modelo está bem ajustado[5-7]. Em QSAR, convencionou-se que, para reduzir a possibilidade de *overfitting*, cada grupo de cinco ou seis compostos incluídos no modelo permite a acomodação de uma variável[3,9,10]. Portanto, se a série possui dezesseis compostos ($n = 16$) o número de variáveis que o modelo pode acomodar é três ($k = 3$).

Devido ao fato de, quase sempre, o número total de variáveis disponíveis ser muito maior do que o número que será efetivamente incluído nos modelos, ou seja $m >> k$, há necessidade de lançar-se mão de algum tipo de procedimento de seleção para a composição dos modelos de QSAR. O processo de seleção consiste em encontrar combinações de *k* variáveis, dentre as *m* disponíveis, capazes de produzir modelos matemáticos que descrevam adequadamente os valores observados da atividade biológica (eq 1), ou seja, modelos em que o somatório (dos quadrados) das diferenças entre $\hat{Y}$ e *Y* seja o mais próximo de zero possível (método dos mínimos quadrados, ou

* e-mail: anderson@npd.ufes.br; Internet: http://tau.cce.ufes.br/anderson

MMQ). Na prática, o valor de *m* pode variar muito. Em QSAR clássico[11], em que as variáveis utilizadas são linearmente relacionadas com a variação da energia livre do processo de interação fármaco-receptor, o valor de *m* pode ser pouco maior do que uma dezena. No entanto, a utilização de índices de conectividade[3], índices de similaridade[3], propriedades físicas do espaço circunvizinho à molécula do fármaco, obtidas pelo método CoMFA[12], e índices eletrônicos derivados de cálculos de química quântica[13], podem facilmente elevar o valor de *m* para além da centena. Além disso, as dificuldades na síntese de compostos e na execução de testes biológicos limitam em cerca de duas ou três dezenas, em média, o número de compostos nos conjuntos de dados utilizados na prática. Isso, por sua vez, limita o valor de $k \leq 6$. A necessidade de selecionar um subconjunto de quatro, cinco ou seis variáveis dentre um conjunto de, por exemplo, uma centena conduz inevitavelmente o pesquisador à busca de técnicas eficientes de seleção.

Este trabalho tem como objetivo a apresentação e discussão dos principais métodos utilizados na seleção de variáveis em QSAR. Serão abordados a busca sistemática, o algoritmo genético e os métodos quimiométricos. Os dois primeiros métodos referem-se à seleção para modelos matemáticos baseados em RLM, ou seja, para modelos baseados no método de Hansch[11]. Este não é o único método disponível para a construção de modelos em QSAR. A construção de modelos baseada na análise de componentes principais (PCA) e regressão por mínimos quadrados parciais (PLS) não utiliza o processo de seleção nos mesmos moldes. Neste, o conjunto original de dados é comprimido gerando um número pequeno de variáveis denominadas componentes principais. Estas têm a vantagem de serem independentes (ortogonais) entre si e, portanto, a correlação entre as variáveis não limita sua aplicabilidade.

SELEÇÃO DE VARIÁVEIS E A QUALIDADE DOS MODELOS DE QSAR

Antes da análise dos métodos de seleção propriamente dita, é preciso destacar um aspecto importante sobre a expressão *qualidade de um modelo matemático* em QSAR. Para que uma equação de regressão seja promovida a modelo matemático é preciso muito mais do que simplesmente possuir elevado coeficiente de correlação. Para validar-se estatisticamente uma equação de regressão, é preciso executar diversos testes de validação, tais como o cálculo do coeficiente de correlação, do desvio-padrão, do teste de Fischer e do nível geral de confiabilidade do modelo (p-valor)[5-7]. Também se deve realizar o teste *F* seqüencial, a validação dos coeficientes da equação através do cálculo de seus limites de confiabilidade de 95%, a análise dos resíduos e a validação cruzada. Bons resultados que eventualmente venham a ser obtidos em todos esses testes de forma alguma garantem que a equação venha a ser útil para descrever a atividade biológica de um grupo de compostos. É preciso que a equação seja consistente com algum mecanismo de ação, em nível molecular, proposto para os compostos e que também sirva para fazer previsões sobre a atividade de compostos que não tenham sido incluídos no modelo[3,4,9]. O leitor interessado em maiores detalhes acerca da elaboração e avaliação de modelos de QSAR é encorajado a consultar a referência 14.

Neste trabalho, não haverá preocupação com os aspectos relacionados ao mecanismo de ação dos compostos envolvidos na construção dos modelos em QSAR. Os métodos de seleção de variáveis que serão descritos, são utilizados apenas para a triagem de equações que possuem alguma chance de tornarem-se verdadeiros modelos. Esses métodos não são mágicos. Eles apenas eliminam as equações inviáveis e deixam ao nosso critério as análises posteriores que poderão ou não transformar uma equação em modelo.

BUSCA SISTEMÁTICA

A busca sistemática é sem dúvida o método mais seguro de seleção que pode ser utilizado na construção de modelos matemáticos baseados em RLM. A busca sistemática consiste em combinar as *m* variáveis disponíveis de forma a construir e analisar todos as possíveis equações de regressão com *k* variáveis e, a partir daí, selecionar as melhores. Por exemplo, desejando-se saber qual a melhor regressão que pode ser obtida com $k = 3$, $\hat{Y} = b_0 + b_i X_i + b_j X_j + b_l X_l$, ou simplesmente $\hat{Y} = f(X_i, X_j, X_l)$, sendo que $m = 15$, X_1, X_2, …, X_{15}, devem-se construir equações utilizando-se todas as possíveis combinações não repetidas das quinze variáveis, três a três. Ou seja, $\hat{Y} = f(X_1, X_2, X_3)$, $\hat{Y} = f(X_1, X_2, X_4)$, …, $\hat{Y} = f(X_1, X_2, X_{15})$, $\hat{Y} = f(X_2, X_3, X_4)$, $\hat{Y} = f(X_2, X_3, X_5)$, …, $\hat{Y} = f(X_2, X_3, X_{15})$, …, $\hat{Y} = f(X_{13}, X_{14}, X_{15})$.

Este é o único método de seleção que pode assegurar que a melhor combinação será encontrada. Considerando-se a necessidade de executar regressões com todas as possíveis combinações de variáveis, há que se perguntar sobre a eficiência do método em termos da duração do tempo de busca. Considere-se o seguinte exemplo.

Seja a matriz de Selwood[15], que apresenta $n = 31$ e $m = 53$. Sabendo-se que o número de possíveis equações de regressão com *k* variáveis é dado pela fórmula $m!/[k!(m-k)!]$, o número total de equações distintas que podem ser construídas com $k \leq 6$ é de cerca de 26 milhões. Um microcomputador capaz de analisar mil regressões por segundo levaria aproximadamente sete horas para completar a tarefa, o que é bastante viável. No entanto, o total de equações que podem ser construídas com $k \leq 10$ é de cerca de 25 bilhões. O mesmo microcomputador levaria aproximadamente 285 dias para concluir a tarefa. Neste caso, o tempo de cálculo acaba por inviabilizar a busca.

O tempo de cálculo varia com a complexidade do modelo. Por exemplo, o tempo requerido para analisar um milhão de combinações de *m* variáveis, quatro a quatro, é maior do que o tempo requerido para analisar um milhão de combinações de *m*, três a três. Considerando-se a matriz de Selwood, um microcomputador Pentium II 233 MHz, com 128 MB RAM, é capaz de analisar cerca de 4500 combinações de 53 variáveis, três a três, em um segundo.

Pelo fato da busca sistemática ser tão onerosa do ponto de vista computacional, devem-se encontrar meios para acelerar sua execução. A experiência mostra que há pelo menos duas maneiras eficientes de acelerar a busca sistemática. A primeira consiste em excluir da busca as regressões que contenham variáveis muito correlacionadas entre si. A rigor, os modelos baseados em RLM deveriam incluir apenas variáveis independentes. Na prática, no entanto, observa-se que há sempre algum grau de correlação entre as mesmas. Ainda em termos de RLM, quando ocorre elevado grau de correlação é porque, teoricamente, os termos envolvidos estão descrevendo a mesma propriedade. Como só há necessidade de uma variável por regressão que descreva dada propriedade, a presença de duas variáveis correlacionadas na mesma equação de RLM é, não somente desnecessária, como também prejudicial. Isso ocorre basicamente por dois motivos. O primeiro e mais importante é devido à dificuldade na interpretação das implicações farmacológicas e bioquímicas do conteúdo do modelo. O segundo motivo é técnico. A solução matemática do MMQ envolve inversão matricial. Esta operação implica na perda de precisão aritmética quando a matriz possui duas colunas correlacionadas (determinante próximo de zero).

Para verificar se X_i e X_j são correlacionadas, basta calcular o coeficiente de correlação entre ambas, R_{ij}, que pode ser feito rapidamente através da eq 2. Por exemplo, antes de construir o modelo $\hat{Y} = f(X_2, X_3, X_5)$, devem-se verificar os valores de R_{23}, R_{25} e R_{35}. Se apenas um deles for maior que o valor limite adotado, o modelo deverá ser descartado. A decisão sobre o valor limite de correlação entre duas variáveis é do pesquisador. Em QSAR, costumam-se evitar as

combinações em que $R_{ij} \geq 0,6$[3]. Considerando-se que algumas classes de variáveis são extremamente correlacionadas, como por exemplo cargas atômicas e índices de orbitais de fronteira, a economia de cálculo pode ser significativa.

$$R_{ij} = \frac{\sum X_i \cdot X_j - \sum X_i \cdot \sum X_j / n}{\sqrt{(\sum X_i^2 - (\sum X_i)^2 / n) \cdot (\sum X_j^2 - (\sum X_j)^2 / n)}} \quad (2)$$

A segunda maneira de acelerar a busca sistemática refere-se ao método de avaliar a qualidade de cada combinação. Ao invés de aplicar o MMQ a cada uma das combinações, deve-se optar pelo cálculo em separado de apenas um dos parâmetros de ajuste, *R*, *s* ou *F*. Isso pode ser vantajoso porque a regressão linear múltipla é muito onerosa do ponto de vista computacional, pois utiliza diversas operações matriciais, incluindo uma operação de inversão. Para que essa tática seja realmente vantajosa, o cálculo de *R*, *s* ou *F* deve ser feito *por fora* do MMQ, ou seja, sem envolver operações matriciais onerosas. Embora isso seja possível para *R*, *s* e *F*, na prática utiliza-se o cálculo de *R* para avaliar a potencialidade de cada combinação antes da regressão linear propriamente dita[16]. Vale dizer que somente será processada a combinação que apresentar coeficiente de correlação acima do valor escolhido pelo pesquisador, como por exemplo $R \geq 0,9$. Sabendo-se que poucas combinações de variáveis formam modelos com esse grau de ajuste, pode-se imaginar a economia no tempo de cálculo.

O formulário para o cálculo do coeficiente de correlação é dado a seguir[16], em que são utilizados os índices *i*, *j*, *k*, *l* e *m* para representar, de forma abreviada, os termos X_i, X_j, X_k, etc. Portanto, os índices *k* e *m* que aparecem em algumas das equações que seguem nada têm a ver com aqueles definidos anteriormente.

O coeficiente de correlação para modelos com $k = 1$, $\hat{Y} = f(X_i)$, é dado pela eq 3.

$$R_{yi} = \frac{S_{yi}}{\sqrt{S_{yy} \cdot S_{ii}}} \quad (3)$$

Os valores de S_{yi}, S_{yy} e S_{ii} são dados pelas eqs 4-6.

$$S_{yi} = \sum Y \cdot X_i - \sum Y \cdot \sum X_i / n \quad (4)$$

$$S_{yy} = \sum Y^2 - (\sum Y)^2 / n \quad (5)$$

$$S_{ii} = \sum X_i^2 - (\sum X_i)^2 / n \quad (6)$$

Para modelos com $k = 2$, $\hat{Y} = f(X_i, X_j)$, o quadrado do coeficiente de correlação é dado pela eq 7.

$$R^2_{y.ij} = \frac{R^2_{yi} + R^2_{yj} - 2R_{yi} \cdot R_{yj} \cdot R_{ij}}{1 - R^2_{ij}} \quad (7)$$

Para modelos com $k = 3$, $\hat{Y} = f(X_i, X_j, X_k)$, o formulário necessário ao cálculo quadrado do coeficiente de correlação, $R_{y.ijk}$, é dado pelas eqs 8-11.

$$R^2_{y.ijk} = 1 - (1 - R^2_{yi}) \cdot (1 - R^2_{yj.i}) \cdot (1 - R^2_{yk.ij}) \quad (8)$$

$$R^2_{yk.ij} = \frac{(R_{yk.i} - R_{yj.i} \cdot R_{jk.i})^2}{(1 - R^2_{yj.i}) \cdot (1 - R^2_{jk.i})} \quad (9)$$

$$R^2_{yj.i} = \frac{(R_{yj} - R_{yi} \cdot R_{ij})^2}{(1 - R^2_{yi}) \cdot (1 - R^2_{ij})} \quad (10)$$

$$R^2_{jk.i} = \frac{(R_{jk} - R_{ij} \cdot R_{ik})^2}{(1 - R^2_{ij}) \cdot (1 - R^2_{ik})} \quad (11)$$

Até onde se pôde averiguar, o cálculo do coeficiente de correlação para modelos com $k = 4$ e $k = 5$ não foi explicitamente descrito na literatura. No entanto, a formulação para esse cálculo pode ser deduzida a partir da comparação entre R_{yi}, $R_{y.ij}$ e $R_{y.ijk}$ (eqs 3-11).

Para modelos $k = 4$, $R_{y.ijkl}$ é dado pelas eqs 12-15.

$$R^2_{y.ijkl} = 1 - (1 - R^2_{yi}) \cdot (1 - R^2_{yj.i}) \cdot (1 - R^2_{yk.ij}) \cdot (1 - R^2_{yl.ijk}) \quad (12)$$

$$R^2_{yl.ijk} = \frac{(R_{yl.ij} - R_{yk.ij} \cdot R_{kl.ij})^2}{(1 - R^2_{yk.ij}) \cdot (1 - R^2_{kl.ij})} \quad (13)$$

$$R^2_{yl.ij} = \frac{(R_{yl.i} - R_{yj.i} \cdot R_{jl.i})^2}{(1 - R^2_{yj.i}) \cdot (1 - R^2_{jl.i})} \quad (14)$$

$$R^2_{kl.ij} = \frac{(R_{kl.i} - R_{kj.i} \cdot R_{jl.i})^2}{(1 - R^2_{kj.i}) \cdot (1 - R^2_{jl.i})} \quad (15)$$

Finalmente, para modelos com $k = 5$, $R_{y.ijklm}$ é dado pelas eqs 16-19.

$$R^2_{y.ijklm} = 1 - (1 - R^2_{yi}) \cdot (1 - R^2_{yj.i}) \cdot (1 - R^2_{yk.ij}) \cdot (1 - R^2_{yl.ijk}) \cdot (1 - R^2_{ym.ijkl}) \quad (16)$$

$$R^2_{ym.ijkl} = \frac{(R_{ym.ijk} - R_{yl.ijk} \cdot R_{lm.ijk})^2}{(1 - R^2_{yl.ijk}) \cdot (1 - R^2_{lm.ijk})} \quad (17)$$

$$R^2_{ym.ijk} = \frac{(R_{ym.ij} - R_{yk.ij} \cdot R_{km.ij})^2}{(1 - R^2_{yk.ij}) \cdot (1 - R^2_{km.ij})} \quad (18)$$

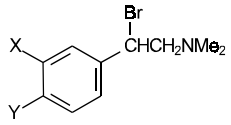
$$R^2_{lm.ijk} = \frac{(R_{lm.ij} - R_{lk.ij} \cdot R_{km.ij})^2}{(1 - R^2_{lk.ij}) \cdot (1 - R^2_{km.ij})} \quad (19)$$

Pode-se ilustrar o método de busca sistemática ao conjunto de dados da Tabela 1 que contém os valores da atividade biológica e valores atualizados[17,18] de π, π_m, σ^+, σ_m e r_v^p de vinte e dois compostos derivados da N,N-dimetil-a-bromo-feniletilamina, substituídos nas posições *meta* e *para* do anel fenila[9,19]. Na Tabela 1, a atividade biológica é representada por log $1/C$[20], em que *C* representa a concentração do fármaco, em moles/kg de peso corporal, capaz de produzir 50% de antagonismo à ação vasopressora de uma dose padrão de epinefrina em ratos.

A execução da busca sistemática ao conjunto de dados da Tabela 1 gerou 31 equações de regressão. Os valores de R, s e F dessas equações são mostradas na Tabela 2. Na construção da Tabela 2, não foi imposta qualquer restrição quanto ao grau de correlação entre as variáveis independentes.

A melhor equação com uma variável é log $1/C = f(r_v^p)$ (No. 5, Tabela 2), cuja avaliação é R = 0,878, s = 0,279 e F = 67,06. Diz-se que r_v^p é capaz de explicar cerca de 77% (R^2 x100%) da variabilidade da atividade. Devido a isto, há boas chances de r_v^p também estar presente nas melhores regressões com maior número de variáveis. Assim, a melhor equação com duas variáveis é log $1/C = f(\pi_m, r_v^p)$ (No. 12, Tabela 2), cuja avaliação é R = 0,936, s = 0,210 e F = 67,51 . Da equação No. 5 para a de No. 12 (Tabela 2) houve significativa melhora do coeficiente de correlação e do desvio-padrão. Para construir a melhor regressão com três variáveis é necessário retirar π_m e acrescentar π e σ^+ na No.12. O resultado dessa mudança é log $1/C = f(\pi, \sigma^+, r_v^p)$ (No. 20, Tabela 2), cuja avaliação é R = 0,963, s = 0,166 e F = 76,32. A comparação dos valores de R e s das equações No. 12 e 20 indica que pode ser mais vantajoso representar a atividade biológica dos compostos da série através de um modelo com três variá-

Tabela 1. Atividade biológica e variáveis descritivas dos derivados da N,N-dimetil-a-bromo-feniletilamina, substituídos nas posições *meta* e *para* do anel fenila



No	X	Y	log 1/C [a]	π [b]	π_m [c]	σ^+ [d]	σ_m [e]	r_v^p [f]
1	H	H	7,46	0,00	0,00	0,00	0,00	1,20
2	H	F	8,16	0,14	0,00	-0,07	0,00	1,47
3	H	Cl	8,68	0,71	0,00	0,11	0,00	1,75
4	H	Br	8,89	0,92	0,00	0,15	0,00	1,85
5	H	I	9,25	1,12	0,00	0,14	0,00	1,98
6	H	Me	9,30	0,58	0,00	-0,31	0,00	1,97
7	F	H	7,52	0,14	0,14	0,35	0,34	1,20
8	Cl	H	8,16	0,71	0,71	0,40	0,37	1,20
9	Br	H	8,30	0,92	0,92	0,41	0,39	1,20
10	I	H	8,40	1,12	1,12	0,36	0,35	1,20
11	Me	H	8,46	0,58	0,58	-0,07	-0,07	1,20
12	Cl	F	8,19	0,85	0,71	0,33	0,37	1,47
13	Br	F	8,57	1,06	0,92	0,34	0,39	1,47
14	Me	F	8,82	0,72	0,58	-0,14	-0,07	1,47
15	Cl	Cl	8,89	1,42	0,71	0,51	0,37	1,75
16	Br	Cl	8,92	1,63	0,92	0,52	0,39	1,75
17	Me	Cl	8,96	1,29	0,58	0,04	-0,07	1,75
18	Cl	Br	9,00	1,63	0,71	0,55	0,37	1,85
19	Br	Br	9,35	1,84	0,92	0,56	0,39	1,85
20	Me	Br	9,22	1,50	0,58	0,08	-0,07	1,85
21	Me	Me	9,30	1,16	0,58	-0,38	-0,07	1,97
22	Br	Me	9,52	1,50	0,92	0,10	0,39	1,97

[a] C representa a concentração do fármaco, em moles/kg de peso corporal, capaz de produzir 50% de antagonismo à ação vasopressora de uma dose padrão de epinefrina em ratos; [b] π é a constante lipofílica de Hansch-Fujita; [c] π_m é a constante lipofílica dos grupos químicos presentes na posição *meta* do anel fenila; [d] σ^+ é a constante eletrônica de Hammett de substituintes capazes de deslocalizar uma carga eletrônica residual positiva; [e] σ_m é a constante eletrônica dos grupos ligados à posição *meta* e; [f] r_v^p é o raio de van der Waals do substituinte na posição *para*.

veis do que com duas. O mesmo não pode ser dito ao considerarem-se as melhores regressões com quatro e cinco variáveis. Os resultados da Tabela 2 mostram que as regressões com mais de três variáveis não são capazes de melhorar a explicação da atividade biológica em relação à equação No. 20 da Tabela 2. Dessa forma, o resultado da busca sistemática indica que a atividade dos compostos da série poderá ser representada por uma equação de três variáveis. Isso não quer dizer que esta seja a de No. 20, pois há outras com três variáveis que possuem avaliações equivalentes, como por exemplo as de No. 16, 23 e 24. Avaliações mais aprofundadas deverão ser executadas sobre essas equações para decidir-se qual é a de melhor qualidade estatística.

É importante verificar o grau de correlação entre as variáveis durante o processo de seleção de equações obtidas através de RLM. Isso é verificado através da construção da matriz de correlação. A Tabela 3 mostra as correlações relativas aos dados da Tabela 1. Pode-se observar que apenas σ^+ e σ_m não devem ser combinadas numa mesma regressão, pois apresentam coeficiente de correlação igual a 0,702.

ALGORITMO GENÉTICO

A evolução da vida na Terra é baseada em três processos fundamentais: *cruzamento*, para perpetuar e aprimorar a qualidade dos indivíduos de uma mesma espécie; *seleção natural*, em que apenas as espécies mais adaptadas ao meio conseguem sobreviver e; *mutação genética*, que é o método que a natureza utiliza para produzir novas espécies. A ação dos três processos evolutivos durante quase um bilhão de anos sobre umas poucas espécies ancestrais resultou na imensa variedade de plantas e animais que hoje conhecemos. Estima-se que para cada espécie que sobrevive ao processo de evolução, cerca de mil outras são eliminadas. Atualmente estão catalogadas aproximadamente dois milhões de espécies de plantas e animais e existem estimativas de que outras 10-30 milhões ainda estejam por ser descobertas. Esses números nos permitem imaginar que o total de espécies que tenham sido extintas ao longo da evolução na Terra seja da ordem de dezena de bilhões. Quando comparada ao tempo de vida médio de um ser humano, pode-se concluir que a evolução é um processo muito lento. Do ponto de vista da razão entre o número de espécies que sobrevivem à evolução e o número total de espécies geradas (cerca de 1/1000), a conclusão obrigatória é que a evolução é um processo pouco eficiente. No entanto, olhando ao nosso redor, percebe-se que, mesmo sendo lenta e pouco eficiente, a evolução foi capaz de produzir resultados espetaculares.

Apesar de ter sido observada apenas em nosso planeta, deve-se notar que a evolução é um processo de otimização, ou seja, espécies altamente evoluídas são desenvolvidas a partir de espécies primitivas e de pouca complexidade estrutural. Na década de 1960, John Holland[21] propôs que o processo de evolução da vida poderia ser utilizado como base para a criação de uma metodologia geral de busca e otimização. Holland imaginou que a lentidão e a pouca efici-

Tabela 2. Resultado da seleção de variáveis através de busca sistemática executada sobre o conjunto de dados mostrado na Tabela 1. Modelos com diferentes números de variáveis estão separados pelas linhas horizontais

No.	π	π_m	σ^+	σ_m	r_v^p	R [a]	s [b]	F [c]
1	•					0,760	0,379	27,29
2		•				0,206	0,570	0,88
3			•			0,152	0,575	0,48
4				•		0,134	0,577	0,37
5					•	**0,878**	**0,279**	**67,06**
6	•	•				0,844	0,320	23,57
7	•		•			0,932	0,217	62,54
8	•			•		0,874	0,290	30,82
9	•				•	0,924	0,228	55,87
10		•	•			0,364	0,556	1,45
11		•		•		0,406	0,546	1,88
12		•			•	**0,936**	**0,210**	**67,51**
13			•	•		0,153	0,590	0,23
14			•		•	0,878	0,286	31,88
15				•	•	0,879	0,285	32,29
16	•	•	•			0,953	0,186	59,15
17	•	•		•		0,886	0,285	21,86
18	•	•			•	0,936	0,215	42,76
19	•		•	•		0,932	0,223	39,53
20	•		•		•	**0,963**	**0,166**	**76,32**
21	•			•	•	0,939	0,211	44,69
22		•	•	•		0,409	0,560	1,21
23		•	•		•	0,953	0,187	58,77
24		•		•	•	0,959	0,174	68,56
25			•	•	•	0,881	0,291	20,73
26	•	•	•	•		**0,964**	**0,167**	**56,55**
27	•	•	•		•	0,963	0,170	54,53
28	•	•		•	•	0,959	0,179	48,57
29	•		•	•	•	0,964	0,169	55,33
30		•	•	•	•	0,959	0,178	48,97
31	•	•	•	•	•	**0,964**	**0,172**	**42,65**

[a] *R* é o coeficiente de correlação, [b] *s* é o desvio-padrão e [c] *F* é teste de Fischer do modelo de regressão.

Tabela 3. Matriz de correlação das variáveis independentes da Tabela 1. As únicas variáveis que apresentam alto grau de correlação são σ^+ e σ_m

	π	π_m	σ^+	σ_m	r_v^p
π	1,000	0,413	0,192	0,127	0,361
π_m		1,000	0,262	0,419	0,018
σ^+			1,000	**0,702**	0,035
σ_m				1,000	0,043
R_v^p					1,000

ência do processo evolutivo como método matemático poderiam ser superadas através da utilização do computador para acelerá-la. O método foi denominado *algoritmo genético* (AG) ou *algoritmo evolucionário*[21-23]. As primeiras aplicações dos AGs foram na otimização de funções matemáticas, como por exemplo na busca de valores de *x* que maximizam ou minimizam uma função do tipo $y = f(x)$[21]. O sucesso das tentativas iniciais logo incentivaram novas aplicações, como simulações do jogo de xadrez e da evolução de populações de bactérias[22].

O esquema geral de execução dos algoritmos genéticos é mostrado na Figura 1. Embora possam ser encontradas muitas variações na estrutura dos algoritmos genéticos, podem-se identificar características fundamentais que são comuns à sua maioria. Dentre estas, a mais importante é a presença das três etapas evolutivas: seleção natural, cruzamento e mutação. O processo denominado *elitismo* também é freqüentemente encontrado em algoritmos genéticos. Este consiste em, uma vez executada dada etapa evolutiva, evitar que indivíduos menos evoluídos que os anteriores tomem seus lugares na próxima geração. Para decidir se os indivíduos da geração seguinte são mais evoluídos que os da geração anterior, é necessário executar a *avaliação*. Os critérios da avaliação variam de acordo com o problema.

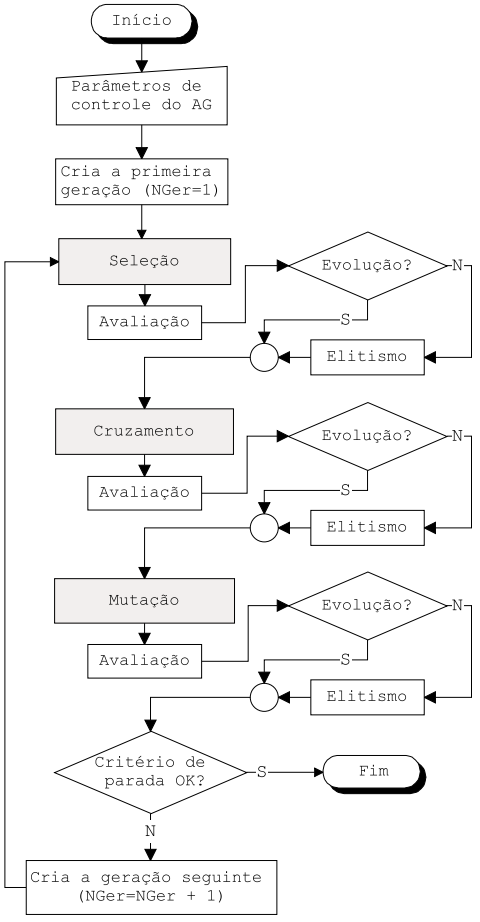


Figura 1. *Esquema geral de funcionamento de um algoritmo genético típico*

O princípio de funcionamento do AG é relativamente simples e pode-se compreendê-lo através de um exemplo prático (ver A Figura 2A-D). Seja o conjunto de dados da Tabela 1, em que se deseja descobrir qual combinação de duas variáveis é capaz de produzir o maior coeficiente de correlação, R. Sabe-se, da Tabela 2, que a solução desse problema é a equação No. 12, $\log 1/C = f(\pi_m, r_v^p)$, que apresenta R = 0,936.

A resolução deste problema através de AG requer algum trabalho. O primeiro passo é a criação de um conjunto inicial constituído por *N* equações de regressão contendo, cada uma, duas variáveis distintas escolhidas aleatoriamente dentre as disponíveis no conjunto de dados. Esse conjunto inicial de equações é denominado *pri-*

meira geração (G_1). Cada uma das equações E representa um indivíduo da geração. Cada indivíduo é uma solução em potencial para o problema. As variáveis de uma equação correspondem aos genes do indivíduo. O conjunto de variáveis que caracteriza uma equação corresponde ao cromossomo do indivíduo.

O valor de *N* pode, em princípio, ser escolhido livremente. Quanto maior for *N* maior será o esforço computacional para resolver o problema. Em contrapartida, quanto mais indivíduos tiver cada geração, mais rapidamente chegar-se-á à solução do problema. O fundamento do AG é aplicar as regras da evolução para produzir gerações seguintes mais evoluídas (que representem melhores soluções para o problema) do que as gerações anteriores. Seja $E_{i,j}$ o j-ésimo indivíduo da iésima geração. Supondo-se em nosso exemplo que cada geração terá número fixo de quatro indivíduos (*N* = 4), a população de G_1 será constituída por $E_{1,1}$, $E_{1,2}$, $E_{1,3}$ e $E_{1,4}$ (Figura 2A). Vamos supor também que para G_1 tenham sido sorteadas as seguintes variáveis para cada indivíduo: $E_{1,1} = (\sigma_m, r_v^p)$, $E_{1,2} = (\sigma^+, r_v^p)$, $E_{1,3} = (\pi, \pi_m)$ e $E_{1,4} = (\pi, \sigma_m)$. Após o sorteio, deve-se proceder à avaliação de G_1. Isso é feito através do cálculo do coeficiente de correlação de cada uma das quatro equações. No presente caso, $R_{1,1} = 0{,}879$, $R_{1,2} = 0{,}878$, $R_{1,3} = 0{,}844$ e $R_{1,4} = 0{,}406$ (ver Tabela 2). Como $R_{1,1}$ é o maior dos quatro, $E_{1,1}$ representa, até o momento, a melhor solução para o problema ou o melhor indivíduo. A avaliação de cada geração é feita com base no somatório de R. Para G_1, $\sum R_1 = 3{,}007$. Além do coeficiente de correlação, a avaliação de cada geração pode ser feita através do cálculo do desvio-padrão (s) ou do teste de Fischer (F)[5-7]. No caso da escolha de F, $\sum F$, deverá apresentar o maior valor possível. No caso de s, $\sum s$ deverá ser o menor possível.

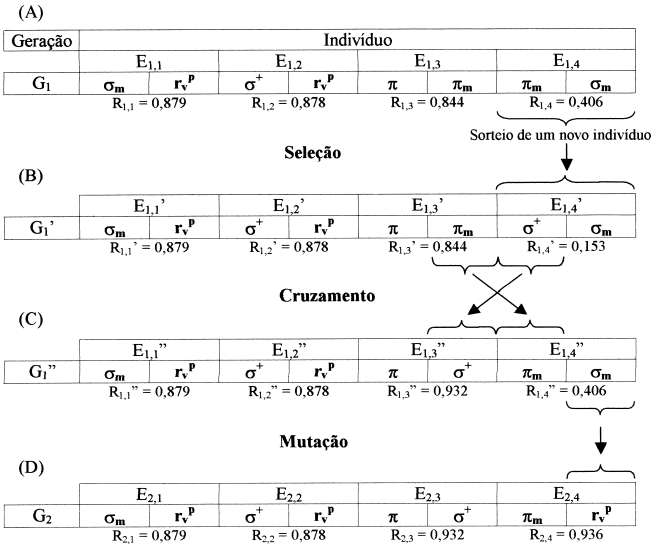


Figura 2. *Exemplo de funcionamento do algoritmo genético. (A) Geração inicial; (B) Geração resultante da aplicação da seleção natural, (C) cruzamento e (D) mutação*

A construção da geração seguinte, G_2, resulta da execução das três operações evolutivas sobre G_1. Como G_2 somente é formada ao final de todo o processo, adotar-se-ão nomes para as gerações intermediárias, utilizando-se apóstrofos. Portanto, partindo-se de G_1, após a seleção natural teremos G_1', após o cruzamento G_1" e após a mutação, que marca o final do processo, G_2.

A primeira das operações evolutivas é a seleção natural. Na seleção, um ou mais indivíduos existentes em G_1 serão copiados (reproduzidos) para G_1', sendo que a escolha destes é feita através de sorteio. O sorteio, porém, não é aleatório, mas sujeito a uma condição. A probabilidade de cada indivíduo de G_1 ser reproduzido para G_1' é proporcional ao valor de sua avaliação. Aqueles capazes de produzir valores elevados de R têm maior probabilidade de serem reproduzidos que os demais. Pode-se permitir que dado indivíduo de G_1 seja reproduzido mais de uma vez para G_1'. No entanto, este procedimento diminui a variabilidade da geração. Alternativamente, pode-se selecionar para a geração seguinte os N - *x* indivíduos com maior R e completar a geração através do sorteio de *x* indivíduos diferentes dos demais.

Vamos supor que sejam selecionados para G_1' os três melhores indivíduos de G_1 ($E_{1,1}$, $E_{1,2}$ e $E_{1,3}$) e que G_1' seja completada pelo sorteio de um novo indivíduo ($E_{1,4}$'). Vamos supor também que este corresponde à equação No. 13 da Tabela 2, $\log 1/C = f(\sigma^+, \sigma_m)$. O resultado da seleção natural é mostrada na Figura 2B. A avaliação de G_1' ($\sum R_1$') vale 2,754, que é menor do que $\sum R_1$. Neste caso, pode-se aplicar diretamente o elitismo e recompor a geração anterior (G_1'= G_1) ou condicionar a aplicação do elitismo a um fator de probabilidade qualquer. No presente caso, o elitismo não será aplicado.

A próxima operação evolutiva é o cruzamento entre pares de indivíduos, também chamada de *crossover*. A idéia central do *crossover*, como o nome sugere, é cruzar os genes (o conjunto das variáveis) de um ou mais pares de indivíduos, fazendo com que a prole gerada nessa operação herde parte dos genes de cada um dos progenitores. No nosso exemplo, vamos supor que os dois piores indivíduos ($E_{1,3}$' e $E_{1,4}$') sofrerão cruzamento. Para este par é sorteada uma posição na seqüência de variáveis que corresponde à posição de *crossover*. Como no presente caso cada indivíduo tem apenas dois genes (variáveis), seus cromossomos serão divididos ao meio e permutados. A Figura 2B-C mostra de forma esquemática o procedimento descrito acima.

Após o cruzamento, $E_{1,3}$' (π, π_m; $R_{1,3}$' = 0,844) foi transformado em $E_{1,3}$" (π, σ^+; $R_{1,3}$" = 0,932) e $E_{1,4}$' (σ^+, σ_m; $R_{1,4}$' = 0,153) foi transformado em $E_{1,4}$" (π_m, σ_m; $R_{1,4}$" = 0,406). A avaliação da geração melhorou de $\sum R_1$' = 2,754 para $\sum R_1$" = 3,095, maior ainda do que a avaliação de G_1 (3,007).

A última operação evolutiva antes de completar-se a construção de G_2 é a mutação. Esta consiste em substituir aleatoriamente uma ou mais variáveis, de um ou mais indivíduos da geração, por outra variável que não exista entre as que não foram substituídas. Supondo-se que apenas $E_{1,4}$" (π_m, σ_m) sofrerá mutação e que esta consistirá na troca de σ_m por r_v^p, o indivíduo assim formado ($E_{2,4}$) será $\log 1/C = f(\pi_m, r_v^p)$, cuja avaliação é $R_{2,4} = 0{,}936$. A configuração da nova geração G2 é mostrada na Figura 2D.

A avaliação de G_2 ($\sum R_2 = 3{,}625$) é superior à avaliação da geração G_1 ($\sum R_1 = 3{,}007$). Isso significa que, de G_1 para G_2, os valores de R evoluíram no sentido da melhor solução para o problema, que é R = 0,936. Segundo o nosso exemplo, o indivíduo $E_{2,4}$ já possui essa avaliação, o que indica que a busca chegou ao fim.

Embora esse exemplo seja fictício e seus resultados tenham sido manipulados para servirem ao seu propósito didático, não está longe da realidade. É evidente que, em sistemas complexos, onde não se conhece a equação que possui o melhor coeficiente de correlação, não haverá garantias de que a melhor equação será encontrada. Para aumentar a probabilidade de encontrar essa equação, deve-se executar a busca utilizando-se muitos indivíduos em cada geração. Isso implicaria em maiores possibilidades de evolução durante a execução dos três processos evolutivos. Além disso, vários esquemas de mutação podem ser adotados, principalmente esquemas de mutações exaustivas sobre os diversos indivíduos de cada geração. Finalmente, o processo evolutivo poderá ser executado através de grande número de gerações, cuja escolha depende da complexidade do problema.

Em QSAR, a utilização de AGs tem-se mostrado promissora na busca de combinações de variáveis que resultam em modelos linea-

res de melhor qualidade[24-26]. Alguns algoritmos pseudo-evolucionários, que não executam cruzamento entre os indivíduos e que utilizam combinação de AG e busca sistemática[16,27,28], também vêm apresentando bons resultados na solução da matriz de Selwood[15].

A aplicação de AG à seleção de variáveis em QSAR somente será vantajosa nos casos em que a busca sistemática mostrar-se inviável. O AG possui a notável propriedade de vasculhar o hiperespaço constituído de modelos lineares multidimensionais e dele extrair o que há de melhor. Pode-se argumentar sobre a real eficiência dos AGs na busca por modelos lineares em conjuntos de dados realmente grandes. É possível aumentar significativamente a eficiência do AG através de técnica de programação apurada e da especificação adequada dos diversos parâmetros que controlam a busca. Aqui também valem os procedimentos de aceleração de cálculo citados na busca sistemática, tais como evitar a combinação de variáveis muito correlacionadas e executar o cálculo de R, s ou F sem utilizar o método dos mínimos quadrados.

MÉTODOS QUIMIOMÉTRICOS

Como já mencionado anteriormente, em QSAR, são duas as principais abordagens utilizadas na construção dos modelos matemáticos. A primeira utiliza o método de RLM, que se adapta ao tratamento de dados através do método clássico de Hansch-Fujita[1-4]. A segunda abordagem utiliza métodos quimiométricos, baseados na análise de componentes principais[29]. O método de RLM tem sido historicamente o mais utilizado devido à sua simplicidade e à facilidade de interpretação dos resultados. No entanto, a aplicação de métodos quimiométricos em QSAR vem crescendo a cada dia e hoje há dúvidas sobre qual das duas abordagens é a mais utilizada[30].

Há diferenças importantes nos fundamentos desses métodos. A RLM apresenta a desvantagem de ser extremamente sensível à presença de colinearidade entre os descritores. Na aplicação do método de RLM a um conjunto de variáveis altamente correlacionadas, os coeficientes da regressão podem tornar-se instáveis e sem significado. Este problema pode ser contornado através da utilização de métodos quimiométricos, que não apresentam essa limitação, como é o caso dos métodos que utilizam a análise de componentes principais como fundamento. Na verdade, esses métodos tiram vantagem da existência de colinearidade, como veremos adiante. Conseqüentemente, os métodos quimiométricos muito têm a oferecer ao processo de construção de modelos de QSAR, especialmente em sistemas complexos, em que se dispõem de muitos descritores (X). Em QSAR-3D, por exemplo, em que o número de variáveis geradas é da ordem de centenas ou milhares, a utilização de métodos quimiométricos que fazem uso da PCA são especialmente indicados.

Os métodos quimiométricos também podem ser úteis no processo de seleção de variáveis. A combinação de diferentes metodologias neste processo pode ser vantajosa. Por exemplo, numa primeira etapa pode-se utilizar PCA para explorar a estrutura dos dados e então, numa segunda etapa podem-se propor modelos quantitativos entre a atividade biológica e os descritores físico-químicos e estruturais. A primeira etapa consiste na visualização das propriedades das moléculas com base nos descritores considerados, sem o objetivo de descobrir ou propor qualquer relação causa-efeito. Isto é importante para visualizar a correlação entre os descritores e agrupamentos entre os compostos. A segunda etapa consiste na utilização desse espaço de descritores para a modelagem, que é feita através de métodos de regressão, como por exemplo o de regressão por mínimos quadrados parciais[8]. Nesta seção, será considerada a aplicação da combinação das metodologias PCA e PLS.

A análise de componentes principais baseia-se na projeção linear do espaço original das variáveis X's, que possui m dimensões (cada dimensão representando uma variável), num subespaço com k dimensões (cada dimensão desse subespaço representando uma componente principal) que preserve a maior variância possível do conjunto de dados. Em outras palavras, PCA é capaz de transformar dados complexos e apresentá-los numa nova perspectiva em que se espera que as informações mais importantes tornem-se evidentes. Na análise de componentes principais, a atividade biológica não é considerada diretamente, visto que PCA utiliza apenas as variáveis descritivas. PCA está fundamentada na correlação entre as variáveis, agrupando aquelas que são mais correlacionadas numa nova variável chamada *componente principal*. As novas variáveis representadas pelas componentes principais são mais informativas e em menor número do que os descritores originais, isto é, $k < m$. As componentes principais são obtidas através da combinação linear dos descritores originais e apresentam as propriedades de serem mutuamente ortogonais e definidas em ordem decrescente da quantidade de variância que são capazes de explicar. Em outras palavras, a informação contida numa componente principal não está presente em outra e a variância que elas descrevem é uma medida da quantidade de informação que as mesmas contém[31]. As relações entre os compostos não são alteradas nessa transformação. Porém, como os novos eixos são ordenados pela sua importância, ou seja, pela ordem de variância citada acima, pode-se visualizar a estrutura do conjunto de dados em gráficos de baixa dimensionalidade (como por exemplo PC1 ´ PC2, PC1 ´ PC3, etc.).

Do ponto de vista matemático a matriz de dados $\mathbf{X}$ é decomposta em duas matrizes, uma de escores ($\mathbf{T}$) e uma de pesos (*loadings*) $\mathbf{L}^T$, ou seja, $\mathbf{X}=\mathbf{T}\,\mathbf{L}^T$. Os escores são as novas coordenadas de cada composto no novo sistema de eixos e a informação de quanto cada descritor original contribui, está contida nos pesos. Os escores $\mathbf{T}$ expressam as relações entre as amostras enquanto que os pesos $\mathbf{L}^T$ mostram as relações entre as variáveis.

No processo de seleção de variáveis, PCA identifica agrupamentos de descritores, proporcionando um entendimento de como eles estão correlacionados e o quanto de informação eles contém. Este é o fundamento de seu uso na seleção das variáveis. No entanto, freqüentemente o número de descritores originais selecionados é maior que o desejado. Embora isso não constitua problema do ponto de vista matemático, corre-se o risco de tornar a interpretação físico-química difícil ou mesmo impossível. Portanto, deve haver equilíbrio entre o aspecto matemático e a possibilidade das correlações encontradas poderem ser explicadas através de hipóteses físico-químico-biológicas. Os resultados da análise de componentes principais costumam ser visualizados em gráficos, facilitando a identificação de agrupamentos. Por exemplo, um gráfico de pesos contém informação sobre as variáveis e é usado para determinar quais delas são mais importantes para descrever a variação dos dados originais. Um gráfico de escores contém informação sobre os compostos tornando visível a similaridade, agrupamentos e diferenças entre os mesmos, com base nas variáveis utilizadas. Portanto, é importante que estes gráficos sejam analisados em conjunto.

Para encontrarmos a relação entre os blocos X dos descritores e Y das atividades, utiliza-se o método PLS[8]. Além de não ser sensível às colinearidades entre os descritores, PLS oferece outras vantagens sobre a RLM: (a) a razão entre o número de descritores e o número de compostos não é limitada como em RLM, que requer mais compostos que descritores, e; (b) RLM está fundamentada no fato de que todas as variáveis selecionadas são importantes para o problema, ou seja, a dimensionalidade está sendo fixada a priori, o que não é o caso quando se usa o método PLS. Portanto, a escolha natural é adotar o método PLS para a modelagem estrutura-atividade.

Um dos métodos centrais desta abordagem consiste em obter uma descrição da variável dependente Y como uma combinação li-

near das variáveis originais, por meio das componentes principais, as quais não são correlacionadas entre si. Neste caso ainda, o algoritmo PLS passa por uma etapa preliminar para assegurar que as componentes principais geradas sejam relevantes para a atividade biológica. A resposta Y é inicialmente usada para encontrar um padrão dentro dos dados X que se correlacione com a mesma. Em outras palavras, as componentes principais são otimizadas para melhor descrever a relação entre os blocos X e Y, simultaneamente (e por isto chamadas de variáveis latentes). Só então estas variáveis latentes são usadas para modelar a atividade Y. O vetor de regressão (coeficientes da regressão) indica quais descritores são importantes na modelagem da resposta biológica Y[31].

Pode-se ilustrar o processo de seleção de variáveis através dos métodos quimiométricos utilizando-se o conjunto de dados da Tabela 1, que contém cinco variáveis independentes (m = 5). Os resultados da análise de componentes principais são mostrados na Tabela 4 e na Figura 3. Os dados foram autoescalados antes da análise (centrados na média e escalados para variância unitária). Os cálculos envolvendo PCA e PLS normalmente são executados através de programas computacionais especializados, ao contrário da RLM, cujas rotinas são encontradas na maioria das calculadoras científicas programáveis. Os cálculos desta seção foram feitos com o programa Pirouette[32], que possui grande flexibilidade em termos de análise multivariada.

Tabela 4. Análise de componentes principais aos dados da Tabela 1

Componente principal	Valor	Variância Percentual	Acumulada
PC1	57,32	54,59	54,59
PC2	32,28	30,74	85,33
PC3	11,37	10,82	96,16
PC4	3,97	3,78	99,94
PC5	0,064	0,061	100,00

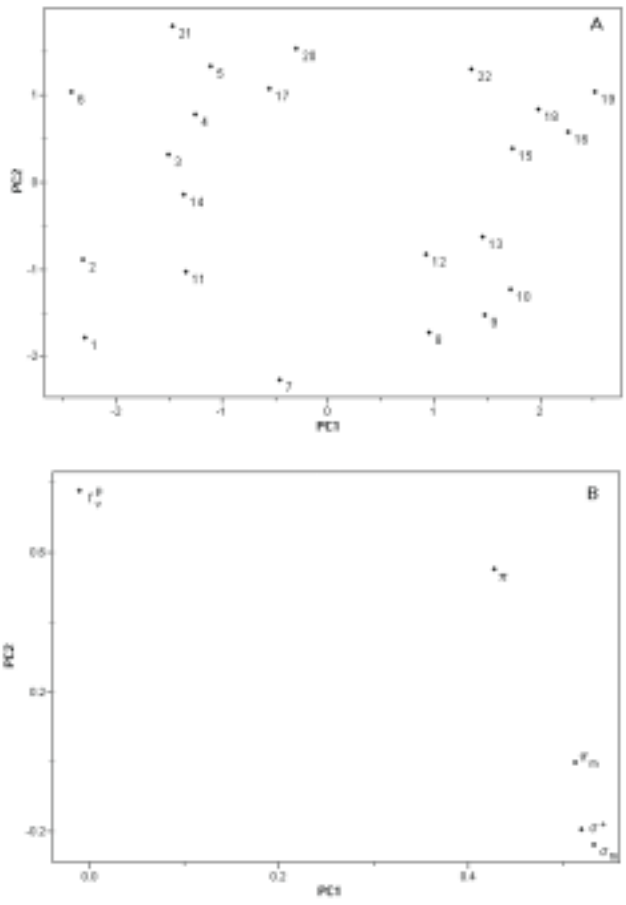


Figura 3. *Gráficos de escores (A) e pesos (B) para as duas primeiras componentes principais (PC1 e PC2)*

A Tabela 4 contém a variância que cada componente principal descreve com a respectiva percentagem, e a variância total acumulada com uma, duas, etc, componentes principais. Pode-se ver que duas ou no máximo três componentes principais são suficientes para uma boa descrição dos dados originais (k = 2 ou 3). Uma equação com duas componentes principais contém 85,33% da informação original dos dados, enquanto que uma com três contém 96,15%. É interessante nesta análise verificar também como os compostos estão agrupados.

O gráfico dos escores na Figura 3A mostra como os compostos se agrupam com base nos descritores considerados. Os pesos na Figura 3B indicam o quanto cada descritor contribui para a formação das novas variáveis (PCs). Pode-se ver que a primeira componente, PC1, tem alta contribuição de todos os descritores, exceto r_v^p, enquanto que a segunda contém informação principalmente das variáveis r_v^p e π. As expressões matemáticas para PC1 e PC2 são mostradas nas eqs 20 e 21

$$PC1 = -0{,}01r_v^p + 0{,}4\pi + 0{,}5\pi_m + 0{,}5\sigma_m + 0{,}5\,\sigma^+ \quad (20)$$

$$PC2 = 0{,}8\,r_v^p + 0{,}6\pi - 0{,}0\pi_m - 0{,}2\sigma_m - 0{,}2\,\sigma^+ \quad (21)$$

Substituindo-se os valores de cada variável para um determinado composto na eq 20, obtém-se o escore daquela amostra no eixo PC1.

A primeira componente principal, PC1 (eq 20), discrimina dois grupos de compostos, um deles contendo os compostos 1, 2, 3, 4, 5, 6, 7, 11, 14, 17, 20 e 21, localizados á esquerda do gráfico de escores (escores negativos). Como mencionado acima, a primeira componente principal tem alta contribuição das variáveis π, π_m, σ_m e σ^+, portanto, aqueles compostos que estão deslocados para a esquerda devem ter valores numéricos pequenos em uma ou mais destas variáveis. A segunda componente principal, PC2 (eq 21), contém a informação que desejamos. Está claro na figura dos escores que PC2 está relacionada com a atividade biológica. Os compostos mais potentes estão na parte inferior do gráfico da Figura 3A (escores negativos) enquanto que os menos potentes estão na parte superior, com escores positivos. Por que os compostos têm esta estrutura em PC2? A resposta a esta pergunta é dada pelos pesos (Figura 3B). Quanto maior a contribuição simultânea das duas variáveis r_v^p e π, isto é, quanto maiores os valores do raio de van der Waals do substituinte na posição *para* e da constante lipofílica de Hansch em um determinado composto, maiores os valores dos escores e, portanto menor a potência do mesmo. Isto não quer dizer que as outras variáveis não contribuem para a atividade biológica. Significa apenas que elas contribuem em grau bem menor. Os descritores σ^+ e σ_m tem pesos negativos em PC2, indicando que os compostos mais ativos (escores negativos) tendem a ter maiores valores destas duas variáveis além das características mencionadas anteriormente. Podemos concluir *a priori* que as variáveis r_v^p e π são muito importantes na modelagem da estrutura-atividade biológica. Feita esta análise preliminar, deve-se construir o modelo de regressão através do método PLS. Em análise quimiométrica, geralmente utiliza-se a validação cruzada para verificar o grau de predizibilidade do modelo. Neste caso, o desvio-pa-

drão (s_{DEP}) é utilizado como principal critério de verificação[31]. Deve-se salientar que s_{DEP} é diferente do desvio-padrão, s, utilizado como parâmetro de avaliação em RLM. O valor de s_{DEP} é dado pela eq 22,

$$s_{DEP} = \sqrt{\frac{\sum_i \left(Y_i - \hat{Y}_i\right)^2}{n}}, \quad (22)$$

onde *n* é o número de amostras do conjunto de dados e $(Y_i - \hat{Y}_i)$ é o desvio da previsão.

Como no método PLS as componentes principais são *otimizadas* para melhor se correlacionarem com a atividade, elas são designadas *variáveis latentes* para se distinguirem das componentes principais. Os resultados obtidos para o modelo PLS com validação cruzada excluindo-se duas amostras de cada vez são mostrados na Tabela 5.

Tabela 5. Desvio-padrão da validação cruzada (s_{DEP}), coeficiente de correlação da validação cruzada (R_{cv}), coeficiente de correlação (R) e desvio-padrão (s) para modelos considerando entre uma e cinco variáveis latentes

	s_{DEP}	R_{cv}	R	s
PC1	0,222	0,919	0,948	0,184
PC2	**0,190**	**0,940**	**0,958**	**0,171**
PC3	0,201	0,933	0,962	0,168
PC4	0,212	0,925	0,963	0,171
PC5	0,220	0,920	0,964	0,172

A equação com duas variáveis latentes (*k* = 2, ou seja, onde a dimensão dos dados originais foi reduzida de cinco para dois) apresenta o menor desvio-padrão de previsão e o mais alto coeficiente de correlação de validação cruzada (PC2; s_{DEP} = 0,190 e R_{cv} = 0,940). Considerando-se o desvio-padrão, s, três variáveis latentes deveriam ser consideradas (PC3; s = 0,168 e R = 0,962). Deve-se notar que os valores de s_{DEP} são sempre maiores que os de s. Isto é esperado, uma vez que as previsões são feitas sobre amostras não incluídas nos cálculos, fazendo com que s_{DEP} seja um parâmetro mais confiável para escolha do melhor candidato a modelo. Comparando os resultados de PLS obtidos utilizando as 5 variáveis originais com os da busca sistemática, o valor de s = 0,168 para três variáveis latentes é menor do que o encontrado para a equação No. 31 com cinco descritores da Tabela 2 cujo desvio-padrão é s = 0,172 e muito semelhante à equação selecionada através da busca sistemática com três descritores (No. 20 em que s = 0,166). Concluindo, o modelo que foi obtido com duas componentes principais e os cinco descritores originais é bem semelhante ao melhor modelo obtido através da busca sistemática, sem que nenhuma seleção de variáveis tenha sido feita.

Por outro lado, caso seja desejável executar seleção de variáveis, devem-se analisar conjuntamente os coeficientes de regressão e os pesos, que são os coeficientes das variáveis latentes uma vez que o modelo PLS foi construído através delas. A Tabela 6 contém os coeficientes de regressão (um para cada variável) e os pesos das duas primeiras variáveis latentes. A equação matemática (modelo) que fornece os valores previstos da atividade biológica (valores autoescalados) em função das variáveis autoescaladas é dada pela eq 23.

$$\log 1/IC = 0{,}462\pi + 0{,}110\pi_m - 0{,}164\sigma^+ - 0{,}132\sigma_m + 0{,}572 r_v^p \quad (23)$$

É possível confirmar que π e r_v^p são os descritores que mais contribuem para a explicação da variabilidade dos valores da atividade biológica. Isto está visível nos altos pesos da primeira variável latente e nos altos coeficientes de regressão mostrados na Tabela 6. A

Tabela 6. Coeficientes de regressão e pesos para o modelo obtido através de PLS com duas variáveis latentes (VL) para todos os descritores

	Coef. regressão	Pesos	
		VL1	VL2
π	**0,462**	**0,635**	-0,290
π_m	0,110	0,172	-0,253
σ^+	-0,164	-0,127	**-0,757**
σ_m	-0,132	-0,112	-0,519
r_v^p	**0,572**	**0,734**	0,099

variável σ^+ apesar de não ter alto coeficiente de regressão como as duas anteriores, tem peso bastante significativo na segunda variável latente e portanto mostra sua contribuição na modelagem. Caso o objetivo fosse construir um modelo com três descritores, π, σ^+ e r_v^p seriam os selecionados. Esta escolha coincide com a melhor equação de três variáveis mostrada na Tabela 2. Para um modelo com quatro descritores, a escolha natural seria a inclusão de σ_m, que apesar de não ter um coeficiente de regressão tão significativo, tem sua contribuição através do peso na segunda variável latente (Tabela 6). O conjunto de variáveis π, r_v^p, σ^+ e σ_m, corresponde à equação 29 da Tabela 2 que apesar de não ter sido a escolhida através da busca sistemática com 4 descritores, apresenta resultados tão bons quanto a equação selecionada (No. 26, Tabela 2).

Concluindo, os métodos quimiométricos podem nos auxiliar na seleção de variáveis ao mesmo tempo em que permite a visualização da estrutura dos dados.

CONCLUSÕES

As três abordagens de seleção variáveis para modelos de relações estrutura-atividade analisadas possuem características distintas. A busca sistemática é o melhor método para modelos com até cinco variáveis. Para modelos com mais variáveis, é aconselhável a utilização de algum método baseado em algoritmo genético, que é mais rápido e eficiente. Nossa experiência mostra que, considerando-se dado número de variáveis, os algoritmos genéticos podem localizar os melhores modelos até seis vezes mais rápido do que a busca sistemática. No entanto, outros métodos que não envolvem a construção de algoritmos específicos para a busca de combinações de variáveis podem ser utilizados. Os métodos quimiométricos baseados em PCA/PLS têm sido muito utilizados ultimamente. Neste caso, a idéia é inicialmente utilizar grande número de variáveis e deixar o próprio procedimento PCA/PLS selecionar aquelas que são mais representativas.

CÁLCULOS

Os cálculos executados na seção *Busca Sistemática* foram executados com o programa BuildQSAR[33], enquanto que os cálculos da seção *Métodos Quimiométricos* foram executados com o programa Pirouette[32].

AGRADECIMENTOS

Os autores são gratos às Fundações de Amparo à Pesquisa do Estado de São Paulo-Fapesp e de Minas Gerais-Fapemig, à Pró-Reitoria de Pesquisa e Pós-Graduação da Universidade Federal do Espírito Santo-PRPPG-UFES e ao Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq pelo auxílio financeiro.

REFERÊNCIAS

1. Gaudio, A. C.; *Quim. Nova* **1996**, *19*, 278.
2. Hansch, C.; Leo, A.; *Exploring QSAR: Fundamentals and Applications in Chemistry and Biology*, American Chemical Society: Washington, DC, 1995.
3. Kubinyi, H. Em *Methods and Principles in Medicinal Chemistry*; Mannhold, R.; Krogsgaard-Larsen, P.; Timmerman, H., eds.; VCH: Weinheim, 1993; vol. 1.
4. Martin, Y. C.; *Quantitative Drug Design: A Critical Introduction*, Marcel Dekker: New York, 1978.
5. Daniel, W. W.; *Biostatistics: A Foundation for Analysis in the Health Sciences*, John Wiley & Sons: New York, 1995.
6. Draper, N. R.; Smith, H.; *Applied Regression Analysis*, John Wiley & Sons: New York, 1981.
7. Myers, R. H.; *Classical and Modern Regression with Applications*, Duxbury Press: Boston, 1986.
8. Geladi, P.; Kowalski, B. R.; *Anal. Chim. Acta* **1986**, *185*, 1.
9. Unger, S. H.; Hansch, C.; *J. Med. Chem.* **1973**, *16*, 745.
10. Topliss, J. G.; Costello, R. J.; *J. Med. Chem.* **1972**, *15*, 1066.
11. Hansch, C.; Fujita, T.; *J. Am. Chem. Soc.* **1964**, *86*, 1616.
12. Cramer, R. D., III; Patterson, D. E.; Bunce, J. D.; *J. Am. Chem. Soc.* **1988**, *110*, 5959.
13. Karelson, M.; Lobanov, V. S.; Katritzky, A. R.; *Chem. Rev.* **1996**, *96*, 1027.
14. Gaudio, A. C.; Zandonade, E.; *Quim. Nova* **2001**, *24*, 658.
15. Selwood, D. L.; Livingstone, D. J.; Comley, J. C. W.; O'Dowd, A. B.; Hudson, A. T.; Jackson, P.; Jandu, K. S.; Rose, V. S.; Stables, J. N.; *J. Med. Chem.* **1990**, *33*, 136.
16. Kubinyi, H.; *Quant. Struct. Act. Relat.* **1994**, *13*, 393.
17. Hansch, C.; Leo, A.; Hoekman, D.; *Exploring QSAR: Hydrophobic, Electronic, and Steric Constants*, American Chemical Society: Washington, DC, 1995.
18. Hansch, C.; Leo, A.; Taft, R. W.; *Chem. Rev.* **1991**, *91*, 165.
19. Cammarata, A.; *J. Med. Chem.* **1972**, *15*, 573.
20. Graham, J. D. P.; Karrar, M. A.; *J. Med. Chem.* **1963**, *6*, 103.
21. Sutton, P.; Boyden, S.; *Am. J. Phys.* **1994**, *62*, 549.
22. Goldberg, D. E.; *Genetic Algoritms in Search, Optimization and Machine Learning*, Addison-Wesley: New York, 1989.
23. Da Costa, L. R.; Gaudio, A. C.; *Rev. Eng. Ciência Tecnol.* **2000**, *3*, 43.
24. Maddalena, D. J.; Snowdon, G. M.; *Exp. Opin. Ther. Patents* **1997**, *7*, 247.
25. Maddalena, D. J.; *Exp. Opin. Ther. Patents* **1998**, *8*, 249.
26. Hasegawa, K.; Kimura, T.; Funatsu, K.; *J. Chem. Inf. Comp. Sci.* **1999**, *39*, 112.
27. Kubinyi, H.; *J. Chemom.* **1996**, *10*, 119.
28. Kubinyi, H.; *Quant. Struct.-Act. Relat.* **1994**, *13*, 285.
29. Wold, S.; Esbensen, K.; Geladi, P.; *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37.
30. Leitão, A.; Montanari, C. A.; Donnici, C. L.; *Quim. Nova* **2000**, *23*, 178.
31. Ferreira, M. M. C.; Antunes, A. M.; Melo, M. S.; Volpe, P. L. O.; *Quim. Nova* **1999**, *22*, 724.
32. *Pirouette for Windows*; Versão 3.0; Infometrix, Woodinville, WA, 2000.
33. De Oliveira, D. B.; Gaudio, A. C.; *Quant. Struct.-Act. Relat.* **2000**, *19*, 599.