Quim. Nova, Vol. 36, No. 4, 554-560, 2013

Artigo

QSAR MODELING: UM NOVO PACOTE COMPUTACIONAL *OPEN SOURCE* PARA GERAR E VALIDAR MODELOS QSAR

João Paulo A. Martins e Márcia M. C. Ferreira*
Instituto de Química, Universidade Estadual de Campinas, CP 6154, 13084-971 Campinas – SP, Brasil

Recebido em 4/6/12; aceito em 15/11/12; publicado na web em 12/3/13

QSAR MODELING: A NEW OPEN SOURCE COMPUTATIONAL PACKAGE TO GENERATE AND VALIDATE QSAR MODELS. *QSAR modeling* is a novel computer program developed to generate and validate QSAR or QSPR (quantitative structure- activity or property relationships) models. With *QSAR modeling*, users can build partial least squares (PLS) regression models, perform variable selection with the ordered predictors selection (OPS) algorithm, and validate models by using **y**-randomization and leave-*N*-out cross validation. An additional new feature is outlier detection carried out by simultaneous comparison of sample leverage with the respective Studentized residuals. The program was developed using Java version 6, and runs on any operating system that supports Java Runtime Environment version 6. The use of the program is illustrated. This program is available for download at lqta.iqm.unicamp.br.

Keywords: QSAR models; OPS variable selection; outlier detection.

INTRODUÇÃO

O estudo das relações quantitativas entre a estrutura química e a atividade biológica ou alguma propriedade físico-química (QSAR/QSPR) é uma área de destaque hoje na comunidade científica. Por exemplo, na área da físico-química estudos de QSPR são essenciais na predição de propriedades que são difíceis de serem medidas experimentalmente. Já na área de química medicinal teórica, a predição da atividade biológica de novos compostos usando relações matemáticas baseadas em propriedades estruturais, físico-químicas e conformacionais de potenciais agentes previamente testados é um campo de pesquisa extremamente ativo e promissor. Relações QSAR são úteis para entender e explicar o mecanismo de ação de fármacos em nível molecular e permite o projeto e o desenvolvimento de novos compostos com propriedades biológicas desejáveis.[1]

Um modelo quantitativo QSAR (ou QSPR) é representado por meio de uma equação matemática que relaciona as propriedades dos compostos investigados com suas atividades biológicas e que possui significância estatística. Essa equação deve não somente possuir um bom poder de predição, mas deve também ser validada mostrando-se robusta e não obtida ao acaso.[2-7]

Existem diversos programas disponíveis na literatura que podem ser utilizados para gerar modelos QSAR. Entre eles, alguns dos mais conhecidos são: MobyDigs,[8] BuildQSAR,[9] VCCLAB,[10,11] QSAR+,[12] BILIN,[13] MOLGEN QSPR,[14] CORAL,[15] CODESSA PRO,[16] WOLF.[17] A Tabela 1 mostra uma comparação das principais características presentes no programa *QSAR modeling* com os programas supracitados. É notório que dentre os programas livres, apenas o *QSAR modeling* incorpora todos os testes sugeridos na literatura para a validação[3] e obtenção de modelos robustos, não obtidos por correlações espúrias e com a avaliação crítica dos compostos com comportamento atípico.

Neste trabalho, é apresentado um novo programa *open source*, denominado *QSAR modeling*, cujo objetivo é construir e validar modelos de QSAR utilizando as ferramentas quimiométricas. Esse é o primeiro programa que implementa o método de seleção de variáveis recentemente desenvolvido *ordered predictors selection* (OPS),[18] incorpora os processos de validação cruzada *leave-N-out* e aleatorização de y (*y-randomization*) além de realizar a detecção de amostras anômalas conhecidas na literatura como *outliers*. A detecção destes compostos, frequentemente negligenciada em programas de QSAR, é implementada combinando os valores de influência (*leverage*) das amostras aos seus respectivos resíduos de Student. Este é um procedimento usual em quimiometria, mas que se mostra ausente nos programas livres citados anteriormente. O programa BuildQSAR é o único que inclui uma metodologia para a detecção de amostras anômalas, analisando o desvio padrão dos resíduos.

O processo de construção de modelos usando o programa *QSAR modeling* é descrito através de um conjunto de dados formado por 37 hidrocarbonetos poliaromáticos tendo o log P (logaritmo do coeficiente de partição octanol-água) como variável dependente.[19] Além dos descritores disponibilizados no conjunto de dados, foram utilizados descritores topológicos calculados com o programa DRAGON 6.[20]

Tabela 1. Comparativo entre as principais características do programa *QSAR modeling* e outros programas disponíveis na literatura

Programa	Teste de robustez[a]	Teste de correlações ao acaso[b]	Detecção de amostras anômalas	Programa livre
MobyDigs	Não	Sim	Não	Não
BuildQSAR	Não	Não	Sim	Sim
VCCLAB	Não	Não	Não	Sim
QSAR+	Não	Sim	Sim	Não
BILIN	Não	Não	Não	Sim
MOLGEN QSPR	Não	Sim	Não	Não
CORAL	Não	Não	Não	Sim
CODESSA PRO	Não	Sim	Sim	Não
WOLF	Não	Não	Sim	Não
QSAR Modeling	Sim	Sim	Sim	Sim

[a] Validação cruzada feita excluindo *N* amostras (*leave-N-out*). [b] Aleatorização de **y**.

*e-mail: marcia@iqm.unicamp.br

PARTE EXPERIMENTAL

O programa *QSAR modeling* foi desenvolvido na linguagem Java[21] e tem uma estrutura orientada a objetos. Foi projetado para ser executado em qualquer sistema operacional (Windows XP, Windows Vista, Windows 7, Linux, Mac OS, Solaris, entre outros), pois a máquina virtual java (JVM) está disponível para esses sistemas. Para executar o programa *QSAR modeling* é necessário ter o ambiente de execução java (JRE) versão 6 instalado no sistema operacional.

RESULTADOS E DISCUSSÃO

As entradas para a execução do programa *QSAR modeling* são dois arquivos texto contendo, respectivamente, a matriz com os valores numéricos dos descritores (geralmente chamada de matriz **X** com *I* linhas e *J* colunas) e o vetor contendo as atividades biológicas (designado vetor **y** com *I* elementos) para os *I* compostos sob investigação. No arquivo contendo os descritores o usuário pode, opcionalmente, adicionar o nome de cada um deles na primeira linha. A tela principal do programa (Figura 1S, material suplementar), assim como as telas de entrada de dados disponíveis para o usuário se encontram no material suplementar.

O programa *QSAR modeling* incorpora as seguintes ferramentas:

1. Pré-processamento dos dados
2. Seleção de variáveis – Algoritmo OPS
3. Construção do modelo de regressão – Método PLS
4. Detecção de amostras com comportamento atípico – influência e resíduos de Student
5. Validações do modelo – Validação cruzada excluindo-*N*-amostras e teste de aleatorização de **y**.

Pré-processamento dos dados

O pré-tratamento dos dados é um procedimento de rotina na construção dos modelos de QSAR. Quando as variáveis têm diferentes unidades ou quando a faixa de variação dos dados é grande, o que ocorre com frequência nos estudos de QSAR, se recomenda o autoescalamento das variáveis. Com este procedimento, a influência de uma variável dominante é minimizada em cálculos posteriores. O autoescalamento implica em subtrair de cada elemento de uma coluna da matriz de dados o valor médio da respectiva coluna e dividir o resultado pelo desvio padrão da mesma, de acordo com a Equação 1,

$$x_{ij(as)} = \frac{x_{ij} - \overline{x}_j}{s_j} \quad (1)$$

onde x_{ij} e $x_{ij(as)}$ são, respectivamente, os valores da *j*-ésima variável, do *i*-ésimo composto antes e depois do autoescalamento; $\overline{x}_j$ é o valor médio da *j*-ésima variável e s_j é o seu desvio padrão. Este tratamento é aplicado à matriz dos descritores e ao vetor contendo as atividades biológicas.

Em alguns casos, a centragem dos dados na média é utilizada ao invés do autoescalamento e, neste caso, de cada elemento de uma coluna da matriz de dados é subtraído do valor médio da respectiva coluna. O programa QSAR *modeling* oferece estes dois tipos de pré-processamento dos dados, se bem que no geral, os dados são autoescalados.

Construção de modelos de regressão com o método PLS

Os modelos matemáticos usados em QSAR são frequentemente obtidos através de uma regressão linear[1,2,22,23] entre a matriz de descritores e a atividade biológica. Geralmente essa regressão pode ser feita de três maneiras diferentes: regressão linear múltipla (MLR); regressão por componentes principais (PCR); e regressão por quadrados mínimos parciais (PLS).

Historicamente, a regressão multivariada era feita usando-se o método MLR, que sempre funcionou bem porque o número de descritores era menor do que o número de amostras. Atualmente, quando se utiliza esta metodologia em estudos de QSAR, é comum fixar um mínimo de 5 ou 6 compostos para cada descritor e considerar que eles não possuem alta correlação entre si (> 0,7). Entretanto, programas modernos de modelagem usados em estudos de QSAR geram milhares de descritores que frequentemente são altamente correlacionados entre si, especialmente em análises de QSAR 3D e 4D.[24-27] Assim, o método MLR não pode ser usado nesses casos, a menos que se faça uma seleção de variáveis criteriosa. Para evitar esses problemas, uma boa alternativa é o uso dos métodos de projeção, também conhecidos como métodos bilineares, como a regressão por componentes principais (PCR) ou a regressão por quadrados mínimos parciais (PLS).[22,28,29] Quando esses métodos são aplicados, o número de descritores e as correlações entre eles deixam de ser um problema. Entre os métodos PLS e PCR, o primeiro é mais popular em estudos de QSAR e foi o método de regressão escolhido para ser implementado no programa *QSAR modeling*. Embora os métodos PLS e PCR apresentem resultados similares, PLS geralmente produz modelos mais parcimoniosos, com um número menor de fatores e mantendo um bom ajuste.

O número ótimo de variáveis latentes (LV) no modelo é comumente determinado pela validação interna cruzada. Esta metodologia é aplicada, pois os métodos de projeção produzem modelos tendenciosos e é necessário evitar o sobreajuste (*overfitting*). Na validação cruzada, o conjunto de dados é dividido em certa quantidade de grupos (de tamanho *N*) e vários modelos são gerados sempre deixando um desses grupos de fora do modelo. Em seguida, o modelo de regressão obtido é usado para prever a variável dependente (atividade biológica ou propriedade físico-química) das amostras deixadas de fora da análise. Esse processo é repetido até que todas as amostras tenham sido excluídas da análise uma vez. Essa estratégia, chamada de validação cruzada *leave-N-out*, é muito importante para se ter uma ideia inicial a respeito da capacidade preditiva e da robustez do modelo. O uso mais comum dessa estratégia é com o valor de *N* igual a 1, ao se fazer a validação cruzada *leave-one-out*.

O programa *QSAR modeling* oferece, como resultado da validação interna cruzada, tabelas contendo os valores dos parâmetros estatísticos listados na Tabela 2, os coeficientes de regressão do modelo PLS ($b(j)$ para $j = 1, 2, ..., J$), os valores previstos para a variável dependente na validação cruzada ($\hat{y}_{cv}$ para $i = 1, 2, ..., I$) e os valores previstos para a variável dependente no modelo ($\hat{y}_{cal}$ para $i = 1, 2, ..., I$).

O procedimento de validação cruzada disponível no programa *QSAR modeling* permite que o usuário escolha o número máximo de variáveis latentes (LV) e o número de amostras a serem removidas durante o processo de validação cruzada (Figura 2S, material suplementar). A Figura 1 mostra os resultados obtidos para o conjunto de dados usado depois de feita a seleção de variáveis.

A Tabela 3 mostra os resultados da validação cruzada aplicados ao conjunto de dados usado depois da seleção de variáveis usando o *QSAR modeling*. O modelo PLS final foi obtido com oito descritores e três variáveis latentes.

Seleção de variáveis com o algoritmo OPS

O algoritmo de seleção de preditores ordenados (OPS) é um algoritmo desenvolvido recentemente para efetuar a seleção das variáveis[18] e já foi usado com sucesso em estudos de QSAR/QSPR.[24,30-34] A ideia básica desse algoritmo é atribuir importância a cada descritor com

Tabela 2. Parâmetros estatísticos calculados pelo programa *QSAR modeling*

Parâmetro	Símbolo	Equação[a]
Soma dos quadrados dos erros de predição da validação cruzada	*PRESScv*	$\sum_{i=1}^{I}[y(i)-\hat{y}_{cv}(i)]^2$
Soma dos quadrados dos erros de predição da calibração	*PRESScal*	$\sum_{i=1}^{I}[y(i)-\hat{y}_{cal}(i)]^2$
Coeficiente de correlação de Pearson da validação cruzada	r_{cv}	$\frac{\sum_{i=1}^{I}[y(i)-\bar{y}]\times[\hat{y}_{cv}(i)-\hat{\bar{y}}_{cv}]}{\sigma_y\sigma_{\hat{y}_{cv}}}$
Coeficiente de correlação de Pearson da calibração	r_{cal}	$\frac{\sum_{i=1}^{I}[y(i)-\bar{y}]\times[\hat{y}_{cal}(i)-\hat{\bar{y}}_{cal}]}{\sigma_y\sigma_{\hat{y}_{cal}}}$
Coeficiente de correlação da validação cruzada	Q^2	$1-\frac{PRESScv}{\sum_{i=1}^{I}[y(i)-\bar{y}]^2}$
Coeficiente de determinação múltipla	R^2	$1-\frac{PRESScal}{\sum_{i=1}^{I}[y(i)-\bar{y}]^2}$
Raiz quadrada do erro da validação cruzada	*RMSECV*	$\sqrt{\frac{PRESScv}{I}}$
Raíz quadrada do erro da calibração	*RMSEC*	$\sqrt{\frac{PRESScal}{I}}$

[a]*I* é o número de amostras do conjunto de treinamento. $\hat{y}_{cv}(i)$ e $\hat{y}_{cal}(i)$ são valores previstos para $y(i)$ na validação cruzada e no modelo final, respectivamente. $\bar{y}$, $\hat{\bar{y}}_{cv}$ e $\hat{\bar{y}}_{cal}$ são os valores médios de $y(i)$, $\hat{y}_{cv}(i)$ e de $\hat{y}_{cal}(i)$, respectivamente.

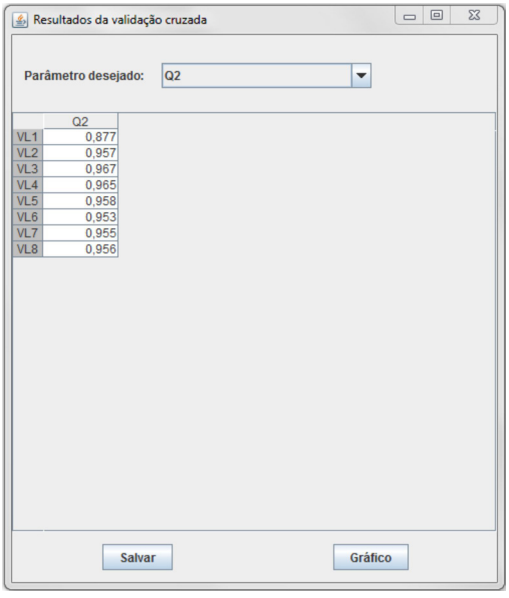

Figura 1. *Janela do programa QSAR modeling na qual os resultados da validação cruzada são mostrados. Todos os parâmetros da Tabela 2, os coeficientes de regressão, os valores previstos para a variável dependente na validação cruzada e os valores previstos para a variável dependente no modelo de regressão podem ser vistos nessa janela*

base em um vetor informativo. As colunas da matriz são rearranjadas de modo que os descritores mais importantes apareçam nas primeiras colunas. Em seguida, são realizadas sucessivas regressões PLS aumentando-se o número de descritores no modelo, com o objetivo de otimizar o modelo PLS. O melhor modelo de regressão pode ser escolhido de acordo com alguns dos parâmetros mostrados na Tabela 2.

O algoritmo OPS está implementado no programa *QSAR modeling* com os seguintes vetores informativos: vetor de correlação; vetor de regressão PLS; e vetor obtido pelo produto elemento a elemento destes dois vetores. A Figura 3S, material suplementar, mostra a janela do programa na qual o usuário escolhe as opções apropriadas para executar o algoritmo OPS.

O usuário tem as seguintes opções para executar o algoritmo OPS (a Figura 3S mostra a tela com essas opções):

- número de variáveis latentes para o algoritmo OPS - número de variáveis latentes do modelo cujo vetor de regressão será usado para ordenar as variáveis. Diferentes números de variáveis latentes produzem vetores de regressão distintos, o que pode afetar a ordenação das variáveis;
- número de variáveis latentes do modelo - número máximo de variáveis latentes dos modelos construídos durante a execução do algoritmo OPS (a ref. 18 traz maiores detalhes);
- número de amostras a serem removidas durante a validação cruzada - valor de *N* no procedimento *leave-N-out*;
- janela - número inicial de descritores na matriz analisada pelo algoritmo OPS;
- incremento - número de descritores que serão adicionados à matriz analisada pelo algoritmo OPS em cada passo;
- porcentagem de variáveis - fração de descritores que serão analisados pelo algoritmo OPS;
- vetor - vetor informativo que será usado para ordenar os descritores;
- critério para classificação do modelo - parâmetro que será usado para avaliar a qualidade do modelo.

Como resultado da seleção de variáveis, o programa *QSAR modeling* mostra uma tabela que lista os melhores modelos obtidos. O programa permite selecionar um dos modelos listados e executar todos os testes de validação disponíveis no programa sobre esse modelo. Além disso, o programa permite salvar a matriz de descritores selecionados para uma análise futura. A Figura 2 mostra a janela do programa *QSAR modeling* que apresenta os resultados obtidos com a aplicação do algoritmo OPS.

Para ilustrar o uso do algoritmo OPS no programa *QSAR modeling*, foi usado um conjunto de dados com 37 compostos e 407 descritores. Dentre os descritores usados nesta análise encontram-se descritores eletrônicos, estéricos, topológicos e eletrotopológicos. Como foram utilizados descritores de diferentes naturezas, o pré-processamento utilizado foi o autoescalamento dos dados. Um corte na correlação, também disponível no programa *QSAR modeling*, foi feito antes da primeira execução do algoritmo OPS. Descritores que apresentaram coeficiente de correlação de Pearson com o vetor de atividades biológicas abaixo de 0,3 foram eliminados, restando 305 descritores. A matriz de dados resultante foi submetida ao algoritmo OPS e o melhor modelo foi obtido com 15 descritores selecionados, 3 variáveis latentes e um valor de Q^2 igual a 0,959. Uma nova tentativa de se obter um melhor modelo foi feita aplicando-se o algoritmo OPS a esta matriz com 15 descritores, o que resultou no modelo final mostrado na Tabela 3 (8 descritores, 3 variáveis latentes e $Q^2 = 0{,}967$).

Detecção de amostras anômalas (*outliers*)

Ao verificar a qualidade do conjunto de treinamento a ser usado para a construção do modelo de regressão, deve-se assegurar que as

Tabela 3. Resultados da validação cruzada obtidos para um modelo com 3 LV após a seleção de variáveis feita com o programa *QSAR modeling*

Parâmetro	*PRESScv*	*PRESScal*	r_{cv}	r_{cal}	Q^2	R^2	*RMSECV*	*RMSEC*
Valor	1,23	0,74	0,98	0,99	0,97	0,98	0,18	0,14

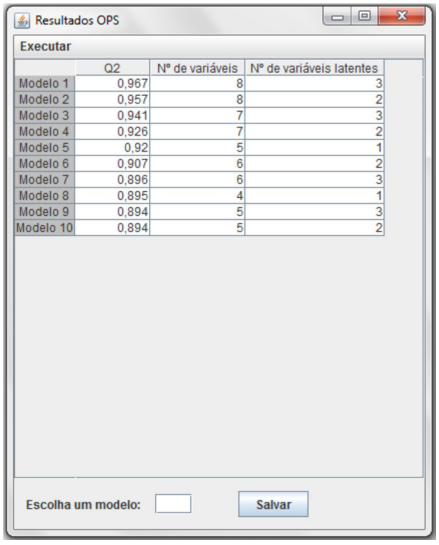

Figura 2. *O valor do parâmetro escolhido para avaliar o modelo, o número de variáveis selecionadas e o número de variáveis latentes dos 10 modelos selecionados são mostrados como resultados do algoritmo OPS*

amostras formam um conjunto homogêneo. Compostos estruturalmente diferentes dos demais ou com valores experimentais atípicos para a atividade biológica podem ter uma influência inadequada no modelo e devem ser removidos do conjunto de treinamento antes da construção do modelo. Um procedimento comum na Quimiometria para se detectar a presença de amostras anômalas (*outliers*) em um conjunto de treinamento é usar os valores de influência (*leverage*) e dos resíduos de Student.[2,22,35] A influência indica exatamente o que o nome diz: a sua capacidade de influenciar na estimativa dos coeficientes de regressão, enquanto que o resíduo de Student é um resíduo (diferença entre o valor experimental da atividade biológica e o valor calculado pelo modelo de regressão) padronizado, obtido dividindo-se o resíduo por uma estimativa de seu próprio desvio padrão. A vantagem de se adotar esta definição para o resíduo é que ele apresenta média igual a zero e desvio padrão igual a um.

A detecção de amostras anômalas feita pelo programa *QSAR modeling* permite que o usuário escolha o número de variáveis latentes que serão usadas pelo modelo PLS e fornece como resultado uma tabela com os valores de influência e do resíduo de Student para cada um dos compostos no conjunto de treinamento (Figura 4S, material suplementar). Amostras com influência maior que $3k/I$, onde k é o número de variáveis latentes e I é o número de amostras, podem ser consideradas suspeitas e devem ser analisadas cuidadosamente, caso a caso.[2,35] Em relação aos resíduos de Student, as amostras devem estar aleatoriamente espalhadas ao redor da origem, indicando que seguem uma distribuição normal. Assumindo-se uma distribuição normal no nível de probabilidade 95% ($\alpha = 0,05$) o valor crítico para um teste bilateral é igual a 1,96, quando os resíduos estão limitados pelo intervalo ± 1,96 (em geral, se usa o intervalo 2,0).

Amostras que apresentam simultaneamente valores de influência e resíduo de Student acima dos limites supraindicados são atípicas e devem ser excluídas do conjunto de dados.

O programa *QSAR modeling* foi usado para verificar a presença de amostras anômalas no modelo depois da seleção de variáveis com o método OPS. A Figura 3 mostra o gráfico de influência *versus* resíduos de Student. A partir dessa figura pode-se observar que não existem amostras que apresentem simultaneamente influência e resíduo de Student acima dos limites aceitos na literatura. No entanto, a partir da Figura 3 pode-se observar que o composto 10 apresenta um alto valor de influência quando comparado ao dos outros compostos, o que o caracteriza como atípico. Outra observação é que os compostos 2 e 23 apresentam um valor de resíduo de Student ligeiramente abaixo do limite inferior. Estes dois últimos devem ser temporariamente excluídos, o modelo refeito e a melhora causada deve ser avaliada; caso ela seja significativa, eles devem ser eliminados e, caso contrário, permanecem no modelo. A amostra 2 tem influência baixa e, portanto, não deve causar alterações no vetor de regressão, o que não ocorre para o composto 23, que tem uma influência mais significativa. Quando os três compostos são removidos, o valor de Q^2 muda de 0,97 para 0,98, melhorando o modelo estatisticamente. Os resíduos altos observados para os compostos 2 e 23 podem ser um indicativo de incerteza nas medidas experimentais. No entanto, a remoção das amostras deve ser realizada cuidadosamente, pois uma explicação química ou biológica deve ser dada para cada amostra classificada como atípica.

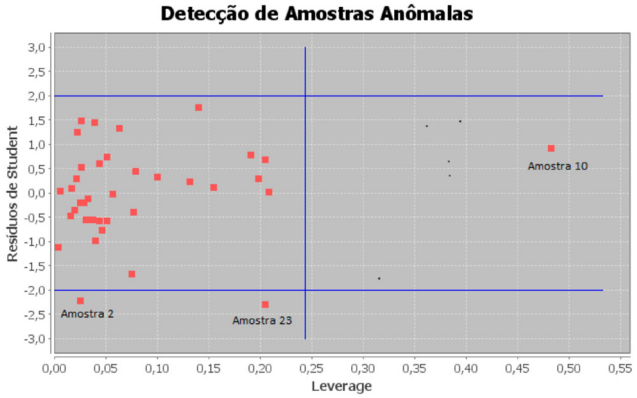

Figura 3. *Gráfico de Leverage versus resíduos de Student para a detecção de amostras anômalas (outliers). As linhas azuis indicam os limites aceitos pela literatura*

Validação cruzada excluindo-*N*-amostras

Se o processo de validação cruzada excluindo-*N*-amostras for repetido inúmeras vezes para diferentes valores de *N*, serão obtidos diferentes valores do coeficiente de correlação de validação cruzada (Q^2) para cada execução. Além disso, mesmo que valores iguais de *N* sejam usados (desde que esse valor não seja igual a 1), diferentes execuções do procedimento *leave-N-out* também levam a valores distintos de Q^2, pois a ordem das amostras na matriz é aleatorizada antes da retirada dos grupos durante o procedimento de validação cruzada.

No entanto, estes valores de Q^2 não deveriam ser muito diferentes entre si. Como o modelo é construído com o objetivo de prever as atividades de novas amostras, não deveria ser sensível às amostras removidas durante a validação cruzada. Assim, para avaliar a robustez do modelo, é altamente recomendável executar repetidos testes da validação cruzada *leave-N-out* para diferentes valores de *N* (variando de 2 até 20 a 30% do número de compostos).[7]

A robustez do modelo é avaliada pelo procedimento *leave-N-out* com o programa *QSAR modeling*. Neste processo é possível escolher

o número máximo de amostras a serem removidas durante a validação cruzada, o número de variáveis latentes, que é mantido fixo durante a validação do modelo, assim como o número de repetições em cada validação para cada número de amostras removidas (Figura 5S, material suplementar). O programa mostra como resultado uma tabela contendo os valores de *RMSECV* ou de Q^2, dependendo da escolha do usuário. Na Figura 4 estão os resultados do teste para um modelo com 3 LV, 3 repetições e um número máximo de 10 amostras removidas para um dos parâmetros estatísticos calculados por este teste.

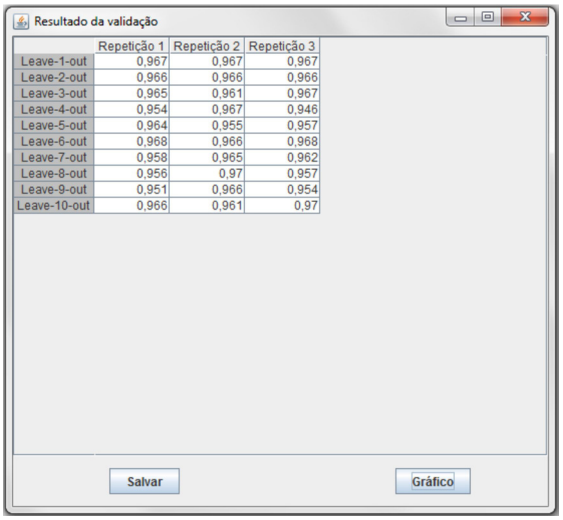

Figura 4. *Resultados obtidos com o procedimento de validação leave-N-out*

O modelo de regressão obtido após a seleção de variáveis com o algoritmo OPS foi submetido ao procedimento de validação *leave-N-out* e os resultados são apresentados na Figura 5. Como pode ser visto, o modelo pode ser considerado robusto, já que pequenas flutuações no valor de Q^2 são observadas com até 10 amostras removidas. Para cada valor de *N* o procedimento foi repetido três vezes (triplicata).

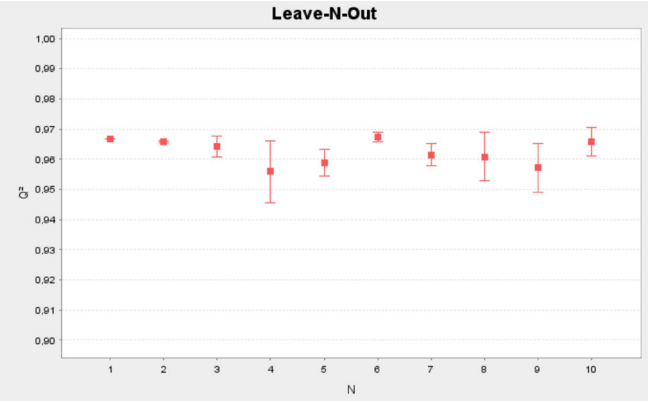

Figura 5. *Validação leave-N-out aplicada ao modelo final obtido depois da seleção de variáveis com o algoritmo OPS. Os pontos representam a média e as barras indicam o desvio padrão de uma triplicata para cada valor de N. O modelo mostrou-se robusto até um valor de N igual a 11 (30% das amostras)*

Teste de aleatorização de y (y-randomization)

A proposta do teste de aleatorização de y é detectar e quantificar correlações ao acaso entre a variável dependente e os descritores.[2,5-7] Para obter uma estimativa da significância de um valor de Q^2 obtido para um dado modelo, deve-se se construir modelos paralelos com os valores de atividade biológica (vetor **y**) permutados, enquanto que os descritores originais (matriz **X**) são mantidos fixos. Espera-se que os modelos paralelos construídos nestas condições sejam de péssima qualidade e com valores de Q^2 bem menores do que o valor obtido para o modelo real, garantindo assim que o modelo real não foi obtido ao acaso.

No processo de aleatorização de **y** usando o programa *QSAR modeling* é possível escolher o número de aleatorizações que serão executadas neste passo da validação (Figura 6S, material suplementar). O programa fornece como resultado uma tabela contendo os valores de R^2 e Q^2, calculados para os modelos obtidos com as atividades biológicas trocadas, e o coeficiente de correlação de Pearson ($r(\mathbf{y}_{al},\mathbf{y})$ entre o vetor **y** com as atividades biológicas corretas e os vetores gerados com as atividades aleatorizadas, $\mathbf{y}_{al}$ (Figura 6). A última linha desta tabela contém os valores de R^2 e Q^2 referentes ao modelo real para que sejam comparados com aqueles obtidos para os modelos paralelos.

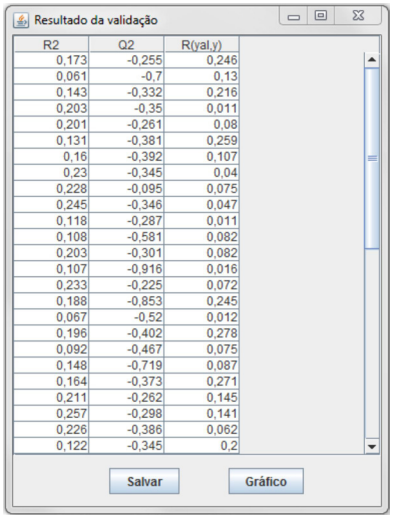

Figura 6. *Resultados do teste de aleatorização de* ***y*** *fornecidos pelo programa QSAR modeling*

O modelo obtido após a seleção de variáveis com o algoritmo OPS foi submetido ao teste de aleatorização de **y**, realizando-se 50 aleatorizações e retirando-se uma amostra (*leave-one-out*) em cada uma delas. Os resultados são apresentados na Figura 7. Como pode ser visto, todos os valores de R^2 *e* Q^2 dos modelos obtidos com o vetor aleatorizado, $\mathbf{y}_{al}$, são menores que 0,4 e 0,0, respectivamente,[2] confirmando que o modelo real não foi obtido ao acaso.

Figura 7. *Valores de R^2 e Q^2 obtidos com o teste de aleatorização de* ***y****. O ponto distante representa os valores de R^2 e Q^2 para o modelo real*

COMPARAÇÃO COM ALGUNS DOS SOFTWARES CITADOS

Com o objetivo de atestar a qualidade do modelo obtido pelo programa *QSAR modeling*, o mesmo conjunto de dados foi utilizado para a construção de modelos com alguns dos softwares citados na Tabela 1 e que também são usados para a construção de modelos QSAR.

O programa VCCLAB,[11] disponível gratuitamente na web, foi utilizado para a construção de um modelo PLS. O programa utiliza PLS como método de regressão e realiza a seleção de variáveis em duas etapas: elimina descritores que são praticamente constantes e seleciona os demais descritores através de algoritmo genético com base nos valores de Q^2 dos modelos gerados. A seleção de variáveis levou a um modelo com 190 descritores e 2 variáveis latentes, com um valor de Q^2 igual a 0,963. Apesar do modelo apresentar um valor de Q^2 próximo ao obtido com o programa *QSAR modeling* (0,967), a interpretação física do modelo é impossível devido ao número excessivo de descritores. Além disso, como o método PLS é tendencioso, a projeção de 190 descritores em apenas 2 variáveis latentes pode levar à perda de informação importante, devido à ocorrência de subajuste. Infelizmente o número de variáveis latentes é selecionado automaticamente pelo programa impedindo, assim, a análise de modelos com outros números de variáveis latentes.

O programa BuildQSAR,[9] disponível para *download* gratuitamente, foi usado para a construção de um modelo MLR. Apesar do programa disponibilizar também o método de regressão PCR, a seleção de variáveis, que pode ser feita através de busca sistemática ou com o algoritmo genético, só pode ser feita utilizando a regressão MLR. No modelo obtido depois da seleção de variáveis feita com o algoritmo genético, foram obtidos 7 descritores, nenhum *outlier* foi detectado e o valor de Q^2 foi de 0,936, também inferior ao valor obtido com o programa *QSAR modeling*. A matriz com estes descritores selecionados foi usada no programa *QSAR modeling* e observou-se que o modelo não passou nos testes de aleatorização de **y** e *leave-N-out*.

O programa Wolf[9] também foi usado para a construção de um modelo MLR. No modelo obtido depois da seleção de variáveis feita com algoritmo genético, foram obtidos 5 descritores e, depois da remoção de uma amostra detectada como anômala (amostra 23), o valor de Q^2 foi de 0,961, também inferior ao valor obtido com o *QSAR modeling*. A matriz com estes descritores selecionados foi usada no programa *QSAR modeling* e observou-se que o modelo não passou no teste de aleatorização de **y**.

CONCLUSÕES

O programa *QSAR modeling* permite a construção de modelos QSAR ou QSPR de uma maneira simples e rápida. Além disso, reúne em um único programa um algoritmo de seleção de variáveis recentemente desenvolvido para construir modelos PLS, um procedimento para detecção de amostras com comportamento anômalo e os principais procedimentos de validação exigidos atualmente pela comunidade científica.

Um conjunto de dados teste foi usado para ilustrar o uso de todas as ferramentas fornecidas pelo programa *QSAR modeling* e os resultados obtidos foram superiores aos obtidos por outros programas, utilizados a título de comparação. Além disso, pôde-se observar que muitas funcionalidades disponíveis no programa *QSAR modeling* não estão disponíveis em outros programas.

Por ser um programa de código aberto, *QSAR modeling* é uma nova ferramenta para estudos de QSAR disponível para qualquer pessoa que desejar usá-la e, assim, pode ser melhorada para necessidades específicas em diversos campos de pesquisa.

O programa se encontra disponível para *download* no site lqta.iqm.unicamp.br.

MATERIAL SUPLEMENTAR

Disponível em http://quimicanova.sbq.org.br, em arquivo pdf, com acesso livre.

AGRADECIMENTOS

À FAPESP e ao CNPq pelo apoio financeiro e à Dra. K. F. M. Pasqualoto do Laboratório de Bioquímica e Biofísica - Instituto Butantan pela ajuda na construção do modelo com o programa Wolf e ao Prof. Dr. E. B. de Melo pela ajuda na construção do modelo com o programa BuildQSAR.

REFERÊNCIAS E NOTAS

1. Deste ponto em diante vamos nos referir à área de QSAR, mas os mesmos procedimentos se aplicam também aos estudos de QSPR; Ferreira, M. M. C.; *J. Braz. Chem. Soc.* **2002**, *13,* 742.
2. Ferreira, M. M. C.; Kiralj, R. Em *Química Medicinal, Métodos e Fundamentos em Planejamento de Fármacos*; Montanari, C., ed.; EDUSP, 2011.3, cap. 12.
3. Guidance Document on the Validation of (Quantitative) OECD Environment Health and Safety Publications Series on Testing and Assessment No. 69. OECD: Paris, 2007, http://search.oecd.org/officialdocuments/displaydocumentpdf/?cote=env/jm/mono(2007)2&doclanguage=en, acessada em Fevereiro 2013.
4. Gramatica, P.; *QSAR & Comb. Sci.* **2007**, *26,* 694.
5. Tropsha, A.; Gramatica, P.; Gombar, V. K.; *QSAR & Comb. Sci.* **2003**, *22*, 69.
6. Eriksson, L.; Jaworska, J.; Worth, A. P.; Cronin, M. T. D.; McDowell, R. M.; Gramatica, P.; *Environmental Health Perspectives* **2003**, *111*, 1361.
7. Kiralj, R.; Ferreira, M. M. C.; *J. Braz. Chem. Soc.* **2009**, *20*, 770.
8. *MobyDigs*, Version 1-2004, Talete srl, Milano, Italy.
9. de Oliveira, D. B.; Gaudio, A. C.; *Quant. Struct.-Act. Relat.* **2000**, *19*, 599.
10. Tetko, I. V.; Gasteiger, J.; Todeschini, R.; Mauri, A.; Livingstone, D.; Ertl, P.; Palyulin, V. A.; Radchenko, E. V.; Zefirov, N. S.; Makarenko, A. S.; Tanchuk, V. Y.; Prokopenko, V. V.; *J. Comput.-Aided Mol. Des.* **2005**, *19*, 453.
11. *VCCLAB*, *Virtual Computational Chemistry Laboratory*, 2005, http://www.vcclab.org, acessada em Fevereiro 2013.
12. *Cerius QSAR+*, 2000, http://www.esi.umontreal.ca/accelrys/pdf/qsarC45.pdf, acessada em Fevereiro 2013.
13. *Bilinear Model*, *BILIN*, 1976, http://www.kubinyi.de/bilin-program.html, acessada em Outubro 2011.
14. *Molecular Structure Generation MOLGEN QSPR*, 2003, http://www.molgen.de/?src=documents/molgenqspr.html, acessada em Fevereiro 2013.
15. *Correlation and Logic*, *CORAL*, 2010, http://www.insilico.eu/coral/, acessada em Fevereiro 2013.
16. *Comprehensive Descriptors for Structural and Statistical Analysis*, *CODESSA PRO*, 2001, http://www.codessa-pro.com/index.htm, acessada em Fevereiro 2013.
17. Rogers, D.; *WOLF Reference Manual Version 5.5*, The Chem21 Group Inc., Chicago, EUA, 1994.
18. Teófilo, R. F.; Martins, J. P. A.; Ferreira, M. M. C.; *J. Chemom.* **2009**, *23*, 32.
19. http://www.moleculardescriptors.eu/dataset/dataset.htm, acessada em Fevereiro 2013.
20 *DRAGON,* Version 6-2010, Talete srl, Milano, Italy.

21. *Java*, version 6 update 10; java development kit; Sun microsystems, Inc: Santa Clara, CA 95054 USA, 2008.
22. Martens, H.; Naes, T.; *Multivariate Calibration*, Wiley: Chichester, 1989.
23. Beebe, K. R.; Pell, R. J.; Seasholtz, M. B.; *Chemometrics: A Practical Guide*, Wiley: Nova York, 1989.
24. Martins, J. P. A.; Barbosa, E. G.; Pasqualoto, K. F. M.; Ferreira, M. M. C.; *J. Chem. Inf. Model.* **2009**, *49*, 1428.
25. Nilsson, J.; de Jong, S.; Smilde, A. K.; *J. Chemom.* **1997**, *11*, 511.
26. Cramer, R. D.; Patterson, D. E.; Bunce, J. D.; *J. Am. Chem. Soc.* **1988**, *110*, 5959.
27. Hopfinger, A. J.; Wang, S.; Tokarski, J. S.; Jin, B.; Albuquerque, M.; Madhav, P. J.; Duraiswami, C.; *J. Am. Chem. Soc.* **1997**, *119*, 10509.
28. Agnar, H.; *J. Chemom.* **1988**, *2*, 211.
29. Geladi, P.; Kowalski, B. R.; *Anal. Chim. Acta* **1986**, *185*, 1.
30. de Melo, E. B.; Ferreira, M. M. C.; *Eur. J. Med. Chem.* **2009**, *44*, 3577.
31. Teófilo, R. F.; Kiralj, R.; Ceragioli, H. J.; Peterlevitz, A. C.; Baranauskas, V.; Kubota, L. T.; Ferreira, M. M. C.; *J. Eletrochem. Soc.* **2008**, *155*, D640.
32. Hernández, N.; Kiralj, R.; Ferreira, M. M. C.; Talavera, I.; *Chemom. Intell. Lab. Syst.* **2009**, *98*, 65.
33. de Melo, E. B.; Ferreira, M. M. C.; *J. Chem. Inf. Model*. **1012**, *52*, 1722.
34. Barbosa, E. G.; Pasqualoto, K. F. M.; Ferreira, M. M. C.; *J. Comput.-Aided Mol. Des.* **2012**, *26*, 1055.
35. Ferreira, M. M. C.; Antunes, A. M.; Melgo, M. S.; Volpe, P. L. O.; *Quim. Nova* **1999**, *22*, 724.