Quim. Nova, Vol. 35, No. 6, 1157-1163, 2012

Artigo

DESENHOS DE ESTRUTURAS QUÍMICAS CORRELACIONAM-SE COM PROPRIEDADES BIOLÓGICAS: MIA-QSAR

Rodrigo A. Cormanich
Instituto de Química, Universidade Estadual de Campinas, CP 6154, 13083-971 Campinas – SP, Brasil
Cleiton A. Nunes
Departamento de Ciência dos Alimentos, Universidade Federal de Lavras, CP 3037, 37200-000 Lavras – MG, Brasil
Matheus P. Freitas*
Departamento de Química, Universidade Federal de Lavras, CP 3037, 37200-000 Lavras – MG, Brasil

Recebido em 20/9/11; aceito em 31/1/12; publicado na web em 11/4/12

CHEMICAL DRAWINGS CORRELATE TO BIOLOGICAL PROPERTIES: MIA-QSAR. Descriptors in multivariate image analysis applied to quantitative structure-activity relationship (MIA-QSAR) are pixels of bidimensional images of chemical structures (drawings), which were used to model the trichomonicidal activities of a series of benzimidazole derivatives. The MIA-QSAR model showed good predictive ability, with r^2, q^2 and $r_{val.\ ext.}^2$ of 0.853, 0.519 and 0.778, respectively, which are comparable to the best values obtained by CoMFA e CoMSIA for the same series. A MIA-based analysis was also performed by using images of alphabetic letters with the corresponding numeric ordering as dependent variables, but no correlation was found, supporting that MIA-QSAR is not arbitrary.

Keywords: multivariate image analysis; QSAR; benzimidazole derivatives.

INTRODUÇÃO

Em 1963, Hansch e Fujita[1] observaram que a atividade biológica de algumas séries de compostos se correlacionava com a lipo-hidrofilicidade das moléculas; a técnica então desenvolvida foi expandida para outras classes de compostos,[2] exibindo correlação igualmente elevada. Esses estudos deram origem à análise quantitativa entre estrutura química e atividade biológica (QSAR - *Quantitative Structure-Activity Relationship*), cujo maior interesse é propiciar o desenvolvimento racional de um novo composto, particularmente um melhor fármaco, evitando síntese aleatória e testes biológicos onerosos de novas moléculas. Um modelo matemático desenvolvido por Free e Wilson[3] também contribuiu para os estudos QSAR dessa época. Atualmente, a maior parte dos modelos QSAR construídos se baseia em descritores (parâmetros que se correlacionam com as atividades biológicas de moléculas) tridimensionais, os quais codificam propriedades moleculares, como efeitos estéricos e eletrostáticos, baseando-se na estrutura espacial de uma classe congênere de moléculas. Os métodos CoMFA (*Comparative Molecular Field Analysis*)[4] e CoMSIA (*Comparative Molecular Similarity Indices Analysis*)[5] destacam-se como métodos QSAR-3D, devido aos inúmeros trabalhos publicados usando essas metodologias. Esses métodos consagrados são os mais amplamente abordados em estudos QSAR-3D e requerem alinhamento tridimensional dos ligantes e, portanto, há a necessidade de similaridade estrutural para sobreposição das estruturas químicas; logo, devem corresponder a uma série congênere. Formalismos 4D,[6] 5D[7] e 6D[8] têm sido aplicados para incorporar novos graus de liberdade (dimensões), de forma que uma análise mais refinada sobre a adaptação do sítio ativo de uma enzima à topologia do ligante, e vice-versa, possa ser mais bem representada. Contudo, descritores moleculares 2D, usualmente descritores físico-químicos referidos em análises QSAR clássicas, não têm se mostrado inferiores aos descritores 3D, sendo extremamente potentes quanto à conveniência e simplicidade dos cálculos.[9] De fato, a necessidade de uma varredura conformacional do ligante e um alinhamento tridimensional exaustivo de estruturas que podem não corresponder às formas bioativas das moléculas, reflete as principais desvantagens das técnicas associadas à metodologia *n*D; portanto, são uma aproximação.

Uma aproximação igualmente preditiva, porém muito mais rápida, barata e simples de operar, foi desenvolvida em 2005 e nomeada MIA-QSAR (*Multivariate Image Analysis applied to QSAR*).[10] Os descritores MIA têm sido aplicados com sucesso não só para correlacionar estruturas químicas com atividades biológicas,[11-16] mas também com propriedades físicas, como temperaturas de ebulição,[17] deslocamentos químicos[18] e perfis eletroforéticos.[19] O método se baseia em utilizar pixels de imagens como descritores; como os pixels podem ser tratados numericamente como binários, a cor branca equivale ao dígito 765 e pixels pretos ao dígito 0, de acordo com o sistema de cores RGB. Em MIA-QSAR, as imagens correspondem a estruturas químicas desenhadas por meio de algum programa para desenho de moléculas, como ChemDraw ou ChemSketch. As modificações estruturais ou mudança na posição dos substituintes em uma série congênere de moléculas correspondem a alterações nas coordenadas dos pixels da imagem, e essas alterações explicam a variância no bloco **Y**, o bloco correspondente às variáveis dependentes (atividades biológicas, por exemplo).

Não é raro alguns pareceres de manuscritos e comentários de bancas examinadoras demonstrarem certo ceticismo sobre a existência de significado físico-químico para os descritores MIA e, portanto, sobre os mesmos poderem se correlacionar com alguma propriedade química, física ou biológica. Alguns, inclusive, relacionaram os resultados de uma análise MIA-QSAR à correlação por acaso que poderia existir entre as notas dos estudantes em uma prova com a ordem alfabética de seus nomes, o que seria uma hipótese completamente arbitrária. Reforçamos a afirmação de que descritores MIA podem codificar propriedades químicas, físicas e biológicas; a descrição físico-química deve estar toda incorporada na maneira com que substituintes são representados. Por exemplo, os descritores MIA podem codificar efeitos estéricos (substituintes de moléculas orgânicas ocupando uma grande área no espaço dedicado ao desenho

*e-mail: matheus@dqi.ufla.br

das estruturas), centros estereogênicos (linhas em cunha ou tracejadas, para representar ligações para frente ou para trás relativas a um carbono quiral) etc. Para comprovar isso, o presente trabalho apresenta uma análise MIA-QSAR e outra, baseada em descritores MIA, em que letras são correlacionadas com sua ordem numérica no alfabeto, isto é, a letra A corresponde ao número 1, a letra B ao 2, a letra C ao 3, e assim sucessivamente. Os pixels das letras *Times New Roman* tamanho 48 (ajustadas às margens superior e esquerda de um espaço de trabalho de tamanho 60×60 pixels do aplicativo Paint do Microsoft Windows) são as variáveis independentes, enquanto os números correspondentes são as variáveis dependentes (bloco **Y**). A correlação das letras do alfabeto com a numeração de 1 a 26 é arbitrária e, portanto, não se espera ajuste razoável, ao contrário do que propõe a análise MIA-QSAR. A análise MIA-QSAR consistiu em correlacionar as estruturas químicas de uma série de derivados benzimidazólicos com suas respectivas atividades tricomonicidas.

PARTE EXPERIMENTAL

O primeiro passo para se construir um modelo MIA-QSAR é escolher um conjunto de dados em que moléculas com determinada propriedade biológica pertençam a uma série congênere; é necessário um mínimo de similaridade entre as estruturas na análise MIA-QSAR, pois o método envolve um alinhamento bidimensional, conforme descrito mais adiante. No presente estudo de caso, uma série com 70 derivados benzimidazólicos com atividade tricomonicida foi obtida da literatura (Tabela 1).[20]

As estruturas químicas foram desenhadas sistematicamente utilizando o programa ChemSketch;[21] tem sido mostrado que pequenas diferenças na maneira de representar um determinado substituinte na molécula (por exemplo, CH_3 ou Me para representar um grupo metila) não afeta estatisticamente o modelo, desde que todos sejam representados da mesma forma para todas as moléculas nas respectivas posições.[22] Cada estrutura química foi transferida para uma área de trabalho do aplicativo Paint do Microsoft Windows e cada imagem foi salva como *bitmaps* (.bmp); é importante que cada estrutura seja salva numa área de trabalho de tamanho definido (no caso, a dimensão da área de trabalho foi 470×265 pixels) e que cada imagem seja movida de tal forma que um determinado pixel, comum a todas as estruturas químicas da série, seja fixado numa determinada coordenada da área de trabalho (no caso, um pixel localizado no carbono ligado ao substituinte R_2 foi fixado na coordenada 200×150 pixels). Esse último passo corresponde ao alinhamento 2D e é feito manualmente (com o auxílio do mouse); é uma etapa fundamental na análise, pois cada imagem (um plano bidimensional) será sobreposta à outra, formando um arranjo tridimensional de tamanho 70×470×265, em que as partes comuns entre as estruturas da série congênere (o esqueleto básico) estejam congruentes. Portanto, a porção variável das moléculas são os substituintes e a orientação de seus pixels explica a variância no bloco **Y** (as atividades biológicas). O arranjo tridimensional pode ser desdobrado para um arranjo bidimensional de tamanho 70×124550, o que permite a regressão dessa matriz com o bloco **Y** por meio de mínimos quadrados parciais (PLS bilinear). Para agilizar os cálculos, as colunas com variância zero (por exemplo, os espaços em branco comuns para todas as imagens ou as partes congruentes das estruturas químicas) foram removidas, dando origem a uma matriz **X** de tamanho 70×2854. A Figura 1 ilustra o procedimento para tratamento e análise das imagens, e os *scripts* a seguir mostram como as imagens podem ser carregadas e convertidas em binários, utilizando o programa Matlab.[23]

```
[filename,MAP]=imread('filename.bmp','bmp');
filename=double(filename);
filename=(filename(:,:,1)+filename(:,:,2)+filename(:,:,3));
```

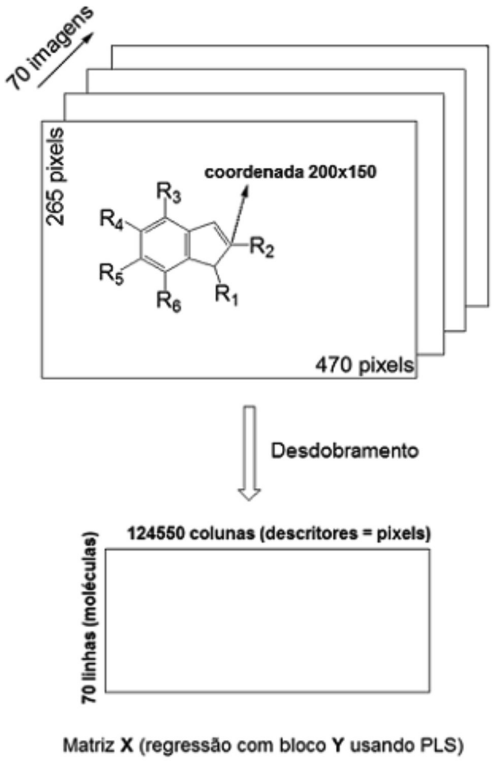


Figura 1. *Construção do arranjo tridimensional e desdobramento para a matriz* ***X***

Segundo análise dos resíduos de Student, as amostras **28** e **31** foram diagnosticadas como *outliers* (provavelmente, por serem as únicas estruturas com R_6 diferente de H), similarmente ao encontrado na literatura de origem,[20] sendo removidas do modelo. As 68 amostras restantes foram divididas em grupos de treinamento e teste, da mesma maneira como descrito na literatura,[20] que utiliza os métodos CoMFA e CoMSIA. Assim, as mesmas amostras do grupo de treinamento dedicado à calibração do modelo (55 moléculas) e do grupo teste à validação externa (13 moléculas), que foram utilizadas na literatura de referência, foram também utilizadas no presente trabalho. É importante que as amostras do grupo treino e teste deste estudo sejam as mesmas das do artigo de referência, para que seja possível a comparação dos resultados obtidos pelo método MIA-QSAR com os obtidos por CoMSIA e CoMFA - os métodos mais amplamente utilizados em análises QSAR-3D.

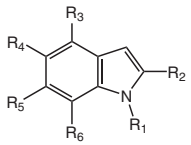
RESULTADOS E DISCUSSÃO

Uma vez construídos a matriz **X** e o bloco **Y**, avaliou-se o número ótimo de variáveis latentes (número de componentes PLS) a ser utilizado no modelo, verificando o menor valor da raiz do erro quadrático médio de validação cruzada *leave-one-out* (RMSECV) em função do número de variáveis latentes do modelo (Figura 2).

Subsequentemente, o modelo com 5 variáveis latentes forneceu valores de r^2 e q^2 de 0,853 e 0,519 para a calibração e validação cruzada LOO, respectivamente. r^2 e q^2 são 1 - $[(\Sigma(y_i-\hat{y}_i)^2/\Sigma(y_i-\bar{y})^2]$, em que y_i são os valores de bioatividade experimentais, $\hat{y}_i$ são os valores de bioatividade estimados/preditos, e $\bar{y}$ são os valores de bioatividade médios. Os valores ajustados e preditos são apresentados na Tabela 2 e suas distribuições gráficas, na Figura 3.

O modelo se mantém preditivo mesmo utilizando-se validação cruzada *leave*-25%-*out* (em que 25% das amostras foram aleatoriamente separadas do conjunto de calibração), cujo valor de q^2 foi de 0,626. O senso comum em QSAR estabelece que valores de r^2

Tabela 1. Derivados benzimidazólicos utilizados na análise MIA-QSAR



Composto	R_1	R_2	R_3	R_4	R_5	R_6
1	H	CF_3	H	H	H	H
2	H	CF_3	H	Cl	H	H
3	H	CF_3	H	F	H	H
4	H	CF_3	H	CF_3	H	H
5	H	CF_3	H	CN	H	H
6	CH_3	CF_3	H	CF_3	H	H
7	CH_3	CF_3	H	H	CF_3	H
8	H	CF_3	H	$SCH_2CH_2CH_3$	H	H
9	CH_3	CF_3	H	$SCH_2CH_2CH_3$	H	H
10	CH_3	CF_3	H	H	$SCH_2CH_2CH_3$	H
11	H	CF_3	H	COC_6H_5	H	H
12	CH_3	CF_3	H	COC_6H_5	H	H
13	CH_3	CF_3	H	H	COC_6H_5	H
14	H	H	H	H	H	H
15	H	CH_3	H	H	H	H
16	H	NH_2	H	H	H	H
17	H	SH	H	H	H	H
18	H	H	H	Cl	H	H
19	H	CH_3	H	Cl	H	H
20	H	NH_2	H	Cl	H	H
21	H	SH	H	Cl	H	H
22	H	SCH_3	H	Cl	H	H
23	H	NH_2	H	Cl	Cl	H
24	H	CF_3	H	Br	H	H
25	H	CF_3	H	Br	Br	H
26	H	CF_3	Br	H	Br	H
27	H	CF_3	Br	Br	Br	H
28	H	CF_3	Br	Br	Br	Br
29	H	C_2F_5	H	Cl	Cl	H
30	H	CF_3	H	NO_2	NO_2	H
31	H	C_2F_5	Br	Br	Br	Br
32	H	SCH_2CH_2OH	Cl	H	Cl	H
33	H	SCH_2CH_2OH	Br	H	Br	H
34	H	$SCH_2CH_2N(CH_3)_2$	Cl	H	Cl	H
35	H	$SCH_2CH_2N(CH_3)_2$	Br	H	Br	H
36	H	$SCH_2CH_2N(C_2H_5)_2$	Cl	H	Cl	H
37	H	$SCH_2CH_2N(C_2H_5)_2$	Br	H	Br	H
38	H	$SCH_2CH_2CH_2N(CH_3)_2$	Cl	H	Cl	H
39	H	$SCH_2CH_2CH_2N(CH_3)_2$	Br	H	Br	H
40	H	SCH_2CH_2-(*N*-piperidil)	Cl	H	Cl	H
41	H	SCH_2CH_2-(*N*-piperidil)	Br	H	Br	H
42	H	SCH_2CH_2-(*N*-morfolinil)	Cl	H	Cl	H
43	H	SCH_2CH_2-(*N*-morfolinil)	Br	H	Br	H
44	H	SCH_2CH_2-(*p*-nitrofenil)	Cl	H	Cl	H
45	H	SCH_2CH_2-(*p*-nitrofenil)	Br	H	Br	H
46	CH_3	$CONH_2$	H	H	Cl	H

Tabela 1. continuação

Composto	R_1	R_2	R_3	R_4	R_5	R_6
47	CH_3	$CONHCH_3$	H	H	Cl	H
48	CH_3	$CON(CH_3)_2$	H	H	Cl	H
49	CH_3	$COOCH_2CH_3$	H	H	Cl	H
50	CH_3	$CONH_2$	H	Cl	H	H
51	CH_3	$CONHCH_3$	H	Cl	H	H
52	CH_3	$CON(CH_3)_2$	H	Cl	H	H
53	CH_3	$COOCH_2CH_3$	H	Cl	H	H
54	CH_3	$CONH_2$	H	Cl	Cl	H
55	CH_3	$CONHCH_3$	H	Cl	Cl	H
56	CH_3	$CON(CH_3)_2$	H	Cl	Cl	H
57	CH_3	$COOCH_2CH_3$	H	Cl	Cl	H
58	CH_3	$CONH_2$	H	H	H	H
59	CH_3	$CONHCH_3$	H	H	H	H
60	CH_3	$CON(CH_3)_2$	H	H	H	H
61	CH_3	$COOCH_2CH_3$	H	H	H	H
62	CH_3	$NHCOOCH_3$	H	H	H	H
63	CH_3	$NHCOOCH_3$	H	Cl	H	H
64	CH_3	$NHCOOCH_3$	H	H	Cl	H
65	CH_3	$NHCOOCH_3$	H	Cl	Cl	H
66	H	SCH_3	H	$CONHCH_3$	H	H
67	H	SCH_3	H	$CON(CH_3)_2$	H	H
68	H	SCH_3	H	$CONHCH_2CH_3$	H	H
69	H	SCH_3	H	$CON(CH_2CH_3)_2$	H	H
70	H	SCH_3	H	$CONHCH_3$	H	H

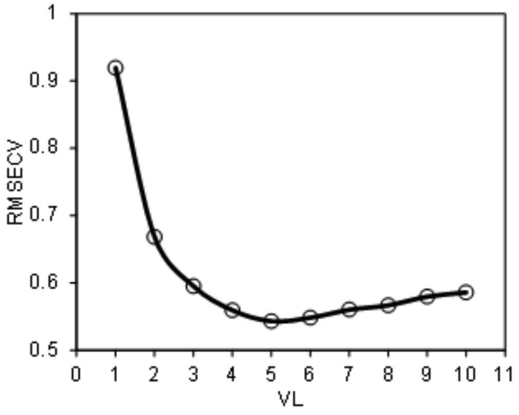


Figura 2. *Gráfico de números de variáveis latentes (VL) do modelo MIA-QSAR vs. RMSECV*

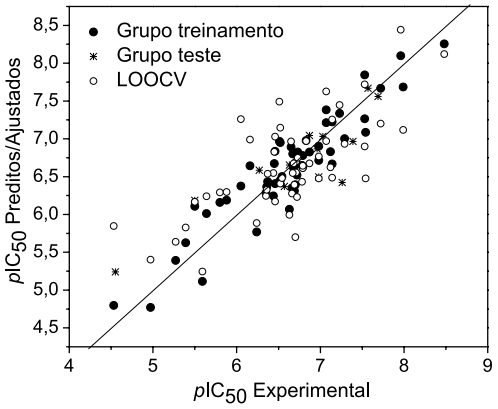


Figura 3. *Gráfico de valores de pIC_{50} experimentais em relação aos preditos/ajustados, obtidos pelo método MIA-QSAR para a série de derivados benzimidazólicos com atividade tricomonicida*

acima de 0,8 e q^2 acima de 0,5 correspondem a modelos preditivos; portanto, o modelo MIA-QSAR obedece a esses parâmetros. Contudo, Golbraikh e Tropsha[24] afirmaram que a única maneira de estabelecer um modelo QSAR confiável é por meio de validação externa, a qual foi realizada no presente trabalho, sendo obtido um valor de $r^2_{\text{val. ext.}}$ para esse teste de 0,778 (valores acima de 0,5 correspondem a modelos QSAR preditivos). Para atestar a robustez do modelo e comprovar que a boa correlação obtida não foi obra do acaso, o bloco **Y** foi aleatorizado (para ambos os grupos de treinamento e teste) e a calibração PLS com 5 variáveis latentes, bem como predição com os parâmetros de regressão obtidos forneceram os valores de r^2 e $r^2_{\text{val. ext.}}$ (média de 10 repetições) de 0,024 ± 0,006 e 0,013 ± 0,006, respectivamente. Esses resultados comprovam que o modelo MIA-QSAR real é preditivo e confiável; os dados estatísticos são sumarizados na Tabela 3.

O mesmo conjunto de dados fora analisado por meio das metodologias de QSAR-3D CoMFA e CoMSIA.[20] Uma variedade de modelos foi criada, dependendo das conformações escolhidas para o alinhamento 3D, dos descritores gerados e do método utilizado para computar as cargas atômicas, dando origem a resultados de correlação bastante variados (Tabela 3). Isso sugere que há uma grande dependência da qualidade do modelo 3D com o procedimento e a escolha de parâmetros a serem utilizados na modelagem. O modelo MIA-QSAR, por outro lado, forneceu resultados de estimativa e predição comparáveis aos melhores modelos CoMFA e CoMSIA obtidos da literatura.[20] Ambas metodologias 3D identificaram os compostos **28** e **31** como *outliers*, excluindo-os das análises; notório é o fato desses *outliers* também terem sido detectados usando os descritores MIA (Figura 4), sugerindo que MIA-QSAR e métodos 3D estejam descrevendo o conjunto de dados estudado da mesma maneira e que, portanto, descritores MIA codificam informação química.

O modelo MIA-QSAR construído pode ser usado para predizer a atividade trocomonicida de novos compostos congêneres da série de derivados benzimidazólicos. Um modo de propor novas estruturas

Tabela 2. Valores de pIC_{50} experimentais, calibrados e preditos por MIA-QSAR para os derivados benzimidazólicos com atividade tricomonicida

Composto	$pIC_{50(Exp)}$	$pIC_{50(cal/pred)}$	$pIC_{50\ (LOO\ CV)}$	Composto	$pIC_{50(Exp)}$	$pIC_{50(cal/pred)}$	$pIC_{50\ (LOO\ CV)}$
1	5,50	5,50	6,17	**36**[a]	7,39	7,29	
2	6,35	6,35	6,25	**37**	7,53	7,23	6,90
3[a]	5,50	6,63		**38**	7,07	7,96	7,63
4	6,63	5,64	6,00	**39**	7,99	8,48	7,12
5	5,64	5,39	6,24	**40**	7,14	6,98	7,26
6	5,39	5,27	5,83	**41**[a]	7,69	7,72	
7	5,27	6,46	5,64	**42**	7,29	6,73	6,93
8	6,46	6,70	7,03	**43**	7,23	6,45	7,45
9	6,70	5,59	5,70	**44**	7,96	6,68	8,44
10	5,59	4,53	5,24	**45**	8,48	6,65	8,12
11[a]	4,55	4,97		**46**[a]	6,96	7,12	
12	4,53	6,44	5,85	**47**	6,98	7,53	6,48
13	4,97	6,52	5,40	**48**[a]	6,63	6,78	
14	6,44	6,54	6,55	**49**	7,72	6,37	7,20
15	6,52	6,71	6,41	**50**	6,73	7,07	6,75
16	6,54	6,69	6,48	**51**	6,45	5,88	6,83
17	6,71	6,79	6,55	**52**	6,68	6,36	6,68
18	6,69	6,71	6,39	**53**[a]	7,57	6,46	
19	6,79	6,87	6,61	**54**[a]	6,87	6,67	
20	6,71	6,98	6,70	**55**	6,65	6,79	6,97
21	6,87	5,80	6,68	**56**	7,12	7,54	6,63
22[a]	7,03	6,66		**57**	7,53	6,05	7,72
23	6,98	6,72	6,77	**58**	6,78	5,50	6,43
24	5,80	6,52	6,29	**59**[a]	6,98	4,55	
25	6,66	6,24	6,28	**60**	6,37	7,03	6,54
26	6,72	6,16	6,23	**61**	7,07	6,57	6,97
27[a]	6,57	6,83		**62**	5,88	7,39	6,30
28[b]	8,70			**63**	6,36	7,69	6,33
29	6,52	7,14	7,15	**64**	6,46	6,96	6,17
30	6,24	6,51	5,89	**65**[a]	6,27	6,63	
31[b]	5,00			**66**	6,67	7,57	6,55
32	6,16	7,53	6,99	**67**	6,79	6,87	6,64
33	6,83	7,07	6,97	**68**	7,54	6,98	6,48
34	7,14	7,99	6,49	**69**[a]	7,26	6,27	
35	6,51	7,14	7,49	**70**	6,05	7,26	7,26

[a] Validação externa. [b] *Outlier*.

Tabela 3. Dados estatísticos das análises MIA-QSAR e modelagem baseada em descritores MIA com as letras do alfabeto, e dados da literatura (de 2 a 5 variáveis latentes)

Conjunto	MIA-QSAR		MIA-alfabeto[b]		CoMFA[20]	CoMSIA[20]
	r^2	RMSE	r^2	RMSE	r^2	r^2
Calibração	0,853	0,281	0,976	1,20	0,927-0,936	0,573-0,915
LOOCV	0,519	0,543	0,137	7,60	0,601-0,634	0,483-0,642
L-25%-Out	0,626	0,718				
Val. Ext.	0,778	0,524	0,115	6,69	0,729-0,890	0,562-0,873
Y-random. (cal.)[a]	0,024 ± 0,006		0,970	1,30		
Y-random. (val. ext.)[a]	0,013 ± 0,04					

[a] Média de 10 repetições. [b] Para 4 variáveis latentes.

potencialmente ativas é desenhar novas moléculas, utilizando os mesmos parâmetros e regras de alinhamento 2D aplicados na construção do modelo, que sejam miscelâneas de subestruturas de duas ou mais moléculas altamente ativas da série congênere. Na sequência, os parâmetros de regressão PLS podem ser utilizados para predizer as atividades biológicas das estruturas propostas. Essa estratégia tem sido aplicada com sucesso e os resultados confirmados por técnicas de *docking*;[25,26] a predição de parâmetros ADME-Tox (absorção, distribuição, metabolismo, excreção e toxicidade) pode auxiliar na modelagem de um fármaco mais seguro, além de mais ativo.

Segundo observação dos dados biológicos, os compostos mais ativos da série (excluindo-se os *outliers*) são o **39** e o **45**, enquanto os menos ativos são o **11** e o **12**. Os substituintes que diferenciam os dois compostos mais ativos da série dos dois menos ativos são R_2, R_3, R_4 e R_5, os quais devem explicar a tendência nos valores de pIC_{50}. Das 14 amostras contendo enxofre e um aceptor de prótons

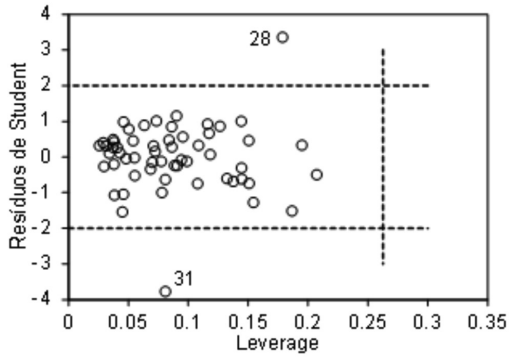


Figura 4. *Detecção de outliers (**28** e **31**) utilizando gráfico de leverage vs. resíduos de Student*

(O ou N) em R_2 (como **39** e **45**), 11 possuem pIC_{50} acima de 7. Por outro lado, das 18 estruturas contendo o grupo eletronegativo CF_3 na posição R_2 (como **11** e **12**), todas apresentam pIC_{50} abaixo de 7 e, dessas, 10 apresentam pIC_{50} inferior a 6. Ainda, dentre os compostos acima com maior atividade biológica, todos possuem halogênios em R_3 e R_5 e hidrogênio em R_4. A maioria dos compostos menos ativos mencionados acima possui hidrogênio na posição R_3 e grupos volumosos em R_4. Portanto, como perfil adequado para um novo tricomonicida, sugere-se a estrutura básica apresentada na Tabela 1, com R_2 hidrofóbico (contendo enxofre e um aceptor de prótons), R_3 e R_5 hidrofóbicos (halogênio – Cl ou Br) e R_4 pouco volumosos (hidrogênio), provavelmente devido à repulsão estérica com o sítio ativo.

Apesar de algumas modelagens MIA-QSAR já terem sido realizadas com sucesso e validadas por meio dos métodos de validação mais rigorosos, por vezes se depara com certo ceticismo a respeito dos descritores MIA de fato codificarem informação química e/ou biológica. Para demonstrar que os bons resultados obtidos pelo método MIA-QSAR não são fortuitos, procurou-se construir um modelo baseado em descritores MIA que, de fato, é arbitrário, e se compararam seus resultados com os da análise MIA-QSAR. No modelo arbitrário, as letras do alfabeto (26 letras em fonte *Times New Roman* tamanho 48, de A a Z) foram as imagens, enquanto as variáveis dependentes corresponderam à sua ordem numérica no alfabeto. O conjunto foi dividido em grupos de treinamento e teste (1/3 das letras do alfabeto: C, F, I, L, O, R, U, X) e os resultados indicam que o modelo construído não é preditivo, conforme esperado para um modelo arbitrário. Inicialmente, o número de variáveis latentes escolhido foi 2, baseado no mínimo RMSECV (Figura 5).

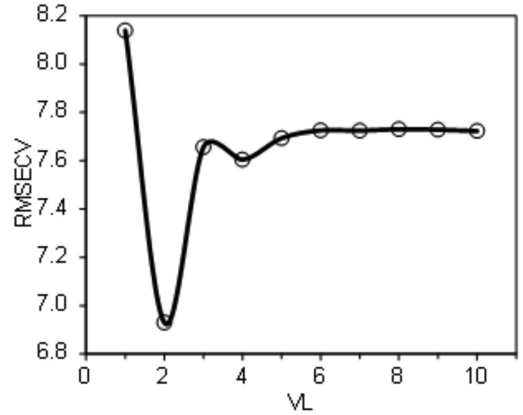


Figura 5. *Gráfico de números de variáveis latentes vs. RMSECV para o modelo baseado em descritores MIA com as letras do alfabeto*

Para esse modelo, os valores de r^2, q^2 e $r^2_{val.\ ext.}$ foram insuficientes para considerar o modelo ao menos razoável; os resultados de validação, inclusive, sugerem que o poder de predição do modelo é desprezível (Tabela 3 e Figura 6a). Ao utilizar um número maior de variáveis latentes (4), visto que o erro de calibração era grande e ainda bastante descendente após 2 variáveis latentes, atingiu-se um valor de r^2 elevado (0,976). Contudo, o modelo continuou nada preditivo (Figura 6b) e, ainda, o teste de randomização do bloco **Y** mostrou um r^2 igualmente elevado (e RMSE igualmente baixo, Figura 7), indicando que, usando 4 variáveis latentes, o modelo é superajustado (*over-fitting*).

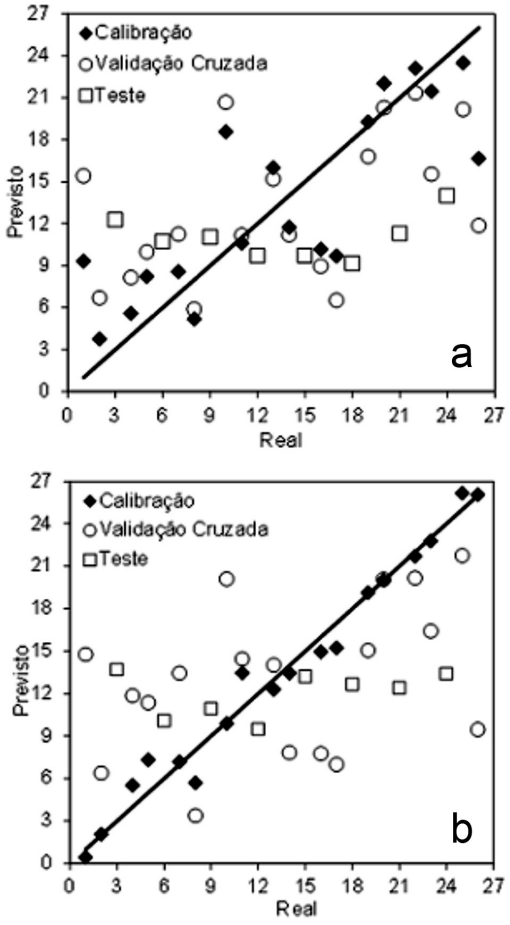


Figura 6. *Gráficos de correlações para os modelos baseados em descritores MIA com as letras do alfabeto. a) Modelo com 2 variáveis latentes e b) modelo com 4 variáveis latentes*

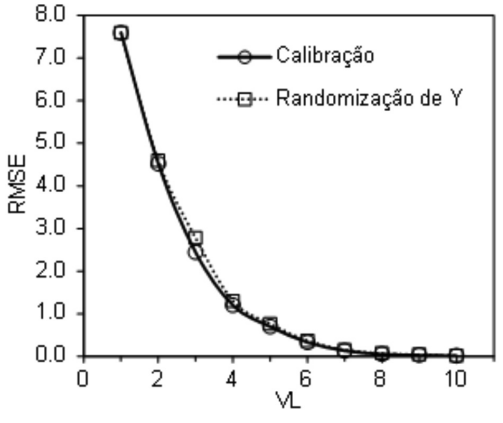


Figura 7. *Gráfico de números de variáveis latentes vs. RMSEC (calibração real e com o bloco Y randomizado) para o modelo baseado em descritores MIA com as letras do alfabeto*

CONCLUSÕES

O método MIA-QSAR pode ser uma ferramenta útil para predizer a atividade biológica de compostos congêneres de uma determinada classe de substâncias bioativas. Para a série de derivados benzimidazólicos com atividade tricomonicida apresentados neste estudo, os resultados estatísticos foram comparáveis aos melhores modelos CoMFA e CoMSIA construídos para essa classe de compostos. Entretanto, o investimento computacional necessário para uma análise MIA-QSAR é modesto, a manipulação dos dados é simples e rápida, e não há necessidade de varredura conformacional e alinhamento tridimensional das moléculas (cuja escolha dos confôrmeros e das regras de alinhamento não deixam de ser arbitrárias). Portanto, além de útil para fins de pesquisa, é perfeitamente praticável em nível de graduação, por exemplo, como aplicação em disciplinas optativas de Química Medicinal, Química Computacional e Quimiometria. Modelos certamente arbitrários comportam-se como tal em uma análise baseada em descritores MIA, mas em QSAR, é possível que os pixels de imagens de estruturas químicas se correlacionem com atividades biológicas.

AGRADECIMENTOS

Ao apoio financeiro na FAPEMIG, CNPq e FAPESP.

REFERÊNCIAS

1. Hansch, C.; Muir, R. M.; Fujita, T.; Maloney, P. P.; Geiger, C. F.; Streich, M. J.; *J. Am. Chem. Soc.* **1963**, *85*, 2817.
2. Hansch, C.; Fujita, T.; *J. Am. Chem. Soc.* **1964**, *86*, 1616.
3. Free, S. M.; Wilson, J. W.; *J. Med. Chem.* **1964**, *7*, 395.
4. Cramer, R. D.; Patterson, D. E.; Bunce, J. D.; *J. Am. Chem. Soc.* **1988**, *110*, 5959.
5. Klebe, G.; Abraham, U.; Mietzner, T.; *J. Med. Chem.* **1994**, *37*, 4130.
6. Hopfinger, A. J.; Wang, S.; Tokarski, J. S.; Jin, B.; Albuquerque, M.; Madhav, P. J.; Duraiswami, C.; *J. Am. Chem. Soc.* **1997**, *119*, 10509.
7. Vedani, A.; Dobler, M.; *J. Med. Chem.* **2002**, *45*, 2139.
8. Vedani, A.; Dobler, M.; Lill, M. A.; *J. Med. Chem.* **2005**, *48*, 3700.
9. Tian, F.; Zhou, P.; Li, Z.; *J. Mol. Struct.* **200**, *871*, **7**140.
10. Freitas, M. P.; Brown, S. D.; Martins, J. A.; *J. Mol. Struct.* **2005**, *738*, 149.
11. Freitas, M. P.; *Org. Biomol. Chem.* **2006**, *4*, 1154.
12. Freitas, M. P.; *Curr. Comput.-Aid. Drug Des.* **2007**, *3*, 235.
13. Freitas, M. P.; *Chemom. Intell. Lab. Sys.* **2008**, *91*, 173.
14. Goodarzi, M.; Freitas, M. P.; *Chemom. Intell. Lab. Sys.* **2009**, *96*, 59.
15. Goodarzi, M.; Freitas, M. P.; *Mol. Simul.* **2010**, *36*, 267.
16. Cormanich, R. A.; Freitas, M. P.; Rittner, R.; *J. Braz. Chem. Soc.* **2011**, *22*, 637.
17. Goodarzi, M.; Freitas, M. P.; *J. Phys. Chem. A* **2008**, *112*, 11263.
18. Goodarzi, M.; Freitas, M. P. Ramalho, T. C.; *Spectrochim. Acta, Part A* **2009**, *74*, 563.
19. Goodarzi, M.; Freitas, M. P.; *Separ. Purif. Technol.* **2009**, *68*, 363.
20. Pérez-Villanueva, J.; Medina-Franco, J. L.; Caulfield, T. R.; Hernández-Campos, A.; Hernández-Luis, F.; Yépes-Mulia, L.; Castillo, R.; *Eur. J. Med. Chem.* **2011**, *46*, 3499.
21. *ACD/ChemSketch Version 12.01*, Advanced Chemistry Development, Inc., Toronto, Canada, 2009.
22. Goodarzi, M.; Freitas, M. P.; Ferreira, E. B.; *QSAR Comb. Sci.* **2009**, *28*, 458.
23. *MatLab, Version 7.5,* MathWorks Inc., Natick, MA, 2005.
24. Golbraikh, A.; Tropsha, A.; *J. Mol. Graphics Modell.* **2002**, *20*, 269.
25. Pinheiro, J. R.; Bitencourt, M.; da Cunha, E. F. F.; Ramalho, T. C.; Freitas, M. P.; *Bioorg. Med. Chem.* **2008**, *16*, 1683.
26. Antunes, J. E.; Freitas, M. P.; da Cunha, E. F. F.; Ramalho, T. C.; Rittner, R.; *Bioorg. Med. Chem.* **2008**, *16*, 7599.