Quim. Nova, Vol. 25, No. 6B, 1061-1066, 2002

Artigo

ESTUDO QSPR SOBRE OS COEFICIENTES DE PARTIÇÃO: DESCRITORES MECÂNICO-QUÂNTICOS E ANÁLISE MULTIVARIADA

Edilson Grünheidt Borges* e Yuji Takahata
Instituto de Química, Universidade Estadual de Campinas, CP 6154, 13083-970 Campinas - SP

Recebido em 20/6/01; aceito em 17/6/01

QSPR STUDY ON PARTITION COEFFICIENTS: QUANTUM-MECHANICAL DESCRIPTORS AND MULTIVARIATE ANALYSIS. Quantum chemistry and multivariate analysis were used to estimate the partition coefficients between n-octanol and water for a serie of 188 compounds, with the values of the q 2 until 0.86 for crossvalidation test. The quantum-mechanical descriptors are obtained with ab initio calculation, using the solvation effects of the Polarizable Continuum Method. Two different Hartree-Fock bases were used, and two different ways for simulating solvent cavity formation. The results for each of the cases were analised, and each methodology proposed is indicated for particular case.

Keywords: partial least squares; aquatic polutants; log *P*.

INTRODUÇÃO

Os compostos com toxidade para peixes têm sua atividade muito ligada à solubilidade em fase aquosa. A classificação em quatro níveis de toxidade destes poluentes com mecanismo de ação semelhante leva este fator em consideração, diretamente, para os compostos de menor toxidade. As classes propostas são as seguintes: Classe 1- Compostos de toxidade mais baixa (também chamados de toxidade basal) são geralmente compostos apolares. Classe 2- Os compostos de toxidade intermediária (também chamados de toxidade aguda) são geralmente compostos polares. Classe 3- Os compostos de toxidade mais alta são classificados como quimicamente reativos. Classe 4- Os compostos como os pesticidas e defensivos agrícolas, de ação específica[1].

Os métodos de classificação tradicionalmente levam em conta a presença ou ausência de grupos funcionais. Esta classificação fica restrita aos compostos que se enquadram nas regras estabelecidas. Contudo, levando-se em conta apenas os coeficientes de partição e mais alguns descritores mecânico-quânticos, Hermens e colaboradores[2] mostraram que se pode realizar uma boa classificação para os poluentes das classes de toxidade 1 e 2.

O trabalho aqui proposto mostra que mesmo o coeficiente de partição (logP) pode ser obtido por modelagem molecular, com os descritores mecânico-quânticos usados por Hermens e mais alguns. Os valores de logP foram modelados com o método dos Mínimos Quadrados Parciais ("Partial Least Squares: PLS"), usando descritores mecânico-quânticos obtidos de cálculos *ab initio*, com aplicação de efeitos de solvatação.

O método empregado para calcular os efeitos de solvatação utiliza-se do conceito do Contínuo Polarizável. Uma revisão sobre os métodos de solvatação foi publicada por Tomasi e Persico[3]. Este procedimento define uma cavidade de solvente com formato ajustado ao soluto, com os efeitos de interação eletrostática entre solvente e soluto incluídos no operador de Fock. Também podem ser incluídos efeitos para considerar a energia de formação da cavidade do solvente utilizado.

Foram escolhidos três solventes: (1) a água, que além de ser uma das fases usadas para obter o valor experimental de logP , é o solvente biológico; (2) o cicloexano, que é um solvente apolar, foi usado para modelar as interações soluto/solvente presentes na parte apolar do octanol; (3) a acetona, que é um solvente polar não prótico, deve ser capaz de modelar bem as interações entre a parte polar do *n*-octanol e o soluto[4]. O *n*-octanol não é um solvente adequado (molécula grande e não esférica) para utilização com o Método do Contínuo Polarizável ("Polarizable Continuum Method: PCM") de solvatação, principalmente por causa da formulação do termo de cavitação.

A correlação entre valores obtidos em cálculos teóricos e coeficientes de partição já foi conseguida por diversos autores, porém no caso aqui apresentado é utilizado um número menor de descritores que nos exemplos já publicados[5], com resultados de reprodução dos valores experimentais comparáveis.

Os métodos utilizados para obtenção dos modelos de regressão são baseados em regressão PLS e redes neurais. Estes métodos foram escolhidos pela sua boa aplicabilidade ao caso. Hermens utilizase de regressão PLS, e Análise Discriminante (também baseada em método PLS) para obter regressão entre valores de toxidade e classificação do conjunto de poluentes.

MÉTODOS

Inicialmente são calculados os descritores quânticos, para então obter uma matriz de dados. Esta matriz será analisada e utilizada para modelar os valores dos coeficientes de partição. Cada uma destas etapas será melhor descrita adiante.

As estruturas das cento e oitenta e oito moléculas mostradas na Tabela 1 foram submetidas a uma busca de conformações com menor energia. Esta busca foi feita por método sistemático para os compostos com substituição direta no anel, usando Hamiltoniano AM1. Este método é implementado no programa MOPAC93[6]. Para os demais casos, o método de geração das estruturas para otimização foi o da matriz de distâncias geométricas implementado no programa TINKER[7]. O método de otimização geométrica das estruturas geradas com o TINKER foi também o AM1. Para descartar as conformações que se tornam idênticas após a minimização de energia foi utilizado o programa TINKER. Um programa desenvolvido por Oliveira[8] faz a comunicação entre os diversos módulos e programas para a busca conformacional e descarte das estruturas repetidas.

Todas as estruturas das moléculas obtidas nos mínimos de energia, pré-otimizadas com AM1, foram completamente otimizadas usan-

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

do o método *ab initio* com base HF(3-21G), com o programa GAMESS99[9]. Propriedades moleculares foram calculadas com método *ab initio*. Para comparar o resultado do aumento do conjunto de base no sistema foram usadas as bases HF(3-21G) e HF(6-311G). O efeito de solvatação foi estudado com aplicação de dois modelos de cavidade progressivamente mais complexos. Para cada nível de complexidade de modelo foram realizados cálculos para os três tipos de solvente: água, acetona e cicloexano.

Usando HF(3-21G) foram calculadas quinze propriedades para o Modelo Um (M1) e trinta para o Modelo Dois (M2). Na Tabela 1 são mostrados os valores previstos em validação cruzada deixando um composto de fora em cada ciclo ("Leave One Out: LO"), para todos os compostos. O modelo M1 aplica efeito de cavitação segundo método de Pierotti[10] e Claverie[11]. Este tipo de cavidade será denominado S1. O modelo M2 aplica efeito de cavitação, o efeito da energia livre de repulsão, e também as forças de dispersão, segundo o método de Amovilli e Mennucci[12]. Este tipo de cavidade será denominado S2. Usando HF(6-321G//3-21G) foram calculadas quinze propriedades usando apenas o efeito de cavitação (cavidade S1) para obter os dados para o Modelo Três (M3).

Para os modelos M1, M2 e M3 foram calculadas as seguintes propriedades: Os Momentos de Dipolo (três descritores: μ_{aquoso} , $\mu_{acetona}$, e $\mu_{cicloexano}$). As cargas calculadas sobre o átomo mais negativamente carregado e as cargas calculadas sobre o hidrogênio mais positivamente carregado[13] (ou átomo de carbono ou nitrogênio mais positivamente carregado, nas moléculas sem hidrogênio) somam mais seis descritores.

Para todos os modelos (M1, M2 e M3) foram definidos três valores de partição (P) relativos aos três pares formados com os três solventes utilizados. Os volumes da cavidade formada por cada solvente também foram utilizados como descritores.

Para o Modelo Dois (M2) foram calculadas mais cinco propriedades, não calculadas para os demais: (1) a variação da energia interna do solvente; (2) a interação eletrostática soluto-solvente; (3) a energia de cavitação de Pierotti; (4) a energia livre de dispersão e (5) a energia livre de repulsão. Para cada uma das cinco propriedades são calculadas três descritores, totalizando mais quinze descritores. Estes descritores foram adicionados aos quinze descritores já mencionados.

As cargas foram calculadas usando o método CHELPG para todos os casos. Para o cálculo foi utilizado um arranjo com alta densidade de pontos para determinação das cargas. Foi determinado que a distância entre os pontos no arranjo não fosse maior que 0,3Å, para minimizar os efeitos conformacionais na magnitude das cargas calculadas[14]. O método CHELPG calcula cargas pontuais, nas posições atômicas, que são capazes de ajustar o valor da carga total molecular. Pode-se definir que as cargas também sejam restritas às capazes de ajustar os vetores do momento de dipolo e de quadrupolo, obtidos diretamente da aplicação do operador na função de onda Hartree-Fock. Em todos os casos foram aplicadas todas as restrições possíveis. O número total de restrições não pode ser maior que o número de átomos presentes na molécula.

Para criação de uma cavidade de solvente tipo S2 são necessários o índice de refratividade em freqüência zero (η), o potencial de ionização do solvente (PI), a densidade do solvente relativa à água (ρ) e a massa molar do solvente (MM). Para a acetona foram usados η =1,359; PI=0,413; ρ =0,788 e MM=58,08. Para o cicloexano foram usados os valores de η =1,426; PI=0,422; ρ =0,750 e MM=84,16. Os valores de η e ρ foram obtidos na literatura[15], o PI foi calculado por método *ab initio* no nível HF(6-311G), e a MM através da fórmula química. Os valores para a água constam no programa GAMESS99.

Os valores de logP para os cento e oitenta e oito compostos orgânicos mostrados na Tabela 1 foram modelados utilizando a matriz de dados obtida com os cálculos quânticos. O conjunto de moléculas utilizado para treinar os modelos de regressão inclui álcoois, halobenzenos, anilinas, fenóis, nitrobenzenos, ésteres, aminas e piridinas. Os valores de logP para os compostos foram obtidos na literatura, publicados por Hermens e colaboradores[2]. Também foram obtidos valores de logP utilizando o programa XLOGP[16] para os compostos derivados de piridina. Os resultados de validação 'LO' deram origem às estimativas dos três modelos (M1-LO1, M2-LO2 e M3-LO3) da Tabela 1.

Nós obtivemos os modelos de regressão com métodos PLS e Redes Neurais de Retropropagação ("Back Propagation Neural Network: BPN"). Na regressão com método PLS foram utilizadas funções implementadas[17] no programa OCTAVE[18], para a obtenção dos modelos de regressão. Os métodos para a validação cruzada tipo 'LO' e para a validação cruzada em bloco (EV) também foram implementados. Para estes modelos obtidos com método PLS foram utilizadas entre cinco e dez Variáveis Latentes (VL) para regressão. O número de VL foi selecionado baseado no valor do coeficiente de correlação em validação cruzada (q^2), obtido comparando-se os valores previstos por 'LO' com os valores publicados. Os valores das variâncias acumuladas pela regressão PLS dos conjuntos de dados experimentais e dos conjuntos de descritores mecânico-quânticos são mostrados na Tabela 2.

Foi treinada uma BPN utilizando os escores de uma Análise por Componentes Principais ("Principal Component Analisys: PCA") realizada para o conjunto de trinta descritores obtidos para o Modelo 2. Os dados de acumulação de variância pela PCA são mostrados na Tabela 2, na sétima coluna. A variância acumulada de 99,19% permite concluir que toda a informação existente no conjunto de trinta descritores foi capturada pelas dez Componentes Principais ("Principal Components: PC"). Esta rede foi treinada com onze neurônios na primeira camada: dez para os dados dos escores das PC, mais um de bias. Na segunda camada foram usados trinta e seis neurônios. O número de neurônios da segunda camada foi determinado por tentativa e erro. Na terceira camada foi usado um neurônio para saída do sinal obtido, que corresponde à resposta do modelo para o valor de logP. O valor limite para o erro de treinamento da rede para os compostos foi de 10^{-4} . O limite de iterações para atingir convergência usado foi de cinqüenta mil. Esta rede foi utilizada para previsão em validação cruzada tipo 'LO' dos compostos desta série. O programa utilizado para treinamento de redes neurais foi o PSDD[19]. As funções utilizadas na rede neural são do tipo sigmóide (não lineares), havendo também um termo linear adicionado, com coeficiente de combinação definido em 0,02.

RESULTADOS E DISCUSSÃO

Foram testados modelos com propriedades calculadas utilizando HF(6-311G//3-21G) com a cavidade S1 (M3) e com HF(3-21G) utilizando as cavidades S1 (M1) e S2 (M2). Os resumos dos resultados obtidos com a metodologia aqui proposta para o cálculo de logP estão listados na Tabela 3. Sob a coluna 'LO1' estão mostrados os indicadores da qualidade da regressão obtidos com M1 para validação cruzada 'LO'. O valor de q^2 de 0,79 indica uma boa capacidade de previsão deste modelo. O valor da estimativa do desvio padrão de previsão ("Standard Deviation of Prediction: SDEP"), de 0,68 unidades de logP, também mostra que este modelo é válido. Este valor é da mesma ordem de grandeza que o esperado em outros métodos publicados para estimativa do valor de logP [20]. Também se aproxima da maior exatidão obtida com método experimental tipo Shake Flask[21], utilizado como um padrão analítico.

Os valores dos indicadores sob a coluna 'EV1' da Tabela 3 são os obtidos na previsão por validação cruzada em bloco. É usado um

Tabela 1. Número atribuído, nomes dos compostos utilizados para teste da metodologia de cálculo do valor de logP aqui proposto, valores experimentais publicados e valores estimados pelos modelos M1, M2 e M3 citados no texto

No.	Nome	Exp.	M1-LO1	M2-LO2	M3-LO3
1	guanidina	-1,510	-1,319	-1,561	-1,728
2	1,2-etanodiol	-1,360	-0,761	-0,927	-0,605
3	2-aminoetanol	-1,310	-0,350	-0,462	-0,275
4	dietilenoglicol	-1,305	-0,121	-0,384	-0,591
5	trietilenoglicol	-1,240	0,575	0,519	0,111
6	1-amino-2-propanol	-0,960	0,556	-0,217	0,340
7	2-metoxietanol	-0,770	0,997	0,689	0,374
8	2-metoxietilamina	-0,672	0,830	0,465	0,177
9	etanol	-0,310	-0,233	-0,485	-0,202
10	acetona	-0,240	-1,029	-1,109	-1,637
11	etilamina	-0,130	0,254	-0,056	-0,016
12	2-etoxietanol	-0,100	0,898	0,841	0,945
13	4-hidroxianilina	0,040	0,518	0,384	0,535
14	2-isopropoxietanol	0,050	2,576	1,734	1,394
15	2-propanol	0,050	0,460	0,034	0,413
16	4-hidroximetilfenol	0,250	1,084	0,932	1,013
17	t-butanol	0,350	0,960	0,236	0,711
18	4-(n-metoximetil) aminofenol	0,475	1,886	2,097	2,215
19	hidroquinona	0,590	-0,168	-0,002	-0,140
20	3-nitropiridina	0,660	1,131	0,899	0,651
21	1,3-diidroxibenzeno	0,800	0,313	0,145	0,355
22	2-butoxietanol	0,830	1,854	1,170	0,895
23	dietiléter	0,870	1,913	1,632	0,980
24	anilina	0,900	0,905	1,091	0,924
25	4-amino-2-nitrofenol	0,960	1,672	1,257	1,461
26	butilamina	0,970	1,393	0,833	0,942
27	4-metilamoniofenol	0,974	1,086	0,854	1,458
28	benzilamina	1,090	1,314	1,540	1,245
29	1,2-dimetilpropilamina	1,102	2,297	1,669	2,057
30	2,2-dimetilpropilamina	1,192	2,003	1,014	1,117
31	3-pentanol	1,210	1,615	1,002	1,490
32	diclorometano	1,250	0,684	0,953	1,482
33	2-metilanilina	1,320	1,528	1,657	1,465
34	4-metoxifenol	1,340	1,420	1,604	1,085
35	tiazol	1,350	0,608	1,046	0,873
36	3-nitroanilina	1,370	1,475	1,388	1,362
37	3-cloropiridina	1,390	1,869	1,985	1,681
38	4-cloropiridina	1,390	1,270	1,761	1,259
39	4-metilanilina	1,390	1,513	1,784	1,566
40	4-nitroanilina	1,390	1,584	1,448	1,222
41	3-metilanilina	1,400	1,798	1,390	1,522
42	1-adamantanoamina	1,436	2,266	1,729	1,564
43	3-etilpiridina	1,460	2,304	2,180	1,974
44	fenol	1,460	1,039	0,825	1,067
45	1,2-dicloroetano	1,480	1,024	1,361	1,270
46	2-cloropiridina	1,480	1,893	2,264	1,713
47	4-bromopiridina	1,570	1,943	1,828	-
48	3-metoxifenol	1,580	1,546	1,340	1,413
49	4-cianofenol	1,600	1,099	1,391	1,333
50	3,3-dimetilbutilamina	1,721	2,258	1,354	1,681
51	1,1-dicloroetano	1,790	1,378	1,622	2,128
52	2-nitroanilina	1,850	1,817	1,564	1,252
53	nitrobenzeno	1,850	1,604	1,356	1,593
54	3-etilanilina	1,855	2,085	2,150	2,022
55	4-etilanilina	1,855	2,046	2,254	2,072
56	3-cloroanilina	1,880	1,956	2,217	2,055
57	4-cloroanilina	1,880	1,961	2,258	2,052
58	1,1,2-tricloroetano	1,890	1,309	1,654	1,737
59	2-cloroanilina	1,900	1,920	2,149	1,862
60	4-nitrofenol	1,910	0,988	0,834	1,125
61	4-metilfenol	1,940	1,661	1,508	1,771
62	2-metilfenol	1,950	1,294	1,754	1,578
63	3-metilfenol	1,960	1,993	1,490	1,800
64	clorofórmio	1,970	1,658	1,819	2,783
65	1,2,3-tricloropropano	1,980	1,695	1,647	1,918
66	1,2-dicloropropano	1,987	1,621	1,844	1,721
67	1,3-dicloropropano	2,000	1,560	1,852	1,738
68	3-nitrofenol	2,000	0,876	0,830	1,048
69	3,4-dicloropiridina	2,010	2,491	2,612	2,230
70	3,5-dicloropiridina	2,010	2,509	2,696	2,416
71	2-cloro-4-nitroanilina	2,055	2,378	2,210	2,023
72	2,4-dicloropiridina	2,100	2,643	2,918	2,506
73	2,5-dicloropiridina	2,100	2,672	2,996	2,397
74	2-clorofenol	2,150	1,878	1,952	2,312
75	benzeno	2,186	1,041	1,372	1,732
76	3,4-dimetilfenol	2,230	2,108	2,065	2,306
77	2-cloronitrobenzeno	2,240	2,451	2,556	2,212
78	3-benziloxipiridina	2,250	3,159	3,876	2,698
79	4-bromoanilina	2,260	2,262	2,219	-
80	2,4-dimetilfenol	2,300	2,077	2,021	2,351
81	2-nitrotolueno	2,300	2,303	2,098	2,094
82	2-cloro-6-nitrofenol	2,326	2,376	2,099	2,040
83	2,6-dimetilfenol	2,360	1,954	1,934	2,307
84	4-nitrotolueno	2,370	2,285	2,255	2,279
85	4-etoxi-2-nitroanilina	2,387	2,916	2,631	2,398
86	1,1,2,2-tetracloroetano	2,390	2,395	2,393	3,186
87	4-clorofenol	2,390	1,555	1,814	1,909
88	4-cloronitrobenzeno	2,390	2,019	2,166	2,013
89	3-nitrotolueno	2,420	2,392	2,103	2,362
90	tricloroeteno	2,420	2,417	2,352	3,313
91	3-cloronitrobenzeno	2,460	2,382	2,381	2,314
92	1,1,1-tricloroetano	2,490	2,295	2,541	2,894
93	3-clorofenol	2,500	1,878	1,926	2,189
94	2-alilfenol	2,548	2,533	2,468	2,712
95	4-etilfenol	2,580	2,084	1,903	2,249
96	1-clorobutano	2,640	2,062	1,852	1,958
97	2-cloro-4-metilfenol	2,654	2,599	2,719	3,020
98	3,4-dicloroanilina	2,690	2,783	3,008	2,785
99	2,3,4-tricloropiridina	2,720	3,418	3,434	3,088
100	2,4,5-tricloropiridina	2,720	3,161	3,516	3,098
101	2,6-diclorofenol	2,750	2,725	2,801	2,686
102	3-benziloxianilina	2,772	3,739	4,097	3,501
103	tolueno	2,786	2,416	2,776	1,912
104	2,3,6-tricloropiridina	2,810	3,594	3,903	3,340
105	4-butilpiridina	2,810	3,480	3,372	2,949
106	1-metileptilamina	2,819	3,540	2,748	3,153
107	2,3-dimetilnitrobenzeno	2,830	2,779	2,619	2,556
108	tetraclorometano	2,830	2,798	2,829	2,285
109	1-naftol	2,840	1,860	2,375	2,350
110	2,3-diclorofenol	2,840	2,380	2,565	2,548
111	clorobenzeno	2,898	2,167	2,578	2,827
112	2,5-dicloronitrobenzeno	2,900	3,059	3,310	2,913
113	3,5-dicloroanilina	2,900	2,903	3,235	2,924
114	2,4-dicloroanilina	2,910	2,862	-	2,844
115	3,4-dimetilnitrobenzeno	2,910	2,439	2,766	2,787
116	4-butilanilina	2,913	2,010	2,166	2,041
117	2,5-dicloroanilina	2,920	2,597	3,000	2,580
118	2,3,6-trimetilfenol	2,922	2,511	2,438	2,796
119	2,3-dicloronitrobenzeno	3,050	2,880	2,988	2,766
120	4-cloro-2-nitrotolueno	3,050	2,914	2,925	2,892
121	2,4-diclorofenol	3,060	2,230	2,613	2,379
122	2,5-diclorofenol	3,060	2,006	2,371	2,036
123	2,4-dicloro-6-nitrofenol	3,066	2,883	2,841	2,635
124	2,4-dicloronitrobenzeno	3,090	3,111	3,158	2,865
125	2-cloro-6-nitrotolueno	3,090	2,480	3,108	2,800
126	2-fenilfenol	3,090	2,515	3,333	3,156
127	4-cloro-3-metilfenol	3,100	2,106	2,125	2,340
128	o-xileno	3,120	2,375	2,528	2,411
129	3,5-dicloronitrobenzeno	3,130	3,017	3,088	3,034

Tabela 1. continuação

No.	Nome	Exp.	M1-LO1	M2-LO2	M3-LO3
130	p-xileno	3,150	2,391	2,750	2,549
131	4-propilfenol	3,200	2,713	2,508	2,769
132	m-xileno	3,200	2,402	2,716	1,698
133	2,3,4,5-tetracloropiridina	3,250	4,081	4,054	4,196
134	4,5-dicloro-2-metoxifenol	3,260	3,466	3,552	3,830
135	3-clorotolueno	3,280	2,912	3,178	3,181
136	4-t-butilfenol	3,310	2,728	2,484	2,923
137	2,3,6-tricloroanilina	3,323	3,567	3,809	3,391
138	3,4-diclorofenol	3,330	2,307	2,495	2,588
139	4-clorotolueno	3,330	2,840	3,222	3,178
140	4-fenoxifenol	3,350	3,376	3,728	3,515
141	tetracloroeteno	3,400	3,127	3,100	2,959
142	1,2-diclorobenzeno	3,433	3,061	3,235	3,724
143	4-cloro-2,3-dimetilfenol	3,433	2,766	2,800	3,012
144	2,3,5,6-tetracloropiridina	3,440	4,287	4,474	4,068
145	1,4-diclorobenzeno	3,444	2,583	3,214	3,238
146	4-cloro-3,5-dimetilfenol	3,483	3,111	2,685	3,004
147	1,3-diclorobenzeno	3,525	3,031	3,429	3,669
148	1,3-diclorobenzeno	3,525	3,032	3,333	3,669
149	4-cloro-2-alilfenol	3,558	3,461	3,524	3,658
150	4-n-butilfenol	3,561	3,501	3,047	3,586
151	4-n-butilfenol	3,561	3,413	3,057	3,501
152	2,3,5-triclorofenol	3,577	3,088	3,303	3,195
153	3,5-diclorofenol	3,620	2,573	2,821	2,908
154	pentacloroetano	3,627	3,242	3,058	4,201
155	2,3,4-tricloroanilina	3,680	3,607	3,758	3,374
156	2,4,5-tricloroanilina	3,690	3,649	3,960	3,574
157	2,4,6-triclorofenol	3,690	3,444	3,679	3,619
158	4-hexiloxianilina	3,694	5,016	4,620	4,340
159	2,4,5-triclorofenol	3,720	2,783	3,079	2,990
160	3,4,5-tricloro-2,6-dimetoxifenol	3,740	3,460	3,844	3,937
161	2,3,6-triclorofenol	3,770	3,076	3,370	3,300
162	3,4,5-tricloro-2-metoxifenol	3,770	4,089	4,038	4,501
163	2-t-butil-4-metilfenol	3,800	3,659	3,522	3,720
164	4-t-pentilfenol	3,830	3,234	2,925	3,373
165	2,3,5,6-tetraclorofenol	3,880	3,968	4,068	3,728
166	2,4,6-tribromofenol	3,917	4,471	3,873	-
167	4-cloro-2-isopropil-5-metilfenol	3,920	4,884	3,288	3,991
168	4-fenilazofenol	3,957	3,949	4,333	4,477
169	2,3,4,5-tetracloroanilina	4,040	4,371	4,442	4,062
170	1,2,4-triclorobenzeno	4,050	3,700	3,999	4,250
171	3,4-diclorotolueno	4,067	4,056	3,986	3,940
172	4-n-pentilfenol	4,090	3,936	3,626	4,104
173	1,2,3-triclorobenzeno	4,139	3,818	3,906	4,368
174	1,3,5-triclorobenzeno	4,189	3,627	4,174	4,206
175	2,3,4,5-tetraclorofenol	4,210	4,284	4,091	4,353
176	2,4-diclorotolueno	4,240	4,013	4,101	3,966
177	3,4,5-triclorofenol	4,280	3,274	3,280	3,429
178	3,4,5,6-tetracloro-2-hidroxifenol	4,290	3,181	3,201	2,857
179	2,3,4,6-tetraclorofenol	4,450	3,696	3,959	3,755
180	2,3,5,6-tetracloroanilina	4,460	4,212	4,548	4,131
181	1,2,4,5-tetraclorobenzeno	4,604	4,005	4,578	4,802
182	1,2,3,4-tetraclorobenzeno	4,635	4,885	4,465	4,996
183	1,2,3,5-tetraclorobenzeno	4,658	4,569	4,718	4,948
184	2,4,5-triclorotolueno	4,780	4,332	4,691	4,553
185	pentaclorobenzeno	5,183	5,110	5,056	5,575
186	4-decilanilina	6,087	6,308	5,654	6,172
187	4-nonilfenol	6,206	6,558	5,965	-
188	4-dodecilanilina	7,145	7,053	-	-

Tabela 2. Variâncias acumuladas nas componentes principais utilizadas. O modelo PLS e PCA para a construção dos modelos propostos

	PLS				PCA	
	Descritores		Atividades		Descritores	
VL #	Esta	Total	Esta	Total	Esta	Total
1	40,55	40,55	54,12	54,12	41,41	41,41
2	14,84	55,38	12,30	66,42	24,69	66,10
3	16,20	71,58	5,83	72,25	15,86	81,96
4	4,28	75,86	5,81	78,05	5,94	87,90
5	13,48	89,35	1,35	79,40	3,48	91,38
6	1,86	91,21	4,01	83,41	2,89	94,27
7	2,04	93,25	1,03	84,44	1,92	96,19
8	2,17	95,42	0,16	84,60	1,57	97,76
9	2,05	97,47	0,09	84,69	1,00	98,76
10	0,50	97,97	0,14	84,83	0,43	99,19

Tabela 3. Resultados obtidos para validação cruzada deixando um composto de fora (LO) e para validação cruzada em bloco (EV) usando 10% do conjunto para validação

	M1		M2		M3	
	LO1	EV1	LO2	EV2	LO3	EV3
PRESS	84,8	109,50	58,4	36,53	49	67,12
q^2	0,79	0,67	0,85	0,77	0,87	0,79
SDEP	0,68	0,77	0,56	0,45	0,52	0,61
n	188	188	186	186	183	183

conjunto com cento e sessenta e nove compostos para treinamento e dezenove compostos para previsão (≈10%) em cada ciclo, escolhidos aleatoriamente. São realizados tantos ciclos quanto necessário para que todos os compostos sejam retirados uma (única) vez do conjunto de treinamento para o conjunto de validação. Os valores dos índices de acerto obtidos para esta previsão aproximam-se dos obtidos na previsão 'LO1', mostrando a estabilidade do modelo em relação à retirada de blocos de amostras para previsão.

O modelo M2 foi obtido com PLS usando dez variáveis latentes. A acumulação de variância das trinta variáveis no bloco dos descritores é alta, como pode ser visto na terceira coluna da Tabela 2. O grande número de VL necessárias para acumular uma variância expressiva mostra que os descritores usados para modelagem de logP contém grande quantidade de informação, que pode ser utilizada pelo modelo PLS. Um aumento do número de VL além das dez mostradas não seria conveniente, porque toda a variância dos dados de atividade biológica possível de ser modelada já foi acumulada pelo modelo PLS. Isto é mostrado na Tabela 2 pela grande diminuição da variância acumulada pela décima VL dos descritores quânticos (na segunda coluna) em relação à nona VL, e também pelos pequenos valores de variância acumulada nas últimas VL das atividades, na quarta coluna de dados.

Os indicadores da qualidade de previsão obtidos em M2 por validação cruzada 'LO' são mostrados na Tabela 3 sob a coluna 'LO2' (este modelo usa a cavidade S2 no nível de cálculo HF(3-21G)). Esta tabela mostra um valor q^2 de 0,85 para M2 com SDEP de 0,56 para 186 compostos. Estes indicadores mostram que este modelo é capaz de prever os valores experimentais de logP com boa precisão. Estes resultados são melhores que os obtidos no modelo M1. Os

indicadores obtidos para estimativa do erro de previsão na validação cruzada em bloco, 'EV2', também mostram um modelo mais robusto que M1. Para este modelo não foi possível realizar cálculos para os compostos 114 e 188 da Tabela 1, por causa de problemas com a obtenção das cavidades de acetona. Para cada caso em que o cálculo quântico para obtenção dos descritores não foi possível, é colocado um traço no valor da atividade biológica prevista para o composto na Tabela 1.

Os indicadores da qualidade da regressão obtidos em validação cruzada 'LO' com M3 são mostrados na Tabela 3 sob a coluna 'LO3'. Este modelo usa a cavidade S1 no nível de cálculo HF(6-311G//3-21G). Para M3 não foram realizados os cálculos para 5 compostos: 47, 79, 166, 187 e 188. Os problemas para obter resultados quânticos com este modelo são: (1) a não disponibilidade da base HF(6-311G) para muitos elementos e (2) a dificuldade para obtenção das cavidades de acetona e cicloexano. Cálculos com cavidade S2 não foram realizados por causa do grande aumento no tempo de execução.

O modelo M3 teve resultados ligeiramente melhores que os obtidos com o modelo M2 na previsão de compostos em validação cruzada, com q^2 de 0,87; SDEP de 0,52 e PRESS de 49. Os resultados de validação cruzada em bloco para este modelo também são ligeiramente melhores que os obtidos no modelo M2, como pode ser observado na Tabela 3. Estes dados mostram que este modelo é capaz de prever os valores experimentais com boa precisão, ligeiramente maior que a de M2.

Comparando os indicadores da capacidade de previsão de M3 com os correspondentes de M1 na Tabela 3 pode-se ter uma idéia da variação na capacidade de previsão em função da base Hartree-Fock utilizada. O aumento da base de HF(3-21G) para HF(6-311G) permite um aumento na capacidade de previsão da ordem de 0,1 unidade em q^2. O aumento da base, com cavidade S1, corresponde também ao aumento do tempo de execução em um fator de 1,7 para o composto 180 da Tabela 1.

A maior sofisticação do efeito do solvente, trocando a cavidade S1 por cavidade S2, proporciona um aumento do valor de q2 em aproximadamente 0,1 unidade, como pode ser visto comparando os resultados dos modelos M2 e M3. A utilização da cavidade S2 torna os cálculos 4,5 vezes mais lentos que com cavidade S1, mantendo a mesma base HF(3-21G), para o composto 180 da Tabela 1.

O modelo considerado mais interessante para uso geral é o modelo M2, pela sua maior disponibilidade em relação ao conjunto de base e pelos resultados obtidos, praticamente tão bons quanto os obtidos com M3. O aumento do tempo computacional do modelo M2 em relação a M3 é razoável, não sendo um grande empecilho para adoção de um ou outro procedimento. A maior disponibilidade do conjunto de base para cálculos HF(3-21G) é um motivo mais importante. Esta base é disponível para elementos químicos do hidrogênio ao xenônio. A utilização da base HF(6-311G) limita as previsão para compostos que tenham elementos do hidrogênio ao neônio, somente. Deve-se levar em consideração que é necessário um grande número de cálculos para o conjunto de calibração. Não é interessante manter dois conjuntos de calibração, para escolher o modelo utilizado no momento de realizar a previsão do logP de uma nova molécula.

Os valores obtidos em validação cruzada 'LO' para logP com M2, lançados em gráfico contra os valores publicados na literatura são mostrados na Figura 1. As médias dos valores experimentais ($\hat{V}$ =2,409) e previstos ($\hat{v}$ =2,408) são praticamente iguais. O coeficiente angular (b = 0,85) da regressão ($y = a+bx$) dos conjuntos de pontos é próximo de um, e a constante de regressão (a = 0,026) é pequena. Estes indicadores da regressão entre valores experimentais e previstos mostram que o modelo é capaz de prever com boa exatidão os valores experimentais.

Na Figura 2 são mostrados os erros de previsão corrigidos pela distribuição de Student de cada composto em validação cruzada tipo 'LO' lançados no gráfico contra os valores de "leverage". Os sete compostos com erro maior que dois desvios padrão são distribuídos homogeneamente, acima e abaixo das linhas tracejadas. Estes compostos são os numerados como 5, 7, 14, 18, 68, 78 e 102 na Tabela 1 e nas Figuras 1 e 2.

Comparando a estrutura de compostos mal previstos (5, 7, 14, 18 e 68) vemos que têm em comum o fato de serem todos funcionalizados com hidroxila. São bastante eficientes na formação de pontes de hidrogênio, e este fator pode ser o responsável pela má previsão de sua atividade. Os modelos utilizados para simulação de efeitos de solvatação não têm como representar perfeitamente a formação de pontes. Uma simulação explícita do solvente seria neces-

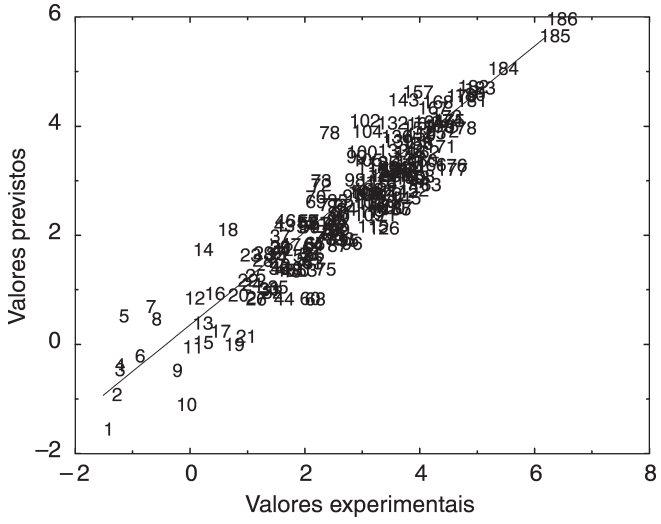

Figura 1. *Valores experimentais e calculados para log $P_{(oc\text{-}aq)}$ para os compostos mostrados na Tabela 1 sob a coluna M1-LO1. A regressão linear entre valores experimentais e previstos é mostrada pela linha pontilhada*

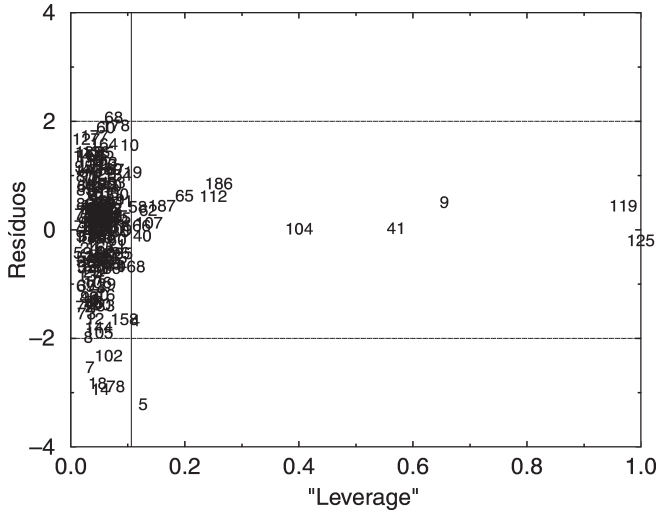

Figura 2. *Valores dos erros de previsão (em unidades de desvio padrão) lançados em gráfico contra os valores de "leverage" obtidos para cada composto com o modelo M2-LO2 da Tabela 1. O limite proposto para os valores de "leverage" é mostrado na linha pontilhada, e o limite de dois desvios padrão pelas linhas tracejadas*

sária para este tipo de representação, com custo computacional proibitivo. A inclusão explícita de solvente, formando um cluster molecular nas regiões onde há maior interação solvente-soluto também introduz erro no cálculo do termo de cavitação, por causa do aumento do volume da cavidade. Os compostos 78 e 102 têm grupo benzilóxi ligado à piridina e anilina. Estes compostos tiveram o valor de logP superestimado, provavelmente por causa do volume das cavidades de solvente necessárias para acomodá-los, interpretada pelo modelo como uma característica associada a compostos mais lipofílicos.

Os valores de "leverage" ajudam a identificar compostos com características anômalas dentro do conjunto de dados. Este valor é uma estimativa da importância de cada ponto para a obtenção dos coeficientes na regressão multivariada. Cada composto com valor de "leverage" superior ao mostrado pela linha pontilhada na Figura 2 deve ser analisado com cuidado. Estes compostos podem ser considerados anômalos em alguns casos. Compostos com valor de logP distante dos extremos da faixa considerada no treinamento do modelo e que tiverem um valor de "leverage" alto, podem ser eliminados do conjunto de treinamento para que se obtenha um modelo mais robusto. Este é o caso dos compostos 41, 65, 104, 112, 119 e 125. Estes compostos foram mantidos porque apesar de ter alto valor de "leverage", a previsão do valor de log P para estes compostos é boa. Significa que apesar de terem ponderação grande na obtenção do modelo de regressão, seus dados não introduzem um grande fator de erro no modelo obtido. Os compostos 9, 186 e 187 estão nos limites inferior e superior da faixa de valores de logP utilizada no treinamento, e o valor de "leverage" para amostras nesta faixa pode ser mais alto sem que isto seja considerado anormal. O limite proposto para o valor de "leverage" não é uma regra fixa, e sim um alerta para verificação cuidadosa dos casos críticos.

Os resultados do modelo obtido com método de treinamento de BPN são desanimadores. Há uma grande capacidade de ajuste de valores por este método, porém as redes treinadas não são capazes de gerar estimativas razoáveis para os valores de logP dos compostos em testes de validação cruzada. O valor baixo de q^2 =0,55 obtido com a rede, com SDEP=1,09 e PRESS de 170 mostra um modelo pouco robusto e incapaz de gerar previsões confiáveis. A pouca capacidade de previsão da rede treinada provavelmente está associada ao seu caráter não linear. Isto permite uma grande capacidade para reproduzir os valores de treinamento, porém pode resultar em valores muito diferentes para a propriedade modelada por causa de pequenas variações nos parâmetros aplicados ou nos dados de entrada. Além de resultados ruins, o treinamento de uma rede neural complicada como esta, é muito demorado se comparado ao necessário para obter um modelo tipo PLS. O tempo é da ordem de 24 h, num computador Pentium II 350Mhz, comparado com alguns minutos para obter os resultados de validação cruzada, em cálculos de regressão com PLS.

CONCLUSÕES

Este trabalho mostra como podem ser calculados valores de logP para compostos orgânicos simples, de tamanho moderado, sem aplicação de parâmetros arbitrários. A aplicação de alguns poucos parâmetros necessários para simulação do efeito de solvatação mais sofisticado (nas cavidade do S2) usa dados experimentais ou calculados com HF(6-311G). Portanto, não podem ser considerados arbitrários. Este tipo de simulação de efeito de solvatação (cavidade S2) aumenta bastante o tempo de processamento.

O modelo M2 foi considerado o mais promissor, pelos resultados obtidos e pela aplicação possível para um número maior de casos. A utilização da cavidade S2 com base HF(3-21G) pode ser considerada como uma proposta geral para modelagem de logP usando método de solvatação tipo PCM. Se não há elementos para os quais a base HF(6-311G) não está disponível este modelo pode ser indicado tanto pelos resultados como pela economia de recursos computacionais. A utilização de base HF(6-311G), com cavidade S2, provavelmente levaria a resultados melhores, com custo computacional alto.

A metodologia proposta é bastante limitada para prever o logP de compostos que realizam pontes de hidrogênio muito eficientemente, ou que têm grande volume molecular e ainda assim são bastante hidrofílicos. Para estes casos este método deve ser usado com cautela.

Uma vantagem do método apresentado é que utiliza somente programas de domínio público. Este é o fator que propulsiona o desenvolvimento desta metodologia, já que os procedimentos e rotinas explicitados podem ser repetidos por qualquer pesquisador, em qualquer universidade, sem nenhum custo na instalação dos programas utilizados. Os cálculos quânticos envolvidos podem ser realizados em computadores pessoais modernos, sem necessidade das grandes estruturas dos centros de alto desempenho, ainda que sejam mais rápidos assim.

AGRADECIMENTOS

Agradecemos à FAPESP pelo apoio financeiro ao projeto, sob os processos 98/06065-5 e 00/02187-0. A Universidade Estadual de Campinas nos fornece apoio institucional e espaço físico para o desenvolvimento do trabalho. A CAPES e o CNPq também colaboram através de subvenções à aquisição de material para pesquisa. O CENAPAD-SP colabora com recursos computacionais. Os colegas de grupo também colaboram na elaboração dos manuscritos, com sugestões e correções importantes.

REFERÊNCIAS

1. Verhaar, H. J. M.; van Leeuwen, C. J.; Hermens, J. L. M.; *Chemosphere* **1992**, 471.
2. Henk, J. M.; Ramos, E. U.; Hermens, J. L. M.; *J. Chemom.* **1996**, *10*, 149.
3. Tomasi, J.; Persico, M.; *Chem. Rev.* **1994**, *94*, 2027.
4. Gramatica, P.; Navas, N.; Todeschini, R.; *Trends Anal. Chem.* **1999**, *18*, 461.
5. Leo, A. J.; *Chem. Rev.* **1993**, *93*, 1281.
6. http://www.home.att.net/ mrmopac, acessada em Outubro 1999.
7. http://www.dasher.wustl.edu/tinker, acessada em Feveriro 2000.
8. de Oliveira, K. M. G.; trabalho não publicado.
9. http://www.msg.ameslab.gov/GAMESS/GAMESS.html, acessada em Outubro 1999.
10. Pierotti, R. A.; *Chem. Rev.* **1976**, *76*, 717.
11. Langlet, J.; Claverie, P.; Caillet, J.; Pullman, A.; *J. Phys. Chem.* **1988**, *92*, 1617.
12. Amovilli, C.; Mennucci, B.; *J. Phys. Chem.* **1997**, *B101*, 1051.
13. Vaes, W. E. U. R.; Verhaar, H. J. M.; Cramer, C. J.; Hermens, J. L. M.; *Chem. Res. Toxicol.* **1998**, *11*, 847.
14. Breneman, C. M.; Wiberg, K.; *J. Comp. Chem.* **1990**, *11*, 361.
15. Weast, R. C.; *CRC Handbook of Chemistry and Phisics*, 67th ed., CRC Press Inc.: Boca Raton, 1987, p 1063.
16. Lai, L.; *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 615.
17. http://www.mitra.iqm.unicamp.br/eborges/links.html, acessada em Maio 2001.
18. ftp://ftp.che.wisc.edu/pub/octave, acessada em Maio 2001.
19. Ichikawa, H.; *PSDD: Perceptron-type Neural Network Simulator*, Hoshi College of Pharmacy, 2-4-41 Ebara, Shinagawa, Tokyo 142 Japan, 1989.
20. Karelson, M.; Lobanov, V. S.; Katritzky, A. R.; *Chem. Rev.* **1996**, *96*, 1027.
21. do Amaral, A.; Malvezzi, A.; Gonçalves, R. S.; *Resumos da 24ª Reunião da Sociedade Brasileira de Química*, Poços de Caldas, Brasil, 2001.