PROPOSIÇÃO, VALIDAÇÃO E ANÁLISE DOS MODELOS QUE CORRELACIONAM ESTRUTURA QUÍMICA E ATIVIDADE BIOLÓGICA

Anderson Coser Gaudio*
Departamento de Física, Centro de Ciências Exatas, Universidade Federal do Espírito Santo, Campus de Goiabeiras, 29060-900 Vitória - ES
Eliana Zandonade
Departamento de Estatística, Centro de Ciências Exatas, Universidade Federal do Espírito Santo

Recebido em 7/4/00; aceito em 15/12/00

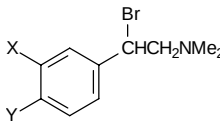
PROPOSITION, VALIDATION AND ANALYSIS OF QSAR MODELS. The present paper aims to bring under discussion some theoretical and practical aspects about the proposition, validation and analysis of QSAR models based on multiple linear regression. A comprehensive approach for the derivation of extrathermodynamic equations is reviewed. Some examples of QSAR models published in the literature are analyzed and criticized.

Keywords: quantitative structure-activity relationships; multiple linear regression; validation of QSAR models.

INTRODUÇÃO

Em sua sétima edição do ano de 1973, o *Journal of Medicinal Chemistry* publicou um artigo de autoria de Unger e Hansch[1] que é considerado por muitos como um marco no desenvolvimento de QSAR, abreviação em inglês para Relações Quantitativas entre Estrutura e Atividade. O artigo tornou-se célebre por estabelecer regras gerais para a elaboração e validação dos modelos matemáticos que correlacionam estrutura química e atividade biológica. A publicação desse artigo foi conseqüência da publicação anterior de dois outros artigos, em que seus autores apresentaram modelos matemáticos distintos para analisar a atividade biológica da mesma série de compostos.

Tudo começou com o artigo de Hansch e Lien[2], em que se analisou a atividade antiadrenérgica de vinte e dois compostos derivados da N,N-dimetil-α-bromo-feniletilamina (**1**), substituídos nas posições meta e para do anel fenila, cujos valores haviam sido determinados cinco anos antes[3].



1

Segundo Hansch e Lien, a atividade antiadrenérgica dos compostos derivados da estrutura **1** poderia ser representada como uma função linear dos efeitos lipofílico e eletrônico que os grupos X e Y proporcionam à estrutura **1** (eq 1).

$$\log 1/C = 1{,}22\ \pi - 1{,}59\ \sigma + 7{,}89 \quad (1)$$
$$(n = 22;\ R = 0{,}918;\ s = 0{,}238)$$

Na eq 1, C representa a concentração do fármaco, em moles/kg de peso corporal, capaz de produzir 50% de antagonismo à ação vasopressora de uma dose padrão de epinefrina em ratos, π é a constante lipofílica de Hansch[4], σ é a constante eletrônica de Hammett[5], n é o número de compostos incluídos no modelo, R é o coeficiente de correlação do modelo e s é o desvio-padrão do modelo. Neste ponto cabe um esclarecimento. Optou-se por apresentar as equações citadas em sua forma original. Assim que o formato apropriado de apresentação dos modelos matemáticos de QSAR for mostrado (ver adiante), o leitor poderá comparar as diversas formas de apresentação já utilizadas ao longo do tempo.

Em 1972, Cammarata[6] apresentou a eq 2 como alternativa para a representação da atividade dos compostos derivados da estrutura **1**.

$$\log 1/C = 0{,}747\ (\pm 0{,}123)\ \pi_m - 0{,}911\ (\pm 0{,}249)\ \sigma_m + 1{,}666\ (\pm 0{,}124)\ r_v^p + 5{,}769 \quad (2)$$
$$(n = 22;\ R = 0{,}961;\ s = 0{,}168)$$

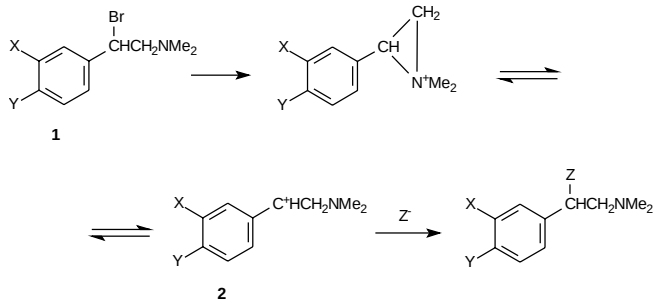
Na eq 2, π_m e σ_m são as constantes lipofílica e eletrônica dos grupos químicos presentes na posição meta do anel fenila da estrutura **1** (X), r_v^p é o raio de van der Waals do substituinte na posição para (Y) e os números entre parênteses correspondem aos desvios-padrão dos coeficientes da equação.

Os valores numéricos de R e s na eq 2 indicam que o modelo de Cammarata consegue explicar maior quantidade da variabilidade dos valores da atividade biológica do que o modelo representado pela eq 1. No entanto, deve-se levar em consideração que o segundo membro da eq 2 contém uma variável a mais do que a eq 1, o que certamente contribui para sua melhor qualidade.

Em 1973, Unger e Hansch[1] reagiram ao modelo proposto por Cammarata, afirmando que o mesmo continha inconsistências relativas às variáveis utilizadas para descrever a atividade biológica e de forma alguma apresentava embasamento bioquímico, o que o invalidava. Alguns dos argumentos citados foram: (a) o modelo não atribui efeito hidrofóbico aos substituintes presentes na posição para. A variável r_v^p dos poucos substituintes (seis) na posição para utilizados no modelo está acidentalmente correlacionada aos efeitos hidrofóbico (R = 0,840) e hidrofóbico/eletrônico (R = 0,983). Portanto, não é possível afirmar com segurança qual é o efeito que realmente é importante nos compostos substituídos na posição para. Além disso, (b) o sinal do coeficiente de r_v^p na eq 2 possui sinal positivo. Isso indica que o efeito estéreo do substituinte intensifica a atividade, o que raramente é observado. Nos casos em que o aumento do tamanho do substituinte intensifica a atividade, a propriedade relevante é a hidrofobicidade e não o efeito estéreo[7] e; (c) utilizou-se apenas o efeito eletrônico dos

*e-mail: anderson@cce.ufes.br.

substituintes na posição meta. Segundo Unger e Hansch[1], isso não está em acordo com o mecanismo de ação proposto para esses compostos, cuja etapa limitante é a interação entre o carbocátion (**2**), produzido rapidamente através da hidrólise do fármaco (**1**) em pH fisiológico, e o provável ambiente nucleofílico localizado no sítio de ação (representado por Z^-).



1 **2**

Unger e Hansch[1] imaginaram que se esse mecanismo de ação estivesse correto, então a constante eletrônica σ^+, apropriada para substituintes capazes de deslocalizar uma carga eletrônica residual positiva, deveria ser mais adequada do que σ. De fato, essa hipótese pôde ser verificada através da eq 3, que claramente possui melhor ajuste do que a eq 1.

$$\log 1/C = 1{,}15\ \pi - 1{,}47\ \sigma^+ + 7{,}82 \quad (3)$$
$$(n = 22;\ R = 0{,}944;\ s = 0{,}197)$$

Nesse mesmo artigo, Unger e Hansch[1] estabeleceram cinco regras gerais para a proposição de modelos matemáticos de relações estrutura-atividade, que são enunciadas a seguir.

(a) Seleção de variáveis independentes: deve-se testar grande número de variáveis, incluindo propriedades de natureza lipofílica, eletrônica, estérea e de polarizabilidade[7,8]. Também devem ser testadas variáveis geradas a partir de cálculos de mecânica quântica[9] e variáveis indicadoras[10]. As variáveis selecionadas na *melhor equação* devem ser essencialmente independentes;

(b) Validação estatística das variáveis selecionadas: cada variável incluída na *melhor equação* precisa ser validada por testes estatísticos apropriados, tais como o teste F, o teste t para os coeficientes de cada variável, etc.;

(c) Princípio da parcimônia (Navalha de Occam): quando houver dúvida na escolha de um entre muitos modelos (aproximadamente) equivalentes, deve-se escolher o mais simples;

(d) Número de variáveis em cada modelo: para minimizar a ocorrência de correlação por coincidência, deve haver, no mínimo, cerca de cinco ou seis compostos para cada variável incluída no modelo;

(e) Modelo qualitativo para o mecanismo de ação dos compostos: é essencial que o modelo quantitativo de relação entre estrutura e atividade seja consistente com o mecanismo de ação, em nível molecular, dos compostos testados.

A idéia por detrás dessas regras era disciplinar a metodologia de elaboração de modelos de QSAR para que essa área de conhecimento, cristalizada por Hansch e colaboradores apenas nove anos antes, não caísse em descrédito pela má utilização dos modelos matemáticos. Apesar disso, o que se tem observado na literatura internacional é que a maioria dos modelos publicada ano após ano é criada sem que essas regras sejam integralmente aplicadas. Acredita-se que há alguns motivos predominantes que colaboram com esse estado de coisas: (a) embora a matemática envolvida na elaboração de um modelo de QSAR seja trivial (regressão linear múltipla), os pressupostos para sua aplicação e a interpretação de suas conseqüências não os são. Além disso, (b) ajustar um conjunto de dados a um modelo linear é fácil, porém, proceder a um conjunto de testes estatísticos consistentes para fazer sua validação requer algum conhecimento de estatística. Muitos químicos medicinais não possuem esse conhecimento. (c) Pelo fato de lidar com modelos matemáticos muito simples, a área de QSAR costuma atrair grande número de entusiastas que vêem na regressão linear (e no seu coeficiente de correlação) a ferramenta ideal para produzir publicações fáceis. A falta de experiência em química medicinal certamente pode dificultar a interpretação apropriada dos modelos criados.

Este trabalho tem como objetivo principal esclarecer alguns aspectos teóricos e, principalmente, práticos sobre proposição, validação e análise dos modelos matemáticos de QSAR. Pretende-se analisar as principais regras de proposição de modelos de QSAR à luz da estatística e analisar alguns exemplos ilustrativos. Assim, espera-se que os alunos e pesquisadores da área de QSAR, especialmente aqueles ainda inexperientes, possam solidificar seu embasamento nessa área, sejam capazes de adotar postura mais crítica em relação aos trabalhos publicados na área de QSAR e, eventualmente, possam melhorar a consistência dos modelos matemáticos que venham a propor.

METODOLOGIA

Os cálculos envolvidos na construção e análise dos modelos de regressão presentes neste trabalho foram executados através do programa Build QSAR, desenvolvido no Departamento de Física da UFES[11].

REGRESSÃO LINEAR MÚLTIPLA

Nos diversos ramos da ciência, freqüentemente deseja-se estabelecer relações quantitativas entre um fenômeno observado e algumas variáveis independentes que se acreditam ter relevância na explicação do fenômeno. Em outras palavras, deseja-se construir um modelo matemático que seja capaz de explicar o fenômeno observado e que também seja capaz de proporcionar previsões dentro e, se possível, fora dos limites investigados. Em QSAR, o fenômeno observado é a atividade biológica e as variáveis independentes são propriedades de natureza lipofílica, eletrônica, estérea e polar. Acreditando-se que essas propriedades sejam relevantes na explicação do nível de atividade biológica, procura-se construir um modelo matemático que estabeleça relação quantitativa entre essas grandezas. O modelo de Hansch-Fujita[7, 8, 12-14] propõe que a medida quantitativa da atividade farmacológica ou toxicológica, genericamente designada de atividade biológica, de uma série de compostos pode ser correlacionada às suas propriedades físico-químicas e estruturais através de um modelo multidimensional linear (eq 4) ou quadrático (eq 5).

$$\log 1/C = a\ X_{Lipofílico} + b\ X_{Eletrônico} + c\ X_{Estéreo} + d\ X_{Polar} + e \quad (4)$$

$$\log 1/C = -a\ X_{Lipofílico}^2 + b\ X_{Lipofílico} + c\ X_{Eletrônico} + d\ X_{Estéreo} + e\ X_{Polar} + f \quad (5)$$

Nessas equações, C é a concentração molar de cada composto capaz de produzir resposta biológica definida (tais como IC50, a concentração molar do fármaco capaz de proporcionar 50% de inibição da atividade fisiológica de um sistema biológico, como por exemplo, a atividade catalítica de uma enzima; LD100, a concentração molar do fármaco capaz de matar 100% dos indivíduos em que é administrado; ED50, concentração molar do fármaco capaz de produzir 50% de seu efeito máximo; etc.), os símbolos X's são variáveis que representam as propriedades físico-químicas e estruturais locais (constantes de substituintes) ou globais (propriedades moleculares) de cada composto analisado e os símbolos *a-f* são coeficientes de ajuste. Embora a

eq 5 seja não linear, o método de obtenção dos seus coeficientes é o mesmo utilizado para a obtenção dos coeficientes dos modelos lineares (eq 4).

O modelo linear é uma combinação linear de variáveis independentes, também chamadas *explicativas*, X_1, X_2, ..., X_k, capaz de reproduzir da melhor forma possível os valores experimentais de um grupo de n observações do fenômeno Y (eq 6).

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon \quad (6)$$

Na eq 6, β_0 é o termo constante de ajuste, β_1, β_2,..., β_k são os coeficientes das variáveis independentes e ε é o erro associado ao modelo. Em estatística, β_0, β_1, ..., β_k são chamados de *parâmetros*. Em QSAR, a designação *parâmetro* costuma ser atribuída às variáveis independentes, como por exemplo parâmetro lipofílico, π, parâmetro eletrônico, σ, etc. Neste trabalho, restringir-se-á o uso do termo *parâmetro* às constantes β_0, β_1, ..., β_k, enquanto que os termos b_o, b_1, ..., b_k (ver abaixo) serão referenciados como estimativas dos parâmetros ou simplesmente coeficientes da regressão.

Na eq 6, são conhecidos apenas os valores de X_1, X_2, ..., X_k e Y e não os de ε. A natureza estocástica do modelo de regressão implica que, para cada valor X_{1i}, X_{2i}, ..., X_{ki}, em que o índice i refere-se ao i-ésimo objeto (composto) incluído no modelo, haja uma distribuição de probabilidade total para os valores de Y. Isto significa que uma dada observação Y_i nunca poderá ser exatamente prevista. A incerteza relativa a Y surge por causa da presença do erro ε.

A eq 6, que também poderíamos chamar de *verdadeiro modelo de regressão*, é exata no sentido de que se os coeficientes β e o erro ε forem conhecidos, o modelo será capaz de reproduzir exatamente o valor observado Y. No entanto, a determinação exata dos valores de β só pode ser feita se todos os possíveis valores de Y forem incluídos no modelo, o que é uma tarefa muito difícil. Em QSAR, isso significaria incluir no modelo todos os compostos com alguma atividade sobre o sistema biológico em estudo. Na prática isso parece inexeqüível, pois de antemão não é possível saber quantos compostos, conhecidos e desconhecidos, apresentam atividade sobre um dado sistema. Além disso, a determinação do erro ε é tarefa muito difícil porque os fatores que contribuem para o seu valor são irregulares, tais como possíveis erros aleatórios inerentes ao fenômeno observado, erros experimentais na medida de Y e X (apesar dos valores de X serem supostamente isentos de erro, na prática não o são) e a própria qualidade do ajuste do modelo, como a ausência de variável explicativa importante. Portanto, na prática os parâmetros verdadeiros da eq 6 permanecerão desconhecidos. Tudo o que se pode fazer é obter uma *estimativa do modelo* através da análise de uma amostra do conjunto de todos os objetos. Em QSAR, isso significa analisar um pequeno subconjunto de compostos, dentre os incontáveis compostos, conhecidos e desconhecidos, que apresentam alguma atividade sobre o sistema biológico em estudo, para construir uma estimativa do modelo que somente seria conhecido se todos aqueles compostos fossem efetivamente analisados.

Apesar de, tecnicamente, o termo correto para referirem-se às equações de regressão, tais como as eqs. 1, 2 e 3, seja *estimativa do modelo*, é usual referirem-se a essas equações apenas como *modelos*. Neste trabalho os autores não se esforçarão em diferenciar esses termos. É importante salientar que, em QSAR, a designação de modelo ou estimativa de modelo é reservada para as equações de regressão que realmente representem alguma relação entre estrutura e atividade em que as regras de proposição de modelos de Unger e Hansch[1] tenham sido observadas.

A estimativa do modelo é uma equação capaz de fornecer *valores previstos* para Y, que são geralmente representados por $\hat{Y}$ (eq 7).

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + ... + b_k X_k \quad (7)$$

Nesta equação, b_0, b_1, ..., b_k são estimativas para os valores dos parâmetros β_0, β_1, ..., β_k, respectivamente. A construção da estimativa do modelo, representada pela eq 7, requer a aplicação do método dos mínimos quadrados, ou MMQ. Este consiste em encontrar o conjunto de valores b_0, b_1, ..., b_k capaz de minimizar os desvios (ao quadrado) entre cada um dos valores observados, Y_i, e os respectivos valores previstos, $\hat{Y}_i$. Ou seja, o somatório $\sum_{i=1}^{n}\left(Y - \hat{Y}_i\right)^2$ deve ser minimizado. A metodologia para a determinação das estimativas dos parâmetros *b*'s pode ser encontrada em livros-texto básicos de estatística[15-21] e não será discutida aqui. Entretanto, é preciso destacar alguns aspectos importantes da construção de modelos através do MMQ.

A obtenção do modelo representado pela eq 7 inicia-se com a construção de um conjunto de dados contendo uma amostra de n observações, ou objetos, e m variáveis explicativas X (Quadro 1). Em QSAR, isso significa selecionar uma amostra de n compostos, determinar experimentalmente as respectivas atividades Y e escolher um conjunto de m descritores físico-químicos e estruturais que se acreditam ser capazes de explicar a atividade biológica observada. O símbolo m refere-se ao número de descritores presentes no conjunto de dados, enquanto que o símbolo k (eq. 6) refere-se ao número de descritores efetivamente incluídos nos modelos de QSAR.

Quadro 1. Conjunto de dados, contendo n objetos e m propriedades descritivas, necessário à construção de modelo linear semelhante à eq. 7.

Y	X_1	X_2	...	X_j	...	X_m
Y_1	$X_{1,1}$	$X_{1,2}$	...	$X_{1,j}$	...	$X_{1,m}$
Y_2	$X_{2,1}$	$X_{2,2}$	...	$X_{2,j}$	...	$X_{2,m}$
⋮	⋮	⋮	⋱	⋮	⋱	⋮
Y_i	$X_{i,1}$	$X_{i,2}$	...	$X_{i,j}$	...	$X_{i,m}$
⋮	⋮	⋮	⋱	⋮	⋱	⋮
Y_n	$X_{n,1}$	$X_{n,2}$	...	$X_{n,j}$	...	$X_{n,m}$

A construção de modelos lineares compreende alguns pressupostos básicos em relação aos componentes do modelo: (a) os valores de X_1, X_2, ..., X_m são fixos, isto é, X_1, X_2, ..., X_m não são variáveis aleatórias. Apesar de muitas das variáveis utilizadas em QSAR originarem-se de medidas experimentais, como π e σ, o erro associado à medida ou ao cálculo dos valores dessas variáveis é, em geral, considerado muito menor do que o erro associado à medida da atividade biológica; (b) o erro ε_i tem distribuição de probabilidade normal; (c) a média de ε_i é igual a zero; (d) para um dado conjunto de valores X_{1i}, X_{2i}, ..., X_{ni}, a variância do erro ε_i é sempre constante; (e) o erro de uma observação é não-correlacionado com o erro de outra observação; (f) duas variáveis independentes quaisquer, X_i, X_j, são não correlacionadas, com $i \neq j$.

Um aspecto importante na construção do conjunto de dados é o comportamento de Y_i, o valor observado do i-ésimo objeto. Caso fossem feitas diversas medidas experimentais de Y_i, dificilmente haveria muitas coincidências. O erro experimental do processo de medição faz com que cada medida resulte num valor ligeiramente diferente de Y_i, ficando todos esses valores agrupados em torno de sua média, $\overline{Y}_i$. Um dos pilares do MMQ pressupõe que os valores obtidos em diversas medições de Y_i apresentam distribuição normal em torno de $\overline{Y}_i$ (Figura 1). Em QSAR, isso significa dizer que a execução de diversas medidas da atividade de um dado composto resultaria numa coleção de valores que apresentaria distribuição normal em torno de sua média. Apesar dessa suposição ser razoável, raramente vê-se comprovação experimental da distribuição normal dos valores

de Y_i. Quando muito, a atividade biológica de cada composto é medida em triplicata, o que não é suficiente para observar qualquer possível padrão de distribuição. Em geral, a aceitação da hipótese da distribuição normal dos valores de Y_i é decorrente da validação estatística e bioquímica dos modelos de QSAR.

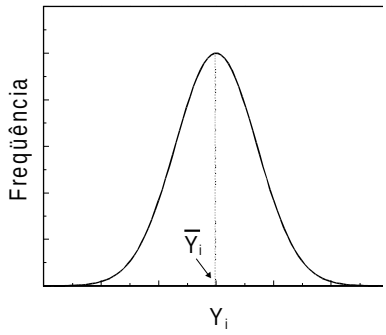


Figura 1, *Distribuição dos valores obtidos em n_i medidas experimentais de Y_i, o i-ésimo valor de Y. Esses valores devem apresentar distribuição normal em torno de sua média para que $\overline{Y}_i$ possa ser apropriadamente utilizado em RLM.*

É importante observar que o fato de se tentar descrever um conjunto de observações experimentais através de um modelo linear não significa que essas observações possam ser bem descritas através desse modelo. A descrição de um conjunto de observações a um modelo linear, bem como a qualquer outro tipo de modelo, é feito por hipótese. Imaginando-se que as observações possam ser descritas por dado modelo linear, cria-se a hipótese "as observações podem ser adequadamente descritas pelo modelo linear". Porém, acreditar-se numa hipótese não a torna necessariamente verdadeira. É preciso testá-la. Uma vez construído o modelo, é preciso submetê-lo a testes para verificar a veracidade da hipótese em que o mesmo está fundamentado.

AVALIAÇÃO DE MODELOS LINEARES

A avaliação consiste em verificar se a especificação do modelo adapta-se convenientemente aos dados observados. A avaliação do modelo pode ser dividida em três partes: (a) avaliação do grau de ajuste; (b) avaliação do grau de significância e; (c) avaliação do grau de previsibilidade.

Avaliação do grau de ajuste

O grau de ajuste do modelo é medido em termos de sua capacidade de reproduzir o valor observado dos objetos. Essa parte da avaliação é feita através do cálculo do coeficiente de correlação (R), do coeficiente de correlação ajustado (R_{Ajust}), que permite comparações entre modelos com número diferente de variáveis, e do desvio-padrão (s), além da análise dos resíduos $\left(Y_i - \hat{Y}_i\right)$. O que se espera de um modelo em relação ao grau de ajuste é que ele apresente R o mais próximo possível de 1, que o valor de s seja o mais próximo possível de zero e que os resíduos apresentem distribuição normal em torno de zero.

A avaliação do ajuste do modelo pode ser feita através da *análise da variância* (ANOVA) da regressão. Será feita breve pausa na análise da avaliação dos modelos de regressão para detalhar o conteúdo da ANOVA.

Análise da variância

Os principais objetivos da análise da variância são (a) verificar se há falta de ajuste no modelo (*lack of fit*); (b) obter estimativa correta para a variância do modelo de regressão (s^2) e; (c) estimar o grau de ajuste e significância do modelo. A análise da variância ajuda a compreender o significado de alguns dos termos que aparecem numa equação de regressão, como por exemplo R, s e F. A ANOVA é geralmente apresentada em forma de tabela e é construída com base nos valores de Y (observado), $\hat{Y}$ (previsto) e $\overline{Y}$ (média global dos valores de Y). A Figura 2 mostra como essas grandezas estão relacionadas para o i-ésimo objeto de um conjunto de dados.

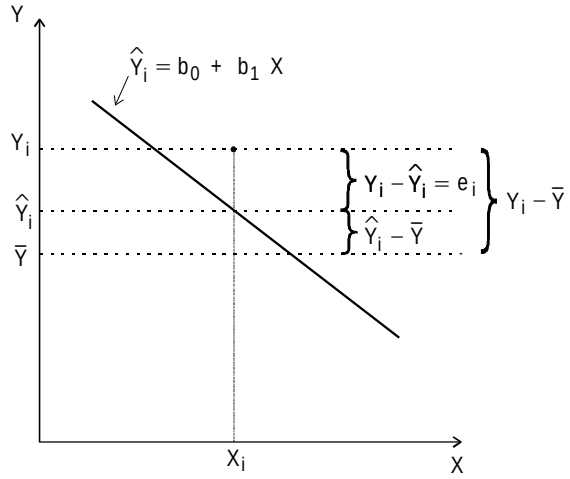


Figura 2. *Representação gráfica do i-ésimo objeto (x_i, y_i) de um conjunto de dados, do valor previsto desse objeto obtido pelo método dos mínimos quadrados ($\overline{Y}_i$) e da média dos valores observados de Y ($\overline{Y}$).*

De acordo com a Figura 2, é válida a identidade representada pela eq 8.

$$\left(Y_i - \overline{Y}\right) = \left(\hat{Y}_i - \overline{Y}\right) + \left(Y_i - \hat{Y}_i\right) \tag{8}$$

Nesta equação, $\left(Y_i - \overline{Y}\right)$ é o desvio do i-ésimo valor observado de Y em relação à média de todos os valores de Y, $\left(\hat{Y}_i - \overline{Y}\right)$ é o desvio do i-ésimo valor previsto de Y em relação à média dos seus valores e $\left(Y_i - \hat{Y}_i\right)$ é o desvio entre o i-ésimo valor observado de Y e o seu respectivo valor previsto, também chamado de i-ésimo resíduo, ou e_i $\left(e_i = Y_i - \hat{Y}_i\right)$.

Pode ser demonstrado[15] que se ambos os membros da eq 8 forem elevados ao quadrado e seus termos forem somados de i = 1, 2, …, n, o resultado será dado pela eq 9.

$$\sum\left(Y_i - \hat{Y}_i\right)^2 = \sum\left(\hat{Y}_i - \overline{Y}_i\right)^2 + \sum\left(Y_i - \hat{Y}_i\right)^2 \tag{9}$$

A eq 9 também pode escrita da seguinte forma: $SS_{Tot} = SS_{Reg} + SS_{Res}$, em que a abreviação SS refere-se à soma dos quadrados dos desvios (*sum of squares*). O termo SS_{Tot} é a variabilidade total da regressão, SS_{Reg} é a variabilidade explicada pelo modelo de regressão e SS_{Res} é a variabilidade que o modelo não consegue explicar e refere-se aos resíduos.

O esquema simplificado da ANOVA é mostrado no Quadro 2.

O Quadro 2 mostra que o quadrado do coeficiente de correlação, $R^2 = SS_{Reg}/SS_{Tot}$, corresponde à fração da variabilidade total que é explicada pelo modelo. Por exemplo, um modelo de QSAR em que R^2 = 0,9 é dito capaz de explicar 90% da variabilidade total dos valores observados da atividade biológica, em torno de sua média $\overline{Y}$. Pode-se definir a média da soma dos quadrados da regressão como $MS_{Reg} = SS_{Reg}/k$ e a média da soma dos quadrados dos resíduos como $MS_{Res} = SS_{Res}/(n\text{-}k\text{-}1)$.

Quadro 2. Análise da variância (ANOVA) do modelo de regressão linear múltipla.

Fonte	[a] *df*	[b] *SS*	[c] *MS*	
Regressão	k	$\sum\left(\hat{Y}_i - \overline{Y}\right)^2$	SS_{Reg}/df_{Reg}	$F = MS_{Reg}/s^2$
Resíduo	$n - k - 1$	$\sum\left(Y_i - \hat{Y}_i\right)^2$	$s^2 = SS_{Res}/df_{Res}$	$R^2 = SS_{Reg}/SS_{Tot}$
Total	$n - 1$	$\sum\left(Y_i - \overline{Y}\right)^2$	SS_{Tot}/df_{Tot}	

[a]*df* = Graus de liberdade (*degrees of freedom*); [b]*SS* = Soma dos quadrados (*sum of squares*); [c]*MS* = Média da soma dos quadrados (*mean square*);

Utilizando-se esta notação, define-se o quadrado do desvio-padrão ou estimativa da variância como $s^2 = MS_{Res}$. Como s^2 é a razão entre a variabilidade não explicada pelo modelo e o número de graus de liberdade relativo aos resíduos da regressão, quanto maior for a variabilidade dos valores de Y que o modelo for capaz de explicar (maior R), menor será o desvio-padrão. O teste F é definido como $F = MS_{Reg}/s^2$, sendo portanto uma razão entre a variabilidade explicada pelo modelo e a variabilidade que permanece sem explicação. Um bom modelo deve apresentar o maior valor possível para F, sendo que o valor mínimo aceitável é dado por tabelas de referência que podem ser encontradas em livros-texto de estatística[15,17,19]. O quadrado do coeficiente de correlação ajustado, citado na Seção anterior, é calculado de acordo com a eq 10.

$$R^2_{Ajust} = R^2 - \left(\frac{k-1}{n-k}\right)\left(1 - R^2\right) \quad (10)$$

Há uma observação importante acerca da variância. A variância de um modelo de regressão, σ^2, somente poderá ser determinada se o verdadeiro modelo de regressão for construído. Como foi visto anteriormente, o verdadeiro modelo de regressão é o que inclui todos os possíveis compostos com atividade sobre o sistema biológico em questão (eq 6). Como isso nunca é possível, o valor de σ^2 sempre será desconhecido. Quando o modelo proposto é correto, a média da soma dos quadrados dos resíduos, s^2, é um estimador *não viesado* (não tendencioso) da verdadeira variância σ^2. Entretanto, quando o modelo não é adequado, s^2 estará estimando algo maior do que σ^2, pois na soma dos quadrados estarão incluídos os vieses devidos à inadequação do modelo.

A estimativa da variância, s^2, pode ser obtida através da construção da estimativa do modelo (eq 7). No entanto, a estimativa correta da variância somente é possível em modelos onde não houver falta de ajuste[15]. Em outras palavras, somente em modelos bem ajustados há possibilidade de obter-se a estimativa correta da variância. Neste ponto parece haver um paradoxo. O desvio-padrão do modelo, s, que é a raiz quadrada da estimativa da variância, é um critério de ajuste do modelo. No entanto, só poderemos saber se s^2 é a estimativa correta de σ^2, ou seja, se s tem significado, se não houver falta de ajuste no modelo. Pode-se romper este ciclo verificando-se, em primeiro lugar, a falta de ajuste do modelo proposto através da utilização dos resíduos da regressão.

Os resíduos de um modelo de regressão contêm toda a informação necessária à compreensão dos motivos que fazem com que o mesmo não consiga explicar 100% da variabilidade dos valores observados de Y. Existem basicamente dois motivos para que isso ocorra: (a) presença de erros aleatórios relativos à determinação experimental dos valores de Y e (b) especificação imprópria do modelo (falta de ajuste). Uma vez que os valores de Y tenham origem em medidas experimentais, os erros aleatórios estarão sempre presentes e, devido a isso, nenhum modelo consegue explicar 100% da variabilidade de Y. Por outro lado, a especificação do modelo é responsabilidade de quem o constrói. A especificação do modelo diz respeito à sua forma final, ou seja, se é linear, parabólico, exponencial, se contém termos cruzados, se o número de termos presentes é adequado, etc. Portanto, deve haver especial cuidado na verificação da falta de ajuste.

Existem duas situações que devem ser bem caracterizadas em relação à verificação da falta de ajuste do modelo. A primeira é quando cada valor Y_i presente no conjunto de dados foi determinado uma única vez, ou seja, quando cada valor de Y_i for o resultado de uma *medida de ponto único*. Neste caso, a verificação da falta de ajuste pode ser feita qualitativamente através da análise da distribuição dos resíduos do modelo. Nos casos em que o modelo é bem ajustado, o conjunto de resíduos e_i contém apenas os erros aleatórios citados anteriormente. Portanto, a análise visual gráfica dos resíduos deverá revelar um padrão estritamente aleatório para a distribuição dos mesmos. Quando o modelo apresenta falta de ajuste, além dos erros aleatórios, os resíduos contêm erros sistemáticos devidos à especificação incorreta do modelo. A presença desses erros pode ser detectada com relativa facilidade na análise visual da distribuição dos resíduos. A Figura 3 mostra quatro situações típicas encontradas na verificação qualitativa da falta de ajuste de modelos lineares[15].

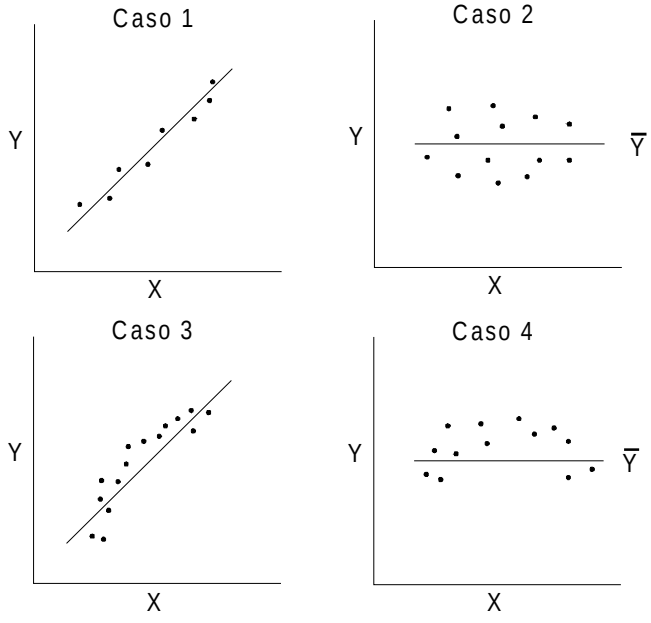


Figura 3. *Casos típicos de distribuição de pontos encontrados na investigação qualitativa de falta de ajuste em modelos de regressão linear.*

No Caso 1 não há falta de ajuste pois os pontos estão dispostos aleatoriamente ao longo da reta ajustada. Portanto, o modelo $\hat{Y} = b_0 + b_1 X$ deverá ser adequado aos dados observados. O Caso

2 também não revela falta de ajuste. No entanto, o modelo de regressão $\hat{Y} = b_0 + b_1X$ não apresentará significância estatística. Neste caso, o modelo $Y = \overline{Y}$ será mais adequado. No Caso 3 observa-se clara falta de ajuste devido ao padrão não aleatório da distribuição dos pontos e, portanto, dos resíduos. O modelo $\hat{Y} = b_0 + b_1X + b_{11}X^2$ deverá representar adequadamente os dados observados. De forma semelhante, no Caso 4 há falta de ajuste, sendo que o modelo $\hat{Y} = b_0 + b_1X + b_{11}X^2$ também poderá ajustar-se adequadamente aos dados observados.

A segunda situação é quando os valores de Y_i presentes no conjunto de dados foram determinados em replicata (duplicata, triplicata, etc.). Em QSAR, isso significa fazer duas ou mais medidas experimentais da atividade biológica para cada composto da série. Neste caso, as repetições das medidas de Y_i podem ser utilizadas para obter a estimativa da variância do modelo. Tal estimativa representa o chamado *erro puro*, pois se o conjunto de valores X_{i1}, X_{i2}, …, X_{ik} (Quadro 1) é o mesmo para duas ou mais observações, somente erros aleatórios podem influenciar os valores de Y_i e gerar diferenças entre eles. Essas diferenças podem proporcionar estimativa da variância mais confiável do que qualquer outra fonte de informação[15]. Nos casos em que os valores de Y_i forem determinados em replicata, o termo SS_{Res} pode ser dividido em duas partes: a soma dos quadrados devida ao erro puro do modelo, SS_e, e a soma dos quadrados devida à falta de ajuste do modelo, SS_{LOF}, sendo que $SS_{Res} = SS_e + SS_{LOF}$. O termo SS_{LOF} é determinado por diferença. O cálculo de SS_e é feito de acordo com a eq 11.

$$SS_e = \sum_{i=1u=1}^{n_X} \sum^{n_i} \left(Y_{iu} - \overline{Y}_i\right)^2 \qquad (11)$$

Na eq 11, Y_{iu} é a u-ésima repetição (u = 1, 2, …, n_i) da medida de Y_i para X_{i1}, X_{i2}, …, X_{ik}, n_X é o número de diferentes valores para X_1, X_2, …, X_k, que é equivalente ao número de diferentes objetos, n_i é o número de repetições feitas para Y_i e $\overline{Y}_i$ é a média das repetições Y_{i1}, Y_{i2}, Y_{in_i}….

O esquema da ANOVA, incluindo o teste para falta de ajuste, é mostrado no Quadro 3.

A verificação da falta de ajuste de um modelo de regressão é feita através da construção da tabela ANOVA, incluindo o teste de falta de ajuste, e verificação da significância estatística do valor encontrado para F_{LOF}. Para que o valor de F_{LOF} seja considerado significante, deverá ser maior ou igual ao respectivo valor de referência (nível de confiança de 95%), $F_{(df_{LOF}, df_e)}$, que pode ser encontrado em livros-texto de estatística[15,17,19]. Se F_{LOF} for significante, então há falta de ajuste no modelo construído e outro tipo de modelo deverá ser testado.

Devido à sua relevância, torna-se importante ilustrar esse assunto com um exemplo prático. A Tabela 1 contém um conjunto de dados com seis derivados do 2-bromo-etanoato, cuja atividade bactericida em *B. diphtheriae* foi medida em replicata. Os valores médios da atividade (log 1/*C*) e os valores do coeficiente de partição (log *P*) foram extraídos da literatura[22]. Os valores das repetições de log 1/*C* são fictícios, uma vez que, quando feitos, raramente são publicados. Na maioria dos casos, os valores da atividade biológica que aparecem nos conjuntos de dados referem-se às médias das repetições. Portanto, o conjunto de dados habitualmente encontrado na literatura incluiria apenas a média de log 1/*C* e log *P* (colunas log $1/C_{Média}$ e log P, Tabela 1). A Tabela 2 mostra a ANOVA para o modelo de regressão log 1/*C* = b_0 + b_1 log *P*, construída de acordo com o Quadro 2, em que foram considerados apenas os valores médios de log 1/C.

Tabela 1. Atividade bactericida de 2-bromo-etanoatos substituídos ($RCHBrCOO^-$) em *B. diphtheriae* (log 1/*C*) em função do logaritmo de seu coeficiente de partição octanol-água (log *P*)[22]. Os valores numéricos apresentados na coluna *Repetições* são fictícios.

R	log 1/*C*		log *P*
	Repetições	Média	
$CH_2(CH_2)_6CH_3$	1,32 1,53 1,95	1,60	0,32
$CH_2(CH_2)_8CH_3$	2,25 2,15	2,20	1,32
$CH_2(CH_2)_{10}CH_3$	2,39 2,21 2,79 2,61	2,50	2,32
$CH_2(CH_2)_{12}CH_3$	3,15 3,84 3,24	3,41	3,32
$CH_2(CH_2)_{14}CH_3$	3,98 4,65 4,17 4,44	4,31	4,32
$CH_2(CH_2)_{16}CH_3$	4,32 3,91 3,80	4,01	5,32

Quadro 3. Análise da variância (ANOVA), incluindo o teste para falta de ajuste, do modelo de regressão linear múltipla.

Fonte	[a] *df*	[b] *SS*	[c] *MS*	
Regressão	k	$\sum\left(\hat{Y}_i - \overline{Y}\right)^2$	SS_{Reg}/df_{Reg}	$F = MS_{Reg}/s^2$
Resíduo	$n - k - 1$	$\sum\left(Y_i - \hat{Y}\right)^2$	$s^2 = SS_{Res}/df_{Res}$	$R^2 = SS_{Reg}/SS_{Tot}$
Falta de ajuste(LOF)	$n-k-1-\sum_{i=1}^{n}(n_i-1)$	$\sum_{i=1}^{n_X} n_i\left(\hat{Y}_i - \overline{Y}_i\right)^2$	SS_{LOF}/df_{LOF}	[a] $F_{LOF} = MS_{LOF}/s_e^2$
Erro puro (e)	$\sum_{i=1}^{n}(n_i-1)$	$\sum_{i=1u=1}^{n_X}\sum^{n_i}\left(Y_{iu} - \overline{Y}_i\right)^2$	[b] $s_e^2 = SS_e/df_e$	
Total	$n - 1$	$\sum\left(Y_i - \overline{Y}\right)^2$	SS_{Tot}/df_{Tot}	

[a]F_{LOF} = Teste de significância para a falta de ajuste do modelo. Se F_{LOF} for significativo em relação ao nível de confiança estipulado, por exemplo 95%, há falta de ajuste no modelo; [b]s_e^2 = Estimativa correta da variância do modelo, caso F_{LOF} seja não significativo;

Tabela 2. Análise da variância (ANOVA) do modelo de regressão linear múltipla log $1/C = b_0 + b_1 \log P$ referente ao conjunto de dados mostrado na Tabela 1 ($n = 6$; $k = 1$). Foram utilizados os valores médios de log $1/C$ na construção da ANOVA.

Fonte	*df*	*SS*	*MS*	
Regressão	1	5,1955	5,1955	F = 47,4067
Resíduo	4	0,4384	s^2 = 0,1096	R^2 = 0,9222
Total	5	5,6339	1,1268	

Analisando-se apenas os números que aparecem na Tabela 2, pode-se acreditar que log $1/C = b_0 + b_1 \log P$ é um excelente modelo de regressão, pois $R = 0,96 \approx 1$, s é pequeno e $F = 47,4$ é grande, comparado com o valor de referência $F_{(1,4)} = 7,71$ (nível de confiança de 95%). No entanto, utilizando-se as replicatas de log $1/C$ para construir a ANOVA, de acordo com o Quadro 3, o resultado é bem diferente (Tabela 3).

Tabela 3. Análise da variância (ANOVA), incluindo o teste de falta de ajuste, do modelo de regressão linear múltipla log $1/C = b_0 + b_1 \log P$ referente ao conjunto de dados mostrado na Tabela 1 ($n = 19$; $k = 1$). Foram utilizados os valores repetidos de log $1/C$ na construção da ANOVA.

Fonte	*df*	*SS*	*MS*	
Regressão	1	16,8569	16,8569	F = 107,0940
Resíduo	17	2,6758	s^2 = 0,1574	R^2 = 0,8630
Falta de ajuste	4	1,5800	0,3950	F_{LOF} = 4,6862
Erro puro	13	1,0958	s_e^2 = 0,0843	
Total	18	19,5327	1,0852	

Na construção da Tabela 3, $SS_e = SS_{e1} + SS_{e2} + \ldots + SS_{e6}$, em que $SS_{e1} = (1,32^2 + 1,53^2 + 1,95^2) - 3 \times [(1,32 + 1,53 + 1,95) / 3]^2 = 0,2058$, com 3 - 1 = 2 graus de liberdade; $SS_{e2} = (2,25^2 + 2,15^2) - 2 \times [(2,25 + 2,15) / 2]^2 = 0,0050$, com 2 - 1 = 1 grau de liberdade; $SS_{e3} = (2,39^2 + 2,21^2 + 2,79^2 + 2,61^2) - 4 \times [(2,39 + 2,21 + 2,79 + 2,61) / 4]^2 = 0,1924$, com 4 - 1 = 3 graus de liberdade; $SS_{e4} = (3,15^2 + 3,84^2 + 3,24^2) - 3 \times [(3,15 + 3,84 + 3,24) / 3]^2 = 0,2814$, com 3 - 1 = 2 graus de liberdade; $SS_{e5} = (3,98^2 + 4,65^2 + 4,17^2 + 4,44^2) - 4 \times [(3,98 + 4,65 + 4,17 + 4,44) / 4]^2 = 0,2610$, com 4 - 1 = 3 graus de liberdade e; $SS_{e6} = (4,32^2 + 3,91^2 + 3,80^2) - 3 \times [(4,32 + 3,91 + 3,88) / 3]^2 = 0,1502$, com 3 - 1 = 2 graus de liberdade. Logo $SS_e = 0,2058 + 0,0050 + 0,1924 + 0,2814 + 0,2610 + 0,1502 = 1,0958$ e $df_e = 2 + 1 + 3 + 2 + 3 + 2 = 13$. Portanto, $SS_{LOF} = SS_{Res} - SS_e = 1,6540 - 0,0740 = 1,5800$ e $df_{LOF} = df_{Res} - df_e = 17 - 13 = 4$. Finalmente, $MS_{LOF} = 1,5800 / 4 = 0,3950$, $s_e^2 = 1,0958 / 13 = 0,0843$ e $F_{LOF} = 0,3950 / 0,0843 = 4,6862$.

Na Tabela 3, $s_e^2 = 0,0843$ seria a estimativa correta da variância da regressão, caso não houvesse falta de ajuste no modelo testado (log $1/C = b_0 + b_1 \log P$). No entanto, F_{LOF} é maior do que o valor de referência, $F_{(4,\ 13)} = 3,18$, indicando haver falta de ajuste no modelo. A explicação para a existência de falta de ajuste neste modelo decorre principalmente da atividade biológica do composto R = $CH_2(CH_2)_{16}CH_3$, cujo valor, log $1/C = 4,01$, sofreu quebra de linearidade em relação aos compostos anteriores. Essa diminuição repentina da atividade, por sua vez, é conseqüência da elevada lipossolubilidade do composto, que faz com que haja dificuldades no transporte e biodisponibilidade do fármaco. Compostos com alta lipossolubilidade tendem a ficar retidos nas membranas celulares que precisam atravessar para atingirem a biofase.

Devido à falta de ajuste verificada no modelo log $1/C = b_0 + b_1 \log P$, outro tipo de modelo deveria ser testado, como por exemplo o parabólico[23], log $1/C = b_0 + b_1 \log P + b_{11} \log P^2$, ou o bilinear[24], log $1/C = b_0 + b_1 \log P + b_{11} \log (\beta P + 1)$. No entanto, a inclusão de mais uma variável num modelo com tão poucos compostos (seis) aumenta a probabilidade de ocorrência de correlação por coincidência (regra *d*, citada na Introdução deste artigo).

Avaliação do grau de significância

O grau de significância é medido através da execução de testes de validação (teste estatístico de hipótese), sendo que cada teste destina-se a verificar a significância de diferentes partes do modelo.

Para testar a significância estatística de R^2, aplica-se um teste de hipótese conhecido como teste F, cujo valor é obtido na tabela ANOVA associada à regressão (Quadro 2). O teste F verifica o quanto da variabilidade de Y pode ser explicada pelas variáveis $X_1, X_2, \ldots, X_k$, e o quanto pode ser atribuída ao efeito do erro aleatório ε. Para validar R^2 através do teste F, é preciso comparar o valor de F obtido no modelo com o valor de referência. Este, em geral, se refere ao nível de confiança de 95% e pode ser obtido em tabelas apropriadas. Por exemplo, seja um modelo linear com as seguintes características: $n = 20$, $k = 3$, $R = 0,85$, $s = 0,32$ e $F = 12,54$. Para saber se o valor de R^2 possui ou não significância estatística, é preciso comparar o valor de F com o valor de referência que, neste caso, vale $F_{(k,n-k-1)} = F_{(3,16)} = 2,28$. Como $F > F_{(3,16)}$, então R^2 é significativo. Os valores do teste F de dois ou mais modelos de regressão, que possuam diferentes valores de n e k, em princípio não podem ser comparados. Por exemplo, sejam dois modelos lineares, M1 e M2, com as características: M1 ($n = 20$, $k = 3$, $R = 0,85$, $s = 0,32$, $F = 12,54$) e M2 ($n = 23$, $k = 4$, $R = 0,91$, $s = 0,30$, $F = 15,23$). Apesar de M2 apresentar as estatísticas R, s e F superiores em relação a M1, não é possível afirmar com segurança que M2 é mais significativo do que M1 apenas com base nessas informações. Nesse caso, deve-se calcular as probabilidades (p-valor) associadas aos valores de F. O p-valor fornece um meio seguro de comparação do nível de significância de modelos com diferentes números de objetos e variáveis. Um valor de $p = 0,0001$, significa que o valor de R^2 é estatisticamente significante e o erro envolvido na afirmação dessa hipótese é de 0,01%. Se para M1 $p = 0,0001$ e para M2 $p = 0,0005$, então M1 terá maior significância estatística do que M2.

A significância estatística dos coeficientes da regressão é testada mediante o cálculo de seus intervalos de confiança (T), geralmente referentes a um nível de confiança de 95% (t). O resultado do teste é mostrado em associação com o respectivo coeficiente (eq 12).

$$\hat{Y} = b_0 (\pm T_0) + b_1 (\pm T_1) X_1 + \ldots + b_i (\pm T_i) X_i + \ldots + b_k (\pm T_k) X_k \quad (12)$$

O valor de T_i é calculado de acordo com a eq 13,

$$T_i = s.t_{(n-k-1,95)} \cdot \sqrt{(X'X)^{-1}_{i,i}} \quad (13)$$

em que s é o desvio-padrão da regressão, $t_{(n-k-1,95)}$ é o valor da distribuição t de Student para a probabilidade de 95% e o argumento da raiz quadrada refere-se ao elemento diagonal (linha i, coluna i) da matriz resultante da operação indicada com a matriz das variáveis independentes, **X**. Se T_i for maior do que o valor do próprio coeficiente b_i, significa que o valor $b_i = 0$ pertence ao intervalo de confiança de 95% considerado. Isso implica em que a variável X_i, em relação à qual b_i está associada, não contribui para a explicação da variabilidade dos valores observados de Y. Naturalmente que quanto mais T_i se aproxima de b_i, menor será a significância estatística de b_i.

Avaliação do grau de previsibilidade

O grau de previsibilidade do modelo é testado através da validação cruzada (*cross validation*)[25-28]. O processo de validação cruzada consiste nas seguintes etapas: (a) excluir um dos objetos do modelo; (b) reconstruir o modelo sem esse objeto; (c) utilizar esse modelo para calcular o valor do objeto excluído; (d) obter o desvio entre o valor observado e o valor previsto para esse objeto; (e) refazer as etapas a-d para os demais objetos do conjunto de dados, um por vez; (f) calcular o valor da estatística *PRESS* (*PREdictive Sum of Squares)*, que corresponde à soma dos quadrados dos desvios obtidos no item d e; (g) calcular o quadrado do coeficiente de correlação da validação cruzada (Q^2) e o desvio-padrão da validação cruzada (s_{PRESS}).

Um modelo com elevado grau de previsibilidade para objetos não incluídos no mesmo apresentará Q^2 próximo de 1 e s_{PRESS} próximo de zero. A forma de calcular Q^2 e S_{PRESS} é mostrada nas eqs 14-16.

$$\mathrm{PRESS} = \sum\left(\mathrm{Y} - \hat{\mathrm{Y}}\right)^2 \tag{14}$$

$$\mathrm{Q}^2 = 1 - \frac{\mathrm{PRESS}}{\sum\left(\mathrm{Y} - \overline{\mathrm{Y}}\right)^2} \tag{15}$$

$$\mathrm{S}_{\mathrm{PRESS}} = \frac{\sqrt{\mathrm{PRESS}}}{\mathrm{n} - \mathrm{k} - 1} \tag{16}$$

ANÁLISE DAS REGRAS DE ELABORAÇÃO DE MODELOS

Nesta seção são analisadas as regras de elaboração de modelos e, sempre que possível, a análise de cada regra será acompanhada de exemplos ilustrativos de aplicação da mesma.

Seleção de variáveis independentes

Parece haver consenso no que diz respeito à utilização de grande número de variáveis explicativas (*m*) na construção do conjunto de dados. Essas variáveis devem abranger ampla gama de propriedades (lipofílica, eletrônica, estérea e polar). Além das constantes de substituintes utilizadas em QSAR clássico[7,8,29-33], devem-se incluir na análise propriedades físico-químicas moleculares tais como área superficial e volume moleculares[34], propriedades derivadas de cálculo de orbital molecular[9, 35-37], variáveis indicadoras[7,8,38], índices de similaridade[39-42] e índices topológicos[34]. A utilização de grandes conjuntos de dados em QSAR pressupõe a necessidade de algum tipo de método de seleção de variáveis, como por exemplo, a busca sistemática[43], as redes neurais[44-50], os algoritmos genéticos e evolucionários[43,45,51-56] e os métodos multivariados[25,45,52,57-60]. Estes métodos são utilizados para detectar combinações de variáveis capazes de fornecer equações de regressão com elevado coeficiente de correlação, baixo desvio-padrão ou elevado teste F, e que tenham algum potencial para tornarem-se modelos de QSAR. Embora o caminho entre uma equação de regressão e um modelo de QSAR seja relativamente longo, a seleção de variáveis é um dos primeiros passos nessa direção.

Alguns conjuntos de dados que podem ser destacados por sua dimensão são os de Supuran e Clare[61] (n = 28, k = 17), Mracec e colaboradores[62] (n = 49, k = 21), Kelder e Greven[63] (n = 55, k = 24), Menziani e colaboradores[64] (n = 29, k = 27), Gaudio[65] (n = 45, k = 37), Selwood[66] (n = 31, k = 53), Cocchi e colaboradores[67] (n = 40, k = 66) e Gaudio[68] (n = 36, k = 92).

Pode-se ilustrar o processo de seleção de variáveis aplicando-se o método da busca sistemática ao conjunto de dados que deu origem às eqs 2 e 3. Para isso, é necessário estimar os valores de *R*, *s*, *F* e *p* das equações de regressão da atividade biológica (log 1/*C*) em função de todas as possíveis combinações das variáveis π, π_m, σ^+, σ_m e r_v^p. Porém, ao recalcular modelos antigos, é importante fazer a revisão e a atualização dos valores das constantes de substituintes presentes no respectivo conjunto de dados. Sendo consistente com essa filosofia, construiu-se a Tabela 4, que apresenta valores revisados e atualizados para essas constantes de substituintes. Esses valores foram obtidos a partir de recentes compilações de constantes de substituintes[30, 69].

A execução da busca sistemática sobre o conjunto de dados da Tabela 4 gerou 31 equações de regressão, sendo cinco equações com uma variável, dez equações com duas variáveis, dez com três variáveis, cinco com quatro variáveis e uma equação com cinco variáveis. Os valores de *R*, *s*, *F* e *p* dessas combinações são mostrados na Tabela 5.

A melhor equação com uma variável é log $1/C = f(r_v^p)$, cuja avaliação é R = 0,878, s = 0,279, F = 67,06 e p < 0,000001 (No. 5, Tabela 5), que é capaz de explicar cerca de 77% da variabilidade da atividade. Como r_v^p é capaz de explicar a maior parte da variabilidade da atividade, é de esperar-se que ela também esteja presente nos melhores modelos com maior número de variáveis. Assim, a melhor equação com duas variáveis é log $1/C = f(\pi_m, r_v^p)$, cuja avaliação é R = 0,936, s = 0,210, F = 67,51 e p < 0,000001 (No. 12, Tabela 5). Da equação No. 5 para a No. 12 houve aumento do valor do coeficiente de correlação e diminuição do desvio-padrão. Para construir a melhor equação com três variáveis é necessário retirar π_m da equação No.12 e acrescentar as variáveis π e σ^+. O resultado dessa mudança é a equação log $1/C = f(\pi, \sigma^+, r_v^p)$, cuja avaliação é R = 0,963, s = 0,166, F = 76,32 e p< 0,000001 (No. 20, Tabela 5). A comparação dos valores de *R*, *s* e *F* das equações No. 12 e 20 indica que a regressão No. 20 é capaz de explicar maior quantidade da variabilidade de log 1/*C* do que a regressão No. 12, apesar daquela conter uma variável a mais do que esta. Como conseqüência, deve ser mais vantajoso representar a atividade biológica dos compostos da série através de uma equação com três variáveis do que com duas. O mesmo não pode ser dito ao considerarem-se as melhores equações com quatro e cinco variáveis. Os resultados da Tabela 5 mostram que os modelos com mais de três variáveis não são capazes de melhorar a explicação da atividade biológica em relação à equação No. 20. Dessa forma, o resultado da busca sistemática indica que a atividade dos compostos da série poderá ser representada por uma equação de três variáveis. Isso não quer dizer que essa equação seja a de No. 20, pois há outras equações com três variáveis que possuem avaliações equivalentes, como por exemplo as equações No. 16, 23 e 24. Avaliações mais aprofundadas deverão ser executadas sobre essas equações para decidir-se qual é a equação de melhor qualidade estatística.

É importante verificar o grau de correlação entre as variáveis ao proceder-se à seleção de variáveis. A construção de modelos através do MMQ exige que as variáveis presentes num dado modelo sejam essencialmente independentes. Além de descreverem a mesma propriedade, duas ou mais variáveis altamente correlacionadas geram problemas de dependência linear no conjunto de dados e imprecisão numérica na construção do modelo. É interessante frisar que a construção de modelos de QSAR através de métodos multivariados, como PCR (*Principal Component Regression*) e PLS (*Partial Least Squares*)[70], não é prejudicada pela presença de correlação elevada entre as variáveis.

O grau de correlação entre as variáveis é verificado através da construção da matriz de correlação. A matriz de correlação das variáveis independentes da Tabela 4 é mostrada na Tabela 6, que revela que as únicas variáveis que não devem ser combinadas numa mesma equação são σ^+ e σ_m, pois apresentam coeficiente de correlação igual a 0,702. No exemplo de seleção de variáveis acima (Tabela 5), as equações de três variáveis selecionadas para posterior estudo, ou seja Nos. 16, 20, 23 e 24, não incluem essas variáveis simultaneamente.

Tabela 4. Conjunto de dados utilizado por Unger e Hansch[1] na dedução da eq. Os valores publicados originalmente foram revisados e atualizados partir de recentes compilações de constantes de substituintes[30,69].

No	Substituinte	log $1/C_{Obs}$	π	π_m	σ^+	σ_m	r_v^p	log $1/C_{Calc}$[a]	Res.
1	H	7,46	0,00	0,00	0,00	0,00	1,20	7,80	-0,34
2	4-F	8,16	0,14	0,00	-0,07	0,00	1,47	8,05	0,11
3	4-Cl	8,68	0,71	0,00	0,11	0,00	1,75	8,48	0,20
4	4-Br	8,89	0,92	0,00	0,15	0,00	1,85	8,67	0,22
5	4-I	9,25	1,12	0,00	0,14	0,00	1,98	8,91	0,34
6	4-Me	9,30	0,58	0,00	-0,31	0,00	1,97	8,85	0,45
7	3-F	7,52	0,14	0,14	0,35	0,34	1,20	7,53	-0,01
8	3-Cl	8,16	0,71	0,71	0,40	0,37	1,20	8,12	0,04
9	3-Br	8,30	0,92	0,92	0,41	0,39	1,20	8,34	-0,04
10	3-I	8,40	1,12	1,12	0,36	0,35	1,20	8,63	-0,23
11	3-Me	8,46	0,58	0,58	-0,07	-0,07	1,20	8,55	-0,09
12	3-Cl, 4-F	8,19	0,85	0,71	0,33	0,37	1,47	8,36	-0,17
13	3-Br, 4-F	8,57	1,06	0,92	0,34	0,39	1,47	8,59	-0,02
14	3-Me, 4-F	8,82	0,72	0,58	-0,14	-0,07	1,47	8,80	0,02
15	3,4-Cl_2	8,89	1,42	0,71	0,51	0,37	1,75	8,79	0,10
16	3-Br, 4-Cl	8,92	1,63	0,92	0,52	0,39	1,75	9,02	-0,10
17	3-Me, 4-Cl	8,96	1,29	0,58	0,04	-0,07	1,75	9,23	-0,27
18	3-Cl, 4-Br	9,00	1,63	0,71	0,55	0,37	1,85	8,98	0,02
19	3,4-Br_2	9,35	1,84	0,92	0,56	0,39	1,85	9,21	0,14
20	3-Me, 4-Br	9,22	1,50	0,58	0,08	-0,07	1,85	9,42	-0,20
21	3,4-Me_2	9,30	1,16	0,58	-0,38	-0,07	1,97	9,60	-0,30
22	3-Br, 4-Me	9,52	1,50	0,92	0,10	0,39	1,97	9,39	0,13

[a] Calculado através da eq.

Tabela 5. Resultado da seleção de variáveis através de busca sistemática executada sobre o conjunto de dados mostrado na Tabela 4. Modelos com diferentes números de variáveis estão separados pelas linhas horizontais.

No.	π	π_m	σ^+	σ_m	r_v^p	R	s	F	$p <$
1	•					0,760	0,379	27,29	0,000041
2		•				0,206	0,570	0,88	0,358276
3			•			0,152	0,575	0,48	0,498423
4				•		0,134	0,577	0,37	0,552142
5					•	**0,878**	**0,279**	**67,06**	**0,000001**
6	•	•				0,844	0,320	23,57	0,000007
7	•		•			0,932	0,217	62,54	0,000001
8	•			•		0,874	0,290	30,82	0,000001
9	•				•	0,924	0,228	55,87	0,000001
10		•	•			0,364	0,556	1,45	0,259901
11		•		•		0,406	0,546	1,88	0,179984
12		•			•	**0,936**	**0,210**	**67,51**	**0,000001**
13			•	•		0,153	0,590	0,23	0,798928
14			•		•	0,878	0,286	31,88	0,000001
15				•	•	0,879	0,285	32,29	0,000001
16	•	•	•			0,953	0,186	59,15	0,000001
17	•	•		•		0,886	0,285	21,86	0,000003
18	•	•			•	0,936	0,215	42,76	0,000001
19	•		•	•		0,932	0,223	39,53	0,000001
20	•		•		•	**0,963**	**0,166**	**76,32**	**0,000001**
21	•			•	•	0,939	0,211	44,69	0,000001
22		•	•	•		0,409	0,560	1,21	0,336497
23		•	•		•	0,953	0,187	58,77	0,000001
24		•		•	•	0,959	0,174	68,56	0,000001
25			•	•	•	0,881	0,291	20,73	0,000005
26	•	•	•	•		**0,964**	**0,167**	**56,55**	**0,000001**
27	•	•	•		•	0,963	0,170	54,53	0,000001
28	•	•		•	•	0,959	0,179	48,57	0,000001
29	•		•	•	•	0,964	0,169	55,33	0,000001
30		•	•	•	•	**0,959**	**0,178**	**48,97**	**0,000001**
31	•	•	•	•	•	**0,964**	**0,172**	**42,65**	**0,000001**

Tabela 6. Matriz de correlação, em termos de R^2, das variáveis independentes da Tabela 4. As únicas variáveis que apresentam alto grau de correlação são σ^+ e σ_m.

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

Validação estatística das variáveis selecionadas

É fundamental que testes de avaliação do modelo e das variáveis selecionadas sejam executados. A avaliação mínima que se exige para um modelo de regressão linear envolve os seguintes testes. (a) Verificação do grau de ajuste do modelo, que envolve o cálculo do coeficiente de correlação (R) e do desvio-padrão (s), análise do gráfico da atividade observada em função da atividade prevista pelo modelo $\left(Y \times \hat{Y}\right)$ e do gráfico dos resíduos da regressão em função da atividade observada $\left(\left(Y - \hat{Y}\right)_X \hat{Y}\right)$; (b) verificação do grau de significância do modelo, que envolve o cálculo do teste F (95% confiança), cálculo do p-valor relativo ao resultado do teste F e cálculo do intervalo de confiança dos coeficientes da regressão (95% de confiança) e; (c) verificação do grau de previsibilidade do modelo, através da execução do teste de validação cruzada e o subseqüente cálculo do coeficiente de correlação (Q^2) e do desvio padrão (s_{PRESS}) das previsões.

Como exemplo de avaliação estatística de uma equação de regressão linear, pode-se avaliar o próprio modelo apresentado por Unger e Hansch[1] (eq 3) como alternativa ao modelo de Cammarata (eq 2). A eq 17 corresponde à versão recalculada da eq 3, em que podem ser notadas pequenas alterações em sua forma original. A avaliação da eq 17 é apresentada a seguir.

$$\log 1/C = 1{,}14\ (\pm 0{,}21)\ \pi - 1{,}25\ (\pm 0{,}40)\ \sigma^+ + 7{,}80\ (\pm 0{,}21) \quad (17)$$
$$(n = 22;\ R = 0{,}932;\ s = 0{,}217;\ F = 62{,}54;\ p < 0{,}000001;\ Q^2 = 0{,}808;\ s_{PRESS} = 0{,}262)$$

Análise do grau de ajuste

O modelo de regressão representado pela eq 17 é capaz de explicar cerca de 87% da variabilidade dos valores observados da atividade ($R^2 \times 100$), o que é um excelente nível de ajuste. A excelência do ajuste é confirmada pelo baixo valor do desvio-padrão (s = 0,217). Esses valores podem ser objetivamente analisados em termos de dois gráficos: log $1/C_{Obs}$ em função de log $1/C_{Prev}$ e resíduos da regressão em função de log $1/C_{Obs}$ (Figura 4). No gráfico da atividade observada em função da atividade prevista (Figura 4a) é importante observar o alinhamento dos pontos em relação à reta ajustada, bem como a distribuição dos pontos ao longo do intervalo de valores de atividades estudado. Caso haja agrupamento de pontos em certas regiões do gráfico e/ou pontos isolados, principalmente nos extremos do gráfico, deve-se estudar com cuidado o impacto que a presença desses pontos tem sobre a estrutura da equação de regressão. A reconstrução do modelo na ausência desses pontos deverá fornecer dados importantes sobre isso. Na Figura 4a, observa-se que, dos 22 compostos estudados, 10 estão fora da área delimitada pelas linhas tracejadas, que corresponde à região do gráfico onde existe 95% de probabilidade de passar a verdadeira reta do gráfico log $1/C_{Obs}$ em função de log $1/C_{Prev}$. Entretanto, oito dos compostos fora da região tracejada encontram-se bem próximos a ela. Apenas dois compostos apresentam desvios apreciáveis, sendo eles os compostos **5** (4-I) e **6** (4-Me), cujos resíduos são, respectivamente, 0,34 e 0,45. Compostos que apresentam grandes resíduos num modelo de regressão são denominados *outliers*. Na maioria dos casos observados na literatura, a detecção da presença de *outliers* é sucedida pela exclusão dos compostos correspondentes e pelo recálculo da equação. Este costuma ser o destino dos *outliers*, pois sua exclusão fatalmente melhora o grau de ajuste da equação. Pode ser importante analisar o motivo do não ajuste de um *outlier* a dado modelo, pois acredita-se que, assim fazendo, informações importantes sobre o mecanismo de ação dos compostos da série podem ser obtidas[7,14].

Na Figura 4b, deve-se observar a distribuição dos resíduos em torno de zero, que corresponde à linha horizontal central. Espera-se que um modelo adequado aos dados observados tenha seus resíduos aleatoriamente dispersos em torno de zero. E é exatamente isso o que se observa na Figura 4b.

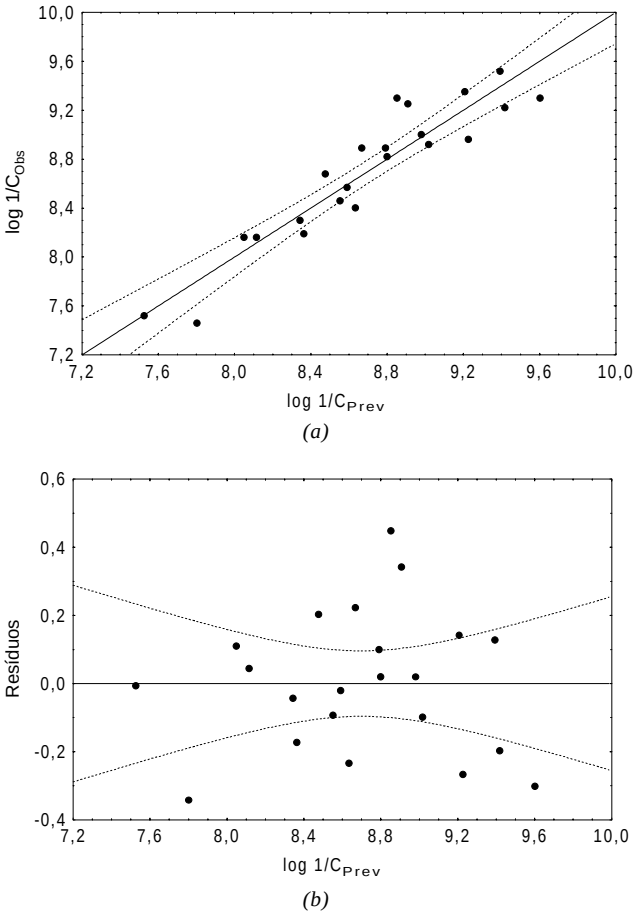


Figura 4. *(a) Atividade biológica observada em função da atividade prevista; (b) resíduos da regressão em função da atividade prevista. As linhas tracejadas delimitam a região dentro da qual o valor previsto da atividade difere, no máximo, 5% em relação ao valor observado.*

Grau de significância do modelo

O valor de referência do teste F para um nível de confiança de 95% (p = 0,05) é $F_{(k,\ n\text{-}k\text{-}1)} = F_{(2,\ 19)}$ = 3,52. Como o teste F da eq 17 (F = 62,54) é bem maior do que o correspondente valor de referência ($F_{(2,\ 19)}$ = 3,52), o nível de significância do modelo também é bem maior do que 95%. Na verdade, como p < 0,000001, o nível de significância do modelo é maior do que 99,9999%. Analisando-se o intervalo de confiabilidade dos coeficientes, percebe-se que todos os coeficientes da regressão são significativos, no nível de confiança equivalente a 95%. Essa constatação decorre do fato de que os intervalos de confiança, mostrados entre parênteses juntos aos respectivos coe-

ficientes, apresentam valores inferiores aos dos próprios coeficientes. Por exemplo, o coeficiente de π é 1,14 e seu intervalo de confiança é ±0,21. Portanto o valor verdadeiro desse coeficiente, que se está tentando descobrir ao construir a equação de regressão, é algum valor entre 1,14 - 0,21 = 0,93 e 1,14 + + 0,21 = 1,35, com 95% de probabilidade. Caso o intervalo de confiança fosse maior do que o próprio valor do coeficiente, o intervalo incluiria o valor zero para o coeficiente.

Grau de previsibilidade do modelo

Valores de Q^2 próximos à unidade e de s_{PRESS} próximos a zero revelam alto grau de previsibilidade do modelo. Infelizmente não existem regras que estabeleçam, em termos absolutos, se o grau de previsibilidade é bom ou ruim a partir do valor de Q^2 e s_{PRESS}. Estes valores têm muito mais utilidade quando se deseja comparar a capacidade de previsão de dois modelos: o que possui maior Q^2 e menor s_{PRESS} possui maior grau de previsibilidade. Na eq 17, o valor de Q^2 (0,808) é muito mais próximo da unidade do que de zero, indicando bom poder de previsão.

Princípio da parcimônia

Trata-se de um princípio fundamental que pode ser utilizado em todas as áreas da ciência. Em QSAR, é comum dispor-se de mais de uma possibilidade em termos de modelos para a escolha daquele que será considerado o melhor modelo de relação estrutura-atividade. A necessidade da escolha de uma entre várias opções de modelos, aproximadamente equivalentes, cria dúvidas em relação a qual deve ser considerado o melhor. Naturalmente que, em se tratando de modelos com o mesmo número de variáveis explicativas, deve-se dar preferência para a equação que apresentar a melhor avaliação (maior *R*, menor *s*, maior *F*, etc.). Nos casos em que os modelos possuírem diferentes números de variáveis explicativas, o princípio da parcimônia aconselha a escolha do modelo com menor número de variáveis. Mas é preciso lembrar que esse princípio deve ser aplicado somente quando se comparam modelos aproximadamente equivalentes. Como saber se dois modelos com número de variáveis diferentes são equivalentes se, nesses casos, os valores de *R*, *s* e *F* não podem ser diretamente comparados? Em situações como essas, o coeficiente de correlação ajustado (R_{Ajust}, eq 10) e o p-valor são mais adequados à comparação, pois seus valores consideram correções para o número de variáveis e para o número de compostos utilizados. Assim, a equação preferida seria aquela com o maior R_{Ajust} e o menor p-valor.

Como exemplo da aplicação do princípio da parcimônia, podem-se comparar as eqs 2 (Cammarata) e 3 (Unger e Hansch), recalculadas, supondo-se que ambas possam explicar adequadamente a atividade dos compostos da série (como foi dito na Introdução deste trabalho, isso não é verdade). Para isso, é necessário recalcular a eq 2, da mesma forma como foi feito para a eq 3. Utilizando-se os dados da Tabela 4, o modelo de Cammarata passa a ser representado pela eq 18.

$$\log 1/C = 0{,}74\ (\pm 0{,}27)\ \pi_m - 0{,}75\ (\pm 0{,}50)\ \sigma_m + 1{,}67\ (\pm 0{,}27)\ r_v^p + 5{,}75\ (\pm 0{,}47) \quad (18)$$

(n = 22; R = 0,959; s = 0,174; F = 68,55; p < 0,000001; Q^2 = 0,874; s_{PRESS} = 0,218)

A pergunta é: qual, dentre as eqs 17 e 18, deve ser considerada como sendo a melhor? O princípio da parcimônia diz que entre dois modelos (aproximadamente) equivalentes deve-se optar pelo mais simples. Seguindo esse princípio, a eq 17 deve ser escolhida como sendo a melhor, pois contém uma variável a menos. No entanto, embora o julgamento do que seja mais simples seja relativamente fácil, o julgamento do que seja (aproximadamente) equivalente pode não ser. Para decidir sobre a equivalência da qualidade estatística das eqs 17 e 18, deve-se proceder à avaliação das mesmas. A seguir é mostrado o resultado dessa avaliação.

Análise do grau de ajuste

Não é possível comparar os coeficientes de correlação das eq 17 (R = 0,932) e 18 (R = 0,959), pois essas equações possuem diferentes números de variáveis. Neste caso, é preciso comparar os coeficientes de correlação ajustados das duas equações. A eq 17 possui R_{Ajust} = 0,924 enquanto que a eq 18 possui R_{Ajust} = 0,952. Isso significa que a eq 18 (Cammarata) possui maior grau de ajuste.

Análise do grau de significância

Também não é possível comparar os valores dos teste F das eq 17 (F = 62,54) e 18 (F = 68,55), devido ao número diferente de variáveis envolvidas. É preciso comparar o p-valor das duas equações. A eq 17 possui p < 0,000001 e a eq 18 também possui p < 0,000001. Assim, ambas as equações possuem aproximadamente o mesmo grau de significância.

Análise do grau de previsibilidade

A eq 17 possui Q^2 = 0,808 e s_{PRESS} = 0,262, enquanto que a eq 18 possui Q^2 = 0,874 e s_{PRESS} = 0,218. Portanto, a eq 18 (Cammarata) possui maior capacidade de fazer previsões acerca da atividade biológica de compostos não incluídos no conjunto de dados do que a eq 17 (Unger e Hansch).

Portanto, conclui-se que a eq 18 é superior à eq 17 em termos de ajuste e previsibilidade e é equivalente à eq 17 em termos de significância. Em termos estatísticos, a conclusão óbvia é que a eq 18 (Cammarata) é melhor do que a eq 17 (Unger e Hansch), mesmo tendo aquela uma variável a mais do que esta. Mas cabe outra pergunta: será que a superioridade estatística da eq 18 é tal que a torna não-equivalente à eq 17? Esta pergunta, como muitas outras semelhantes, não tem resposta exata pois não se dispõe de uma tabela contendo valores de referência para auxiliar a tomada de decisão. Porém, deve-se notar que, ao menos intuitivamente, os valores de R_{Ajust} e Q^2 das duas equações são próximos (os desvios de R_{Ajust} e Q^2, relativos aos seus maiores valores nas eqs 17 e 18, são 2% e 8%, respectivamente) e, não sendo muito exigente, podem-se considerá-las estatisticamente equivalentes. Uma vez consideradas equivalentes, aplica-se o princípio da parcimônia. Neste caso, a melhor equação é a eq 17, pois é mais simples. Mas cabe uma observação final. O fato da eq 18 não ser consistente com o mecanismo de ação dos compostos envolvidos[1], torna a mesma não equivalente à eq 17, sendo, portanto, desnecessária a aplicação do princípio da parcimônia. Pela inconsistência com o mecanismo de ação, a eq 18 não poderia se tornar um modelo de QSAR. Como foi dito anteriormente, a análise comparativa das eqs 17 e 18 foi feita desprezando-se essa observação. Outra análise deste mesmo caso foi feita por Kubinyi[7].

Número de variáveis em cada modelo

Baseado no trabalho de Topliss e Costello[71], Unger e Hansch[1] sugerem que, para cada variável explicativa incluída em modelos de QSAR, devem haver, no mínimo, cerca de cinco ou seis compostos. Essa regra tenta evitar a ocorrência de correlação por coincidência. No entanto, Kubinyi[7] ressalta que, para conjuntos de dados com poucos compostos, pode-se violar essa regra de forma controlada (por exemplo incluindo-se duas variáveis quando se dispõe de apenas oito compostos), desde que a avaliação do modelo justifique a presença dessas

variáveis. Além disso, Kubinyi acrescenta que a disponibilidade de muitos compostos pode gerar modelos com grande número de variáveis (um conjunto de dados com 36 compostos permitiria a construção de modelos com sete variáveis, o que pode ser um exagero), aumentando a probabilidade de ocorrência de correlação por coincidência. Enfim, vale dizer que esta é apenas uma regra geral. Serve apenas para guiar os autores com pouca experiência em estatística na elaboração de modelos de QSAR baseados em regressão linear múltipla.

Exemplo extremo de utilização dessa regra é fornecido por Kim e colaboradores[72], que analisaram a atividade antimalarial em ratos de nada menos do que 646 compostos derivados de fenantrenos, quinolinas e piridinas (eq 19, em que não será mencionado o significado das variáveis citadas).

$$\log 1/C = 0{,}576\ (+0{,}09)\ \Sigma\sigma + 0{,}168\ (+0{,}05)\ \Sigma\pi + 0{,}105\ (+0{,}05)\ \log P - 0{,}167\ (+0{,}07)\ \log(\beta P + 1) - 0{,}169\ (+0{,}10)\ \textit{c-side} + 0{,}319\ (+0{,}136)\ CNR_2 - 0{,}139\ (+0{,}06)\ \textit{AB} - 0{,}795\ (+0{,}06)\ \textit{<3-cures} + 0{,}278\ (+0{,}11)\ \textit{MR-4´-Q} + 0{,}252\ (+0{,}18)\ \textit{Me-6,8-Q} + 0{,}084\ (+0{,}10)\ \textit{2-Pip} + 0{,}151\ (+0{,}19)\ \textit{NBrPy} - 0{,}683\ (+0{,}22)\ \textit{Q2P378} + 0{,}267\ (+0{,}11)\ \textit{Py} + 2{,}726\ (+0{,}15) \quad (19)$$
$(n = 646;\ R = 0{,}898;\ s = 0{,}309;\ \log b = -3{,}959;\ \log P_{Ótimo} = 4{,}19)$

Neste modelo foram incluídas 14 variáveis, número que certamente é justificado pela imensa quantidade de compostos analisados. Se a eq 19 for consistente com algum mecanismo de ação, fato que os autores não analisaram, esse mecanismo deverá ser extremamente complexo. Pode-se afirmar com alguma segurança que a eq 19 é o mais complexo modelo de QSAR já apresentado na literatura.

A literatura também apresenta algumas equações que devem servir de exemplo negativo quanto ao número de variáveis explicativas em relação ao número de compostos analisados. É o caso da eq 20, construída por Jha e colaboradores[73] para explicar a atividade antineoplásica de derivados do ácido glutâmico.

$$\log \%\ ITW = 1{,}4111\ (\log P)^2 - 0{,}5971\ \log P - 0{,}1714\ \pi^2_{Ali} - 3{,}2293\ \sigma_{Ali} + 0{,}9595\ E_{s\ Ali} - 6{,}6199\ \sigma_I + 1{,}3249 \quad (20)$$
$(n = 8;\ R = 0{,}9912;\ s = 0{,}0310)$

Na eq 20, nota-se a presença de seis variáveis explicativas numa equação envolvendo apenas oito compostos. Pode-se observar que os autores avaliaram apenas o grau de ajuste da equação (cálculo de R e s). A avaliação do grau de significância da equação certamente revelaria que a mesma não possui qualidade estatística mínima e, portanto, não pode representar um modelo de QSAR. Kubinyi[7] apontou outros erros nessa equação: (a) atividade biológica em escala imprópria (escalas aceitas são log 1/C, C = IC50, ED50, LD100, etc.; log 1/K_i, K_i = constante de inibição enzimática; log k, k = constante cinética ou de equilíbrio de uma reação; etc.) ; (b) pequena variabilidade dos valores da atividade (a diferença entre o composto mais ativo e o menos ativo é de apenas 0,24 unidade logarítmica; o mínimo aconselhável é cerca de duas unidades logarítmicas); (c) sinal dos coeficiente de $(\log P)^2$ e log P incorretos (parábola invertida); (d) desvio-padrão inconsistente com o tipo de atividade (modelos que descrevem a atividade antineoplásica costumam gerar valores de $s >> 0{,}3$); (e) não há intervalo de confiança para os coeficientes; (f) casas decimais em excesso.

Outro exemplo de aberração foi publicado recentemente por Kong e colaboradores[74] (eq 21).

$$\log EC_{50} = 0{,}116\ \log K_{ow} - 2{,}502\ E_{homo} - 6{,}269\ E_S + 3{,}121\ \pi_X - 0{,}3898\ {}^1\chi^v \quad (21)$$
$(n = 4;\ R = 0{,}99;\ s = 0{,}27)$

Nessa equação percebe-se que o número de variáveis explicativas excede o número de compostos (!). Neste caso, não é possível sequer fazer testes de avaliação do grau de significância e de previsibilidade da equação, pois não há graus de liberdade disponíveis. Aliás, embora o cálculo dos coeficientes da equação e do coeficiente de correlação ainda sejam possíveis, o cálculo do desvio-padrão não o é. De acordo com o Quadro 2, $s^2 = SS_{Res}/(n\text{-}k\text{-}1)$ e, nesse caso, $n\text{-}k\text{-}1 = -2$, implicando em $s^2 < 0$ (!). O número de graus de liberdade para o cálculo do desvio-padrão (n-k-1) pressupõe a existência de um termo constante na equação de regressão (b_0), que não foi incluído na eq 21.

Modelo qualitativo para o mecanismo de ação dos compostos

Além da validação estatística, é fundamental que uma equação de regressão que pretende ser promovida a modelo de QSAR deve ser validada em termos de sua capacidade de explicação do mecanismo de ação dos compostos da série analisada. É o caso da eq 3, que é consistente com o mecanismo de

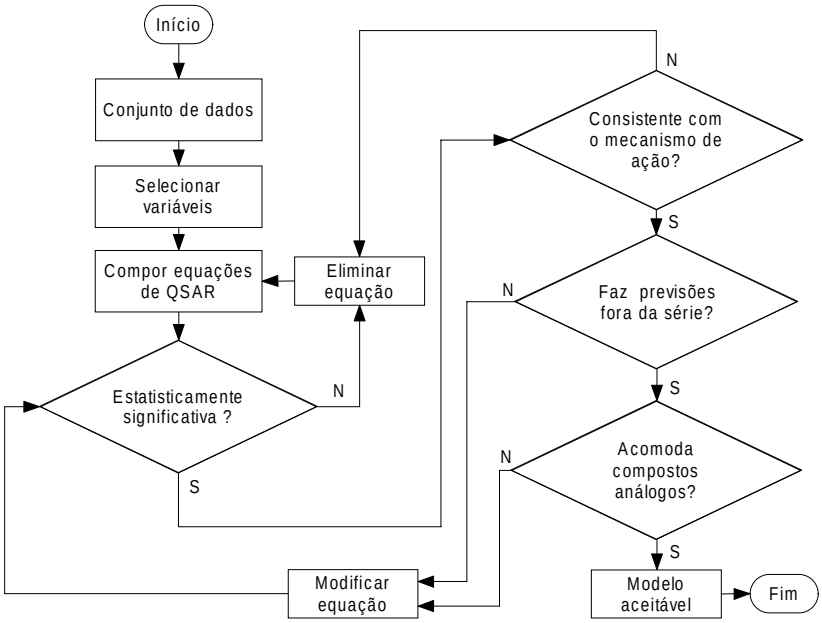


Figura 5, *Esquema geral de proposição de modelos de QSAR.*

bloqueio da atividade adrenérgica exercido pelas β-halo-β-arilalquilaminas[1], mostrado na Introdução deste trabalho.

Pode-se resumir o conteúdo desta seção através de um diagrama de blocos mostrando as principais etapas da elaboração de modelos de QSAR (Figura 5). Nesse esquema percebe-se o caminho relativamente longo entre a construção da equação de regressão e o modelo propriamente dito. Para ser considerada como um modelo de QSAR, além de ser estatisticamente aceitável, a equação precisa ser consistente com algum mecanismo de ação aceitável para os compostos da série, caso contrário deverá ser descartada. Também a equação deverá ser capaz de fazer previsões fora da série de compostos testada. Esse aspecto nem sempre é fielmente verificado pois implica na síntese de compostos adicionais. O que se costuma fazer é utilizar o resultado da validação cruzada como verificação da capacidade de previsão da equação. Finalmente, a equação deverá ser capaz de acomodar compostos com estrutura semelhante aos já incluídos na série sem que a equação seja apreciavelmente modificada.

CONCLUSÕES

A proposição de modelos quantitativos de relações entre estrutura química e atividade biológica é baseada nas cinco regras gerais de proposição de modelos de Unger e Hansch[1]: (a) selecionar as variáveis independentes do modelo dentre grande número de variáveis; (b) validar estatisticamente as variáveis selecionadas; (c) aplicar o princípio da parcimônia; (d) cada modelo deve apresentar cerca de cinco compostos para cada variável independente e; (e) o modelo deve ser consistente com o mecanismo de ação dos compostos. A validação estatística das equações de regressão linear é feita através da avaliação do modelo, dividida em três níveis: (a) avaliação do grau de ajuste; (b) avaliação do grau de significância e; (c) avaliação do grau de previsibilidade da equação. Em cada uma das etapas da avaliação, testes estatísticos específicos são executados e seus resultados avaliados. Uma equação de regressão que sobrevive às regras de proposição de modelos e à avaliação completa pode ser promovida a modelo quantitativo de relação estrutura-atividade.

AGRADECIMENTOS

Os autores agradecem à Pró-Reitoria de Pesquisa e Pós-Graduação da Universidade Federal do Espírito Santo, PRPPG-UFES, e ao Conselho Nacional de Desenvolvimento Científico e Tecnológico, CNPq, pelo auxílio financeiro.

REFERÊNCIAS

1. Unger, S. H.; Hansch, C.; *J. Med. Chem.* **1973**, *16*, 745.
2. Hansch, C.; Lien, E. J.; *Biochem. Pharmacol.* **1968**, *17*, 709.
3. Graham, J. D. P.; Karrar, M. A.; *J. Med. Chem.* **1963**, *6*, 103.
4. Fujita, T.; Hansch, C.; Iwasa, J.; *J. Am. Chem. Soc.* **1964**, *86*, 5175.
5. Hammett, L. P.; *J. Am. Chem. Soc.* **1937**, *59*, 96.
6. Cammarata, A.; *J. Med. Chem.* **1972**, *15*, 573.
7. Kubinyi, H.; *QSAR: Hansch Analysis and Related Approaches*. In: *Methods and Principles in Medicinal Chemistry*; R. Mannhold, P. Krogsgaard-Larsen e H. Timmerman Eds.; Vol. 1; VCH; Weinheim, 1993.
8. Gaudio, A. C.; *Quim. Nova* **1996**, *19*, 278.
9. Karelson, M.; Lobanov, V. S.; Katritzky, A. R.; *Chem. Rev.* **1996**, *96*, 1027.
10. Kubinyi, H.; *J. Med. Chem.* **1976**, *19*, 587.
11. De Oliveira, D. B.; Gaudio, A. C.; *Quant. Struct. - Act. Relat* **2000**, *19*, 599.
12. Hansch, C.; Fujita, T.; *J. Am. Chem. Soc.* **1964**, *86*, 1616.
13. Tute, M. S.; *Adv. Drug. Res.* **1971**, *6*, 1.
14. Martin, Y. C.; *Quantitative Drug Design: A Critical Introduction*. Marcel Dekker; New York, 1978.
15. Draper, N. R.; Smith, H.; *Applied Regression Analysis*. John Wiley & Sons; New York, 1981.
16. Kirsten, J. T.; *Teoria dos Modelos*. Universidade de São Paulo; São Paulo, 1983.
17. Myers, R. H.; *Classical and Modern Regression with Applications*. Duxbury Press; Boston, 1986.
18. Kmenta, J.; *Elemento de Econometria*. Vol. 2; Atlas; São Paulo, 1988.
19. Daniel, W. W.; *Biostatistics: A Foundation for Analysis in the Health Sciences*. John Wiley & Sons; New York, 1995.
20. Hoffman, R.; Vieira, S.; *Análise de Regressão: Uma Introdução à Econometria*. Hucitec; São Paulo, 1998.
21. Bussab, W. O.; *Análise de Variância e Regressão: Métodos Quantitativos*. Atual; São Paulo, 1999.
22. Hansch, C.; Dunn, W. J., III; *J. Pharm. Sci.* **1972**, *61*, 1.
23. Hansch, C.; Clayton, J. M.; *J. Pharm. Sci.* **1973**, *62*, 1.
24. Kubinyi; *J. Med. Chem.* **1977**, *20*, 625.
25. Baroni, M.; Clementi, S.; Cruciani, G.; Costantino, G.; Riganelli, D.; Oberrauch, E.; *J. Chemom.* **1992**, *6*, 347.
26. Cruciani, G.; Baroni, M.; Bonelli, D.; Clementi, S.; Ebert, C.; Skagerberg, B.; *Quant. Struct.-Act. Relat.* **1990**, *9*, 101.
27. Cruciani, G.; Baroni, M.; Clementi, S.; Constantino, G.; Riganelli, D.; Skagerberg, B.; *J. Chemom.* **1992**, *6*, 335.
28. Cramer, R. D.; Bunce, J. D.; Patterson, D. E.; Frank, I. E.; *Quant. Struct.-Act. Relat.* **1988**, *7*, 18.
29. Hansch, C.; Leo, A.; *Exploring QSAR: Fundamentals and Applications in Chemistry and Biology*. American Chemical Society; Washington D. C., 1995.
30. Hansch, C.; Leo, A.; Hoekman, D.; *Exploring QSAR: Hydrophobic, Electronic, and Steric Constants*. American Chemical Society; Washington D.C., 1995.
31. Hansch, C.; *Annu. Rep. Med. Chem.* **1966**, *34*, 347.
32. Hansch, C.; Rockwell, S. D.; Jow, P. Y. C.; Leo, A.; Steller, E. E.; *J. Med. Chem.* **1977**, *20*, 304.
33. Hansch, C.; *Annu. Rep. Med. Chem.* **1967**, *35*, 348.
34. Katritzky, A. R.; Gordeeva, E. V.; *J. Chem. Inf. Comp. Sci.* **1993**, *33*, 835.
35. Ertl, P.; *Quant. Struct.-Act. Relat.* **1997**, *16*, 377.
36. Vaz, R. J.; *Quant. Struct.-Act. Relat.* **1997**, *16*, 303.
37. Clare, B. W.; *Theor. Chim. Acta* **1994**, *87*, 415.
38. Dearden, J. C.; Ghafourian, T.; *J. Chem. Inf. Comp. Sci.* **1999**, *39*, 231.
39. Benigni, R.; Cotta-Ramusino, M.; Giorgi, F.; Gallo, G.; *J. Med. Chem.* **1995**, *38*, 629.
40. Good, A. C.; So, S.-S.; Richards, W. G.; *J. Med. Chem.* **1993**, *36*, 433.
41. Fradera, X.; Amat, L.; Besalú, E.; Carbó-Dorca, R.; *Quant. Struct.-Act. Relat.* **1997**, *16*, 25.
42. Lobato, M.; Amat, L.; Besalú, E.; Carbó-Dorca, R.; *Quant. Struct.-Act. Relat.* **1997**, *16*, 465.
43. Kubinyi, H.; *Quant. Struct.-Act. Relat.* **1994**, *13*, 393.
44. Kovalishyn, V. V.; Tetko, I. V.; Luik, A. I.; *J. Chem. Inf. Comp. Sci.* **1998**, *38*, 651.
45. Maddalena, D. J.; *Exp. Opin. Ther. Patents* **1998**, *8*, 249.
46. Waller, C. L.; Bradley, M. P.; *J. Chem. Inf. Comp. Sci.* **1999**, *39*, 345.
47. Lucic, B.; Trinajstic, N.; *J. Chem. Inf. Comp. Sci.* **1999**, *39*, 121.
48. So, S. S.; Karplus, M.; *J. Med. Chem.* **1997**, *40*, 4360.
49. Tetko, I. V.; Villa, A. E. P.; Livingstone, D. J.; *J. Chem. Inf. Comp. Sci.* **1996**, *36*, 794.
50. So, S. -S.; Karplus, M.; *J. Med. Chem.* **1997**, *40*, 4347.
51. Kubinyi, H.; *Quant. Struct.-Act. Relat.* **1994**, *13*, 285.
52. Murtaugh, P. A.; *Commun. Stat.-Simul.* **1998**, *27*, 711.
53. Maddalena, D. J.; Snowdon, G. M.; *Exp. Opin. Ther. Patents* **1997**, *7*, 247.

54. Hasegawa, K.; Kimura, T.; Funatsu, K.; *J. Chem. Inf. Comp. Sci.* **1999**, *39*, 112.
55. Hasegawa, K.; Funatsu, K.; *J. Mol. Struct. (Theochem)* **1998**, *425*, 255.
56. Kimura, T.; Hasegawa, K.; Funatsu, K.; *J. Chem. Inf. Comp. Sci.* **1998**, *38*, 276.
57. Tominaga, Y.; Fujiwara, I.; *J. Chem. Inf. Comp. Sci.* **1997**, *37*, 1152.
58. Norinder, U.; Rivera, C.; Undén, A.; *J. Pept. Res.* **1997**, *49*, 155.
59. Kubinyi, H.; *J. Chemom.* **1996**, *10*, 119.
60. Schmidli, H.; *Chemom. Intell. Lab. Syst.* **1997**, *37*, 125.
61. Supuran, C. T.; Clare, B. W.; *Eur. J. Med. Chem.* **1995**, *30*, 687.
62. Mracec, M.; Muresan, S.; Mracec, M.; Simon, Z.; Náray-Szabó, G.; *Quant. Struct.-Act. Relat.* **1997**, *16*, 459.
63. Kelder, J.; Greven, H. M.; *Rec. Trav. Chim. Pays-Bas - J. Royal Netherl. Chem. Soc.* **1979**, *98*, 168.
64. Menziani, M. C.; De Benedetti, P. G.; Karelson, M.; *Bioorg. Med. Chem.* **1998**, *6*, 535.
65. Gaudio, A. C.; *Dissertação de Mestrado*; Unicamp, Campinas, SP, 1992.
66. Selwood, D. L.; Livingstone, D. J.; Comley, J. C. W.; O'Dowd, A. B.; Hudson, A. T.; Jackson, P.; Jandu, K. S.; Rose, V. S.; Stables, J. N.; *J. Med. Chem.* **1990**, *33*, 136.
67. Cocchi, M.; Menziani, M. C.; Fanelli, F.; De Benedetti, P. G.; *J. Mol. Struct. (Theochem)* **1995**, *331*, 79.
68. Gaudio, A. C.; *Tese de Doutorado*; Unicamp, Campinas, SP, 1998.
69. Hansch, C.; Leo, A.; Taft, R. W.; *Chem. Rev.* **1991**, *91*, 165.
70. Ferreira, M. M. C.; Antunes, A. M.; Melo, M. S.; Volpe, P. L. O.; *Quim. Nova* **1999**, *22*, 724.
71. Topliss, J. G.; Costello, R. J.; *J. Med. Chem.* **1972**, *15*, 1066.
72. Kim, K. H.; Hansch, C.; Fukunaga, J. Y.; Steller, E. E.; Jow, P. Y. C.; Craig, P. N.; Page, J.; *J. Med. Chem.* **1979**, *22*, 366.
73. Jha, T.; Debnath, A. K.; Mazumdar, A.; Sengupta, C.; De, A. U.; *Indian J. Chem.* **1986**, *25*, 169.
74. Kong, F. X.; Hu, W.; Liu, Y.; *Environ. Exp. Bot.* **1998**, *40*, 105.