Journal of Medicinal Chemistry

Volume 20, Number 5 May 1977

Theory of Linear Equations as Applied to Quantitative Structure–Activity Correlations

L. J. Schaad* and B. A. Hess, Jr.*

Department of Chemistry, Vanderbilt University, Nashville, Tennessee 37235. *Received June 7, 1976*

The theory of linear equations is reviewed and the results used to examine the types of linear dependence difficulties that can occur in quantitative structure–activity relationships.

There have been a number of errors[1] in the literature on quantitative structure–activity relationships (QSAR) caused by a misunderstanding of points in the theory of linear equations. Because of the large and increasing practical importance of QSAR in drug design, a simple exposition of this theory, starting at an elementary level, is presented here.

Determinants and matrices are mathematical constructs which allow a compact notation in treating linear equations. Both require building up a certain background of definitions and theorems. The treatment here will be based primarily on Cramer's rule and will make use of the five properties of determinants in the appendix of ref 2; but, at the cost of a slightly inelegant notation, matrices will be avoided.

Cramer's Rule. Consider a set of n linear equations in the n unknowns $x_1, x_2, \ldots x_n$.

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= c_1 \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n &= c_2 \\ \ldots \qquad\qquad \ldots & \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n &= c_n \end{aligned} \tag{1}$$

According to Cramer's rule, which is easily proved,[3,4] the unknown x's may be written as the ratio of two determinants

$$x_1 = \frac{\begin{vmatrix} c_1 & a_{12} & \ldots & a_{1n} \\ c_2 & a_{22} & \ldots & a_{2n} \\ & \ldots & & \\ c_n & a_{n2} & \ldots & a_{nn} \end{vmatrix}}{\det|a_{ij}|} \quad \ldots \quad x_n = \frac{\begin{vmatrix} a_{11} & a_{12} & \ldots & c_1 \\ a_{21} & a_{22} & \ldots & c_2 \\ & \ldots & & \\ a_{n1} & a_{n2} & \ldots & c_n \end{vmatrix}}{\det|a_{ij}|} \tag{2}$$

where $\det|a_{ij}|$ is a determinant formed from the coefficients a_{ij} in eq 1, and the determinant in the numerator of x_i is obtained by replacing the ith column of $\det|a_{ij}|$ by the column of c_i's from eq 1.

Thus the solution of eq 1 is given explicitly and is unique. However, it is clear that there will be trouble if the denominators vanish, that is, if

$$\det|a_{ij}| = 0 \tag{3}$$

and this case must be examined more closely. If any row of $\det|a_{ij}|$ is a linear combination of the others, then by property (e) of ref 2 $\det|a_{ij}| = 0$. The converse is also true, so that eq 3 implies that some row (suppose it is the last) is a linear combination of the others.

$$(\text{row } n) = \sum_{\alpha=1}^{n-1} \lambda_\alpha \times (\text{row } \alpha) \tag{4}$$

It follows at once that the left-hand side of the last line in eq 1 is the same linear combination of the other left-hand sides.

$$(a_{n1}x_1 + \ldots + a_{nn}x_n) = \sum_{\alpha=1}^{n-1} \lambda_\alpha (a_{\alpha 1}x_1 + \ldots + a_{\alpha n}x_n) \tag{5}$$

There are then two cases; either eq 6

$$c_n = \sum_{\alpha=1}^{n-1} \lambda_\alpha c_\alpha \tag{6}$$

holds, or it does not. If it does, then the last equation in eq 1 is a linear combination of the first $(n-1)$ equations. It imposes no further restriction on the n variables so that one has in effect $(n-1)$ equations in n unknowns which cannot be solved uniquely for all n variables. All one can do is discard the superfluous line in eq 1 and solve the remaining $(n-1)$ equations for $(n-1)$ of the unknowns in terms of the other.

$$\begin{aligned} a_{11}x_1 + \ldots + a_{1,n-1}x_{n-1} &= c_1 - a_{1n}x_n \\ a_{21}x_1 + \ldots + a_{2,n-1}x_{n-1} &= c_2 - a_{2n}x_n \\ \ldots \qquad\qquad \ldots & \\ a_{n-1,1}x_1 + \ldots + a_{n-1,n-1}x_{n-1} &= c_{n-1} - a_{n-1,n}x_n \end{aligned} \tag{7}$$

The solution thus determined is unique only if there is some extraneous way of finding the last variable. Equation 7 was written with x_n taken to the right, but any of the other variables might as well have been used.

It was assumed above that the last line in eq 1 was a combination of the others. Condition 3, of course, only implies that some line is such a combination; and one must be careful to discard such a line, not one that is independent. For example, if the second equation equals the first minus twice the third, any of these three may be discarded; the fourth must not be.

In the second case, that is, when eq 6 does not hold, then the equations are inconsistent and cannot be solved. A simple example is

$$x_1 + x_2 = 2$$
$$2x_1 + 2x_2 = 6 \quad (8)$$

The left side of the second equation in eq 8 is double the first, but the right sides are not related in the same way. One arrives at the obvious inconsistency $(x_1 + x_2) = 2 = 3$. Inconsistent sets of equations will not be encountered in the QSAR applications below except through error.

The difficulties above are compounded, but with no new principles involved, if there is more than one linear dependence connecting the lines in eq 1. For example, the second equation might equal the first minus twice the third and the seventh equal the sum of the eighth, ninth, and tenth. Discarding one equation as in eq 7 will give a set of $(n - 1)$ equations whose $(n - 1) \times (n - 1)$ determinant of coefficients still vanishes. Two (or more) equations must be discarded and $n - 2$ (or fewer) unknowns be determined in terms of two (or more) arbitrary x's.

If, in eq 1

$$c_1 = c_2 = \ldots = c_n = 0 \quad (9)$$

the equations are said to be homogeneous. Applying Cramer's rule gives an entire column of zeros in the determinant in the numerator of each x. Using property (a) of ref 2 shows

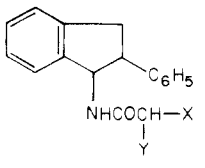
$$x_1 = x_2 = \ldots = x_n = 0 \quad (10)$$

This trivial solution, as it is called, is the only one unless $\det|a_{ij}| = 0$. Homogeneous linear equations are important in many applications. There the trivial solution is usually of no interest, and one looks for conditions on the a_{ij} to make $\det|a_{ij}|$ vanish. However, the homogeneous case will not appear in the QSAR applications below.

Application to QSAR Equations. QSAR are of two main kinds. In the Hansch approach[5,6] the biological activity of a set of molecules is correlated, usually by the least-squares method, with some other physical property or properties. The partition ratio between water and octanol is most often used; but Hammett's σ constants, steric constants, calculated electron density, and pK_a are among the other physical parameters that have also been employed. The second technique, by Free and Wilson,[7,8] makes no connection between biological activity and other molecular properties but assumes that the biological activity equals a sum of additive contributions from the various parts of the active molecule. Best values of the fragment contributions are determined by least squares. Which physical parameters and what function of them should be used, the best form of the biological activity to be fit, and the validity of the assumption of additivity are all currently under active discussion. The reader is referred to important papers by Craig[9] and Kubinyi and Kehrhahn[10,11] for these points which will not be considered here.

The linear dependence problems encountered in the Free–Wilson method include all those in the Hansch method plus an additional type built into the model. It will therefore be sufficient to consider only the Free–Wilson method in detail.

To be specific, consider Free and Wilson's[7] first example.



I

Here, although the two substituent sites are identical, for the moment treat them as distinguishable. We shall return to this point in the last section below. The substituent X may be any of the set $(X_1, X_2, \ldots, X_N)$ and Y any of the set $(Y_1, Y_2, \ldots, Y_M)$. The following discussion will generalize easily to any number of substituent sites. Suppose some biological activity, f, has been measured for K of these molecules. Then by assumption one may write

$$f_k(\text{obsd}) \approx f_k(\text{calcd}) \equiv \mu + \sum_{i=1}^{N} x_i n_{ki} + \sum_{j=1}^{M} y_j m_{kj} \quad (11)$$
$$k = 1, 2, \ldots, K$$

where f_k(obsd) is the measured activity of the kth molecule, x_i is the activity contribution of X_i, n_{ki} is the number of times (0 or 1) that X_i occurs in molecule k, and y_j and m_{kj} are defined analogously for substituent Y_j. In each molecule one and only one of the n_{ki} and one and only one of the m_{kj} will equal 1; all others equal zero. The interpretation of the parameter μ will depend upon arbitrary choices yet to be specified.

The parameters $\mu, x_1 \ldots x_N, y_1 \ldots y_M$ are to be varied to match calculated activities to be observed in the best way possible, that is to minimize the quantity

$$S = \sum_{k=1}^{K} [f_k(\text{obsd}) - f_k(\text{calcd})]^2 \quad (12)$$

Setting the derivatives of S with respect to each of these $(N + M + 1)$ parameters equal to zero gives

$$\mu K + x_1 \sum_k n_{k1} + \ldots + y_M \sum_k m_{kM} = \sum_k f_k(\text{obsd})$$
$$\mu \sum_k n_{k1} + x_1 \sum_k n_{k1}^2 + \ldots + y_M \sum_k n_{k1} m_{kM}$$
$$= \sum_k n_{k1} f_k(\text{obsd})$$
$$\ldots \qquad \ldots$$
$$\mu \sum_k m_{kM} + x_1 \sum_k m_{kM} n_{k1} + \ldots + y_M \sum_k m_{kM}^2$$
$$= \sum_k m_{kM} f_k(\text{obsd}) \quad (13)$$

to be solved for the unknowns $\mu, x_1, \ldots, y_M$ by Cramer's rule. But consider the resulting denominator

$$\begin{vmatrix} K & \sum_k n_{k1} & \ldots & \sum_k m_{kM} \\ \sum_k n_{k1} & \sum_k n_{k1}^2 & \ldots & \sum_k n_{k1} m_{kM} \\ \ldots & \ldots & \ldots & \ldots \\ \sum_k m_{kM} & \sum_k m_{kM} n_{k1} & \ldots & \sum_k m_{kM}^2 \end{vmatrix} \quad (14)$$

Adding rows 3 through $N + 1$ to row 2, which by property (d) of ref 2 does not change the value of the determinant, and using the fact that each molecule contains one and only one substituent in the X position and therefore

$$n_{k1} + n_{k2} + \ldots + n_{kN} = 1 \quad (15)$$
$$k = 1, 2, \ldots, K$$

gives

$$\sum_k n_{k1} + \sum_k n_{k2} + \ldots + \sum_k n_{kN} = \sum_k (n_{k1} + \ldots n_{kN})$$
$$= \sum_k 1 = K \quad (16)$$

Thus the first element of the modified second row equals the first element of the first row. Continuing in the same way shows that the entire second row equals the first row, and hence the determinant vanishes by property (c) of ref 2. Using

$$m_{k1} + m_{k2} + \ldots + m_{kM} = 1 \quad (17)$$
$$k = 1, 2, \ldots, K$$

Table I. Least-Squares Types (1-Substituent Site)

Function used	Condition on z's: Linear: $a_1z_1 + a_2z_2 + \ldots + a_nz_n = 0$			Affine: $a_1z_1 + \ldots + a_nz_n + a_0 = 0$		
	Use condition on z's directly	Use linear condition on α's	Use affine condition on α's	Use condition on z's directly	Use linear condition on α's	Use affine condition on α's
Linear: $f = \alpha_1z_1 + \ldots + \alpha_nz_n$	(1) equivalent (2) equivalent (3) Trouble if one ignores condition on z's			(4) Can ignore condition on z's	(5) Too restrictive not equivalent to original problem	(6)
Affine: $f = \alpha_1z_1 + \ldots + \alpha_nz_n + \alpha_0$	(7) equivalent (8) equivalent (9) equivalent (10) Trouble if one ignores condition on z's			equivalent	(11) equivalent	(12)

in an analogous way shows that rows 1 and $(N + 2)$ through $(N + M + 1)$ are also linearly dependent.

It is convenient to introduce here a distinction that will be important later. If the quantities $z_1, z_2, \ldots, z_n$ are related by

$$a_1z_1 + a_2z_2 + \ldots + a_nz_n = 0 \tag{18}$$

they are said to be linearly dependent; but if

$$a_1z_1 + a_2z_2 + \ldots + a_nz_n + a_0 = 0 \tag{19}$$

where a_0 is a nonzero constant, we shall say that they are affinely dependent. This last is not common terminology; but the distinction between eq 18 and 19 is analogous to that between linear and affine[12] transformations.

The two affine dependencies, eq 15 and 17, lead to two linear dependencies in the rows of eq 14, and hence the least-squares equations, eq 13, do not have a unique solution. To remedy this, eq 15 and 17 can be solved for n_{k1} and m_{k1} and the results substituted into eq 11 to give

$$f_k(\text{calcd}) = (\mu + x_1 + y_1) + \sum_{i=2}^{N}(x_i - x_1)n_{ki} + \sum_{j=2}^{M}(y_j - y_1)m_{kj} = \mu' + \sum_{i=2}^{N}x_i'n_{ki} + \sum_{j=2}^{M}y_i'm_{kj} \tag{20}$$

Best values of the $(N + M - 1)$ independent parameters $\mu', x_2', \ldots, y_M'$ can then be determined by least squares.[2]

The molecule with $X = X_1$ and $Y = Y_1$ has $n_{k2} = n_{k3} = \ldots = n_{kN} = m_{k1} = \ldots = m_{kM} = 0$, and eq 20 shows that the calculated activity of this compound is simply μ'. X_1 and Y_1 might be taken to be hydrogen so that constants $x_2 \ldots y_M$ give, analogously to the Hammett treatment of σ constants, activity of a substituent relative to hydrogen. This is the Fujita–Ban[13] variant of the Free–Wilson model and can be derived alternatively by imposing the two conditions

$$x_H = y_H = 0 \tag{21}$$

on the substituent constants in place of the conditions on the variables n_{ki} and m_{kj} given by eq 15 and 17. Then eq 11 and 20 become identical, and $\mu = \mu'$ = calculated activity of the unsubstituted compound.

In the Cammarata variant,[14] μ of eq 11 is set equal to the observed activity of the unsubstituted compound. Such an assumption might appear to be quite similar to the Fujita–Ban condition, but there are important differences that can be understood by studying first a simpler problem. There is a further point that this will also clarify: the actual physical relation imposed by nature in QSAR is on the variables n_{ki} and m_{kj}, but it is ignored in practice and a restriction placed instead on the least-squares parameters μ, x_i, y_j. It must be shown when and why the two are equivalent. Suppose the function to be fit by least squares in eq 12 is either a linear function

$$f_k(\text{calcd}) = \alpha_1z_{k1} + \alpha_2z_{k2} + \ldots + \alpha_nz_{kn} \tag{22}$$

or an affine function

$$f_k(\text{calcd}) = \alpha_1z_{k1} + \alpha_2z_{k2} + \ldots + \alpha_nz_{kn} + \alpha_0 \tag{23}$$

where $z_{k1}, \ldots, z_{kn}$ is the kth set of measurements of the experimental variables $z_1, \ldots, z_n$ and $\alpha_1, \ldots, \alpha_n, \alpha_0$ are parameters whose best values are to be found. Suppose further that the variables are restricted either by the linear relation of eq 18 or the affine relation of eq 19 and that these are either to be used directly or replaced by linear or affine relations on the α's. The possibilities to be examined are shown in Table I.

If the linear relation, eq 18, were ignored in case 1 of Table I, the resulting least-squares equations (summation over the index k is implied)

$$\begin{aligned}
(\Sigma z_{k1}^2)\alpha_1 + (\Sigma z_{k1}z_{k2})\alpha_2 + \ldots + (\Sigma z_{k1}z_{kn})\alpha_n &= \Sigma z_{k1}f_k \\
(\Sigma z_{k2}z_{k1})\alpha_1 + (\Sigma z_{k2}^2)\alpha_2 + \ldots + (\Sigma z_{k2}z_{kn})\alpha_n &= \Sigma z_{k2}f_k \\
\ldots \qquad\qquad \ldots & \\
(\Sigma z_{kn}z_{k1})\alpha_1 + (\Sigma z_{kn}z_{k2})\alpha_2 + \ldots + (\Sigma z_{kn}^2)\alpha_n &= \Sigma z_{kn}f_k
\end{aligned} \tag{24}$$

would be linearly dependent like eq 13, their determinant would vanish, and they could not be solved uniquely for all the α_i. Using instead eq 18 to eliminate z_1 gives the independent least-squares equations

$$\begin{aligned}
(\Sigma z_{k2}^2)\alpha_2' + (\Sigma z_{k2}z_{k3})\alpha_3' + \ldots + (\Sigma z_{k2}z_{kn})\alpha_n' &= \Sigma z_{k2}f_k \\
\ldots \qquad\qquad \ldots & \\
(\Sigma z_{kn}z_{k2})\alpha_2' + (\Sigma z_{kn}z_{k3})\alpha_3' + \ldots + (\Sigma z_{kn}^2)\alpha_n' &= \Sigma z_{kn}f_k
\end{aligned} \tag{25}$$

where

$$\alpha_i' = \alpha_i - \alpha_i\alpha_1/\alpha_1 \qquad i = 2, 3, \ldots, n \tag{26}$$

The parameter α_1' is not defined, but it is not needed. If the coefficients a_i in the linear relation eq 18 are changed, there is no change in the least-squares equation eq 25 and, hence, no change in the computed values of the α_i'. The original parameters will depend upon the coefficients of

eq 18 because of eq 26. Nevertheless, the equality

$$f(\text{calcd}) = \alpha_1 z_1 + \ldots + \alpha_n z_n = \alpha_2' z_2 + \ldots + \alpha_n' z_n \quad (27)$$

(derived by substituting eq 18 into eq 22 and using eq 26) shows that the computed values of f are independent of the particular linear relation eq 18.

Now consider case 2 where the condition on the z's (eq 18) is replaced by a linear constraint on the α's. These parameters must then satisfy this constraint plus the $(n - 1)$ linearly independent conditions of eq 24

$$\begin{aligned} &c_1\alpha_1 + c_2\alpha_2 + \ldots + c_n\alpha_n = 0 \\ &(\Sigma z_{k2}z_{k1})\alpha_1 + (\Sigma z_{k2}^2)\alpha_2 + \ldots + (\Sigma z_{k2}z_{kn})\alpha_n = \Sigma z_{k2}f_k \\ &\qquad\ldots \qquad\qquad\qquad \ldots \\ &(\Sigma z_{kn}z_{k1})\alpha_1 + (\Sigma z_{kn}z_{k2})\alpha_2 + \ldots + (\Sigma z_{kn}^2)\alpha_n \\ &\quad = \Sigma z_{kn}f_k \end{aligned} \quad (28)$$

which can be solved uniquely for the α's by Cramer's rule. It is not at sight obvious that the f's computed with these α's will be independent of the arbitrary coefficients c_i nor that the result will be the same as obtained in case 1. Notice that the $(n - 1)$ variables $z_2, z_3, \ldots, z_n$ in the right equality of eq 27 can be specified independently. Therefore, in order that f be independent of the c_i, each α_i' must be independent of these c's. Solving eq 28 for α_1 and α_2, then using eq 26 gives

$$\alpha_2' = -\frac{\begin{vmatrix} 0 & (a_1c_1 + a_2c_2) & c_3 & \ldots & c_n \\ \Sigma z_{k2}f_k & (a_1\Sigma_{k2}z_{k1} + a_2\Sigma z_{k2}^2) & \Sigma z_{k2}z_{k3} & \ldots & \Sigma z_{k2}z_{kn} \\ & \ldots & \ldots & & \\ \Sigma z_{kn}f_k & (a_1\Sigma_{kn}z_{k1} + a_2\Sigma z_{kn}z_{k2}) & \Sigma z_{kn}z_{k3} & \ldots & \Sigma z_{kn}^2 \end{vmatrix}}{\alpha_1 \begin{vmatrix} c_1 & c_2 & \ldots & c_n \\ \Sigma z_{k2}z_{k1} & \Sigma z_{k2}^2 & \ldots & \Sigma z_{k2}z_{kn} \\ & \ldots & & \\ \Sigma z_{kn}z_{k1} & \Sigma z_{kn}z_{k2} & \ldots & \Sigma z_{kn}^2 \end{vmatrix}} \quad (29)$$

Expanding these determinants by their first rows gives

$$\alpha_2' = \frac{A_1c_1 + A_2c_2 + \ldots + A_nc_n}{B_1c_1 + B_2c_2 + \ldots + B_nc_n} \quad (30)$$

where $A_1 \ldots A_n$ and $B_1 \ldots B_n$ involve $(n - 1) \times (n - 1)$ cofactors. Examination of these quantities using eq 18 shows

$$\frac{A_1}{B_1} = \frac{A_2}{B_2} = \ldots = \frac{A_n}{B_n} = R$$

where

$$R = \frac{\begin{vmatrix} \Sigma z_{k2}f_k & \Sigma z_{k2}z_{k3} & \ldots & \Sigma z_{k2}z_{kn} \\ \Sigma z_{k3}f_k & \Sigma z_{k3}^2 & \ldots & \Sigma z_{k3}z_{kn} \\ & \ldots & \ldots & \\ \Sigma z_{kn}f_k & \Sigma z_{kn}z_{k3} & \ldots & \Sigma z_{kn}^2 \end{vmatrix}}{\begin{vmatrix} \Sigma z_{k2}^2 & \Sigma z_{k2}z_{k3} & \ldots & \Sigma z_{k2}z_{kn} \\ \Sigma z_{k2}z_{k3} & \Sigma z_{k3}^2 & \ldots & \Sigma z_{k3}z_{kn} \\ & \ldots & \ldots & \\ \Sigma z_{kn}z_{k2} & \Sigma z_{kn}z_{k3} & \ldots & \Sigma z_{kn}^2 \end{vmatrix}} \quad (31)$$

Thus

$$\alpha_2' = \frac{R(B_1c_1 + \ldots + B_nc_n)}{(B_1c_1 + \ldots + B_nc_n)} = R \quad (32)$$

which is independent of the c's. A similar treatment holds for the other α_i''s showing that the computed values of f are, in fact, independent of the linear relation imposed on the α's. Further, R, the value of α_2' in case 2 equals α_2' in case 1 as seen by applying Cramer's rule to eq 25. Cases 1 and 2 are therefore equivalent; either can be used and will give the same computed values of f.

The analysis of case 3 is similar to that of case 2. The only change is that the zeros in the first line of eq 28 and in the first element of eq 29 are replaced by $-c_0$. This adds to the numerator of eq 32 a term in c_0 whose coefficient is

$$-\begin{vmatrix} (a_1\Sigma z_{k2}z_{k1} + a_2\Sigma z_{k2}^2) & \Sigma z_{k2}z_{k3} & \ldots & \Sigma z_{k2}z_{kn} \\ (a_1\Sigma z_{k3}z_{k1} + a_2\Sigma z_{k3}z_{k2}) & \Sigma z_{k3}^2 & \ldots & \Sigma z_{k3}z_{kn} \\ & \ldots & \ldots & \\ (a_1\Sigma z_{kn}z_{k1} + a_2\Sigma z_{kn}z_{k2}) & \Sigma z_{kn}z_{k3} & \ldots & \Sigma z_{kn}^2 \end{vmatrix} = 0 \quad (33)$$

Hence computed f values are independent of c_0 and case 3 is equivalent to case 2.

Case 4 is unique. Solving the affine relation eq 18 for z_1 gives

$$f_k(\text{calcd}) = \alpha_2'z_{k2} + \ldots + \alpha_n'z_{kn} + \alpha_0' \quad (34)$$

where

$$\alpha_0' = \alpha_1a_0/a_1 \text{ and } \alpha_i' = \alpha_i - \alpha_1a_i/a_1 \quad (35)$$
$$i = 2, \ldots, n$$

Minimizing with respect to the n parameters $\alpha_0, \alpha_2, \ldots, \alpha_n$ leads to the normal equations

$$\begin{aligned} &(\Sigma z_{k2})\alpha_2' + (\Sigma z_{k3})\alpha_3' + \ldots + (\Sigma z_{kn})\alpha_n' + K\alpha_0' = \Sigma f_k \\ &(\Sigma z_{k2}^2)\alpha_2' + (\Sigma z_{k2}z_{k3})\alpha_3' + \ldots + (\Sigma z_{k2}z_{kn})\alpha_n' \\ &\quad + (\Sigma z_{k2})\alpha_0' = \Sigma z_{k2}f_k \\ &\qquad\ldots \qquad\qquad\qquad \ldots \\ &(\Sigma z_{kn}z_{k2})\alpha_2' + (\Sigma z_{kn}z_{k3})\alpha_3' + \ldots + (\Sigma z_{kn}^2)\alpha_n' \\ &\quad + (\Sigma z_{kn})\alpha_0' = \Sigma z_{kn}f_k \end{aligned} \quad (36)$$

The values of $\alpha_2', \ldots, \alpha_n', \alpha_0'$ and, hence, of $f_k(\text{calcd})$ are clearly independent of $a_1, \ldots, a_n$. If instead, eq 18 is ignored, eq 24 is obtained. Then using eq 18, and treating the first row of eq 24 in the same way as the first row of eq 14 was treated above, gives for the modified first row of eq 24

$$(\Sigma z_{k1})\alpha_1 + \ldots + (\Sigma z_{kn})\alpha_n = \Sigma f_k \quad (37)$$

which is not a linear combination of the other rows. Substituting $z_{k1} = -[a_2z_{k2} + \ldots + a_nz_{kn}]/a_1$ into the first column of modified eq 24 and collecting like terms give a set of equations identical with eq 36. Thus in case 4 the conditions on the z's can be ignored and still give correct results.

However, imposing further conditions on the α's as in cases 5 and 6 is not equivalent to the original problem. Such restriction on the parameters must lead to poorer (i.e., larger) values of S in eq 12.

The remaining cases can now be treated easily. Cases 10 and 12 are like cases 1 and 2 with an additional variable (z_0) which always has the value 1. Case 10 is therefore equivalent to case 12. The linear condition on the α's in case 11 is a special case of the affine condition

$$c_0\alpha_0 + c_1\alpha_1 + \ldots + c_n\alpha_n + c_{n+1} = 0 \quad (38)$$

in case 12 with $c_{n+1} = 0$. But computed f's in case 12 are independent of these c's; therefore cases 11 and 12 are equivalent. Further, cases 7, 8, and 9 are just special cases of 10, 11, and 12 with $a_0 = 0$; but results in 10, 11, and 12 are independent of the a's. Hence cases 7, 8, and 9 are all equivalent to each other and to cases 10, 11, and 12.

Results in Table I can now be applied to QSAR. The Fujita–Ban variant[13] of the Free–Wilson model for molecules with one substituent site corresponds to case 11 with the actual affine condition on the substituent numbers (eq 15) replaced by the linear condition

$$x_H = 0 \quad (39)$$

on the substituent parameters (corresponding to the α's in Table I). With more substituent sites, there is an affine relation like eq 15 for each; and each is replaced by setting the corresponding hydrogen substituent equal to zero. As shown in the derivation of Table I, the computed drug activities are all equal to those that would have been obtained had eq 15 been used directly. In the Fujita–Ban variant, as we have seen, the parameter μ equals the calculated activity of the hydrogen-substituted compound.

In the Cammarata variant (method 3 of ref 14) μ is set equal to the observed activity of the hydrogen-substituted compound in addition to using eq 39. Fixing μ is equivalent to doing a least-squares analysis of the linear function

$$f_k(\text{calcd}) - \mu = \alpha_1 z_{k1} + \ldots + \alpha_n z_{kn} \quad (40)$$

with an affine condition on the z's. With one substituent site this corresponds exactly to case 4 in Table I. The further use of eq 39 gives the unnecessarily restrictive case 5. With more than one substituent site, one could fix μ and set $X_H = 0$ for all but one site; putting $X_H = 0$ for all sites is again too restrictive.

The original Free–Wilson variant[7] is also like case 11 with the conditions on the parameters being

$$x_1\sum_k n_{k1} + \ldots + x_N\sum_k n_{kN} = 0 \quad (41)$$

$$y_1\sum_k m_{k1} + \ldots + y_M\sum_k m_{kM} = 0 \quad (42)$$

for the two-site example discussed above. With one site, eq 42 is dropped, while for more sites further analogous conditions are added. If eq 11 is now summed over k using eq 41 and 42

$$\sum_{k=1}^{K} f_k(\text{calcd}) = \sum_{k=1}^{K}\mu + \sum_{k=1}^{K}\left(\sum_{i=1}^{N} x_i n_{ki} + \sum_{j=1}^{M} y_j m_{kj}\right)$$

$$= K\mu + 0 + 0$$

$$\mu = \frac{1}{K}\sum_{k=1}^{K} f_k(\text{calcd}) \quad (43)$$

it is seen that μ equals the average of the calculated activities where the average is taken over the set of measured compounds. Applying the same conditions to the first line of eq 13 shows that μ also equals the average of the observed activities of the same compounds.

If instead the conditions

$$x_1 + x_2 + \ldots + x_N = 0$$

$$y_1 + y_2 + \ldots + y_M = 0 \quad (44)$$

are used[2] in the two-site case, the parameter μ again equals the average calculated activity, but the average is taken over all possible molecules that can be constructed with the given set of X and Y substituents, not over the molecules actually measured as in the previous case. To see this, it is convenient to change notation from f_k to f_{ij} where the first of the two subscripts indicates the X substituent, and the second the Y. Then

Table II. Free–Wilson Matrix for Compounds I

	X		Y		
Compd	H	CH_3	$N(CH_3)_2$	$N(C_2H_5)_2$	f(obsd)
1	1	0	1	0	2.13
2	1	0	0	1	1.28
3	0	1	1	0	1.64
4	0	1	0	1	0.85

$$f_{ij}(\text{calcd}) = \mu + x_i + y_j \quad (45)$$

summing over all $N \times M$ possible compounds and using eq 44

$$\sum_{i=1}^{N}\sum_{j=1}^{M} f_{ij}(\text{calcd}) = \sum_{i=1}^{N}\sum_{j=1}^{M}(\mu + x_i + y_j)$$

$$= NM\mu + M\sum_{i=1}^{N} x_i + N\sum_{j=1}^{M} y_j$$

$$= NM\mu + 0 + 0 \quad (46)$$

completes the proof. Analogous results hold for an arbitrary number of substituent sites.

Note also that, as Kubinyi and Kehrhahn[10] point out, the parameter μ is often described as the activity of the molecule minus its substituents, but this is never the exact interpretation.

Linear dependence of the type above is inherent in the Free–Wilson model and can always be anticipated and removed as described. There is a second type of linear dependence that may or may not occur in both the Free–Wilson and the Hansch methods. Consider again a two-site Free–Wilson example. Suppose X = Cl and Y = CH_3 occur only once in the molecules tested, and suppose they are both in the same molecule. The single observed activity for this molecule obviously cannot yield unique values of the two parameters X_{Cl} and Y_{CH_3}. The two columns corresponding to these parameters in the determinant of eq 24 will be identical, and the determinant will vanish. In this simple case the difficulty can be removed by combining the substituents to make a single group contribution. More complicated examples of this kind of linear dependence, which has been discussed by Craig,[9] are not easily recognized on inspection and cannot be treated with combined substituent parameters; but in all cases continued vanishing of the least-squares determinant, after dependencies of the first type have been removed, will indicate such difficulty.[2] The origin of type two linear dependence is in the choice of experimental data and can be removed either by making more measurements or by excluding some of those originally considered.

Examples. The set of four analgesics (I) studied by Free and Wilson[7] makes a convenient example since all calculations can easily be done by hand. Following Free and Wilson and treating the two substituent sites as nonequivalent gives the Free–Wilson matrix of Table II.

In the notation of eq 11, the calculated activity is

$$f_k(\text{calcd}) = \mu + x_1 n_{k1} + x_2 n_{k2} + y_1 m_{k1} + y_2 m_{k2} \quad (47)$$

where $x_1 = x_H$, $x_2 = x_{CH_3}$, $y_1 = y_{N(CH_3)_2}$, and $y_2 = y_{N(C_2H_5)_2}$. If conditions eq 15 and 17 on the independent variables are ignored, and one attempts to find best values of μ, x_1,

Table III. Comparison of QSAR Modifications

Compd	f(obsd)	f(calcd): Direct use of eq 49 or Fujita-Ban or Free-Wilson conditions	f(calcd): Cammarata conditions	Parameter values: Using eq 49	Free-Wilson	Fujita-Ban	Cammarata
1	2.13	2.115	2.13	$x_2' = -0.46$	$x_1 = 0.23$	$x_1 = 0$	$x_1 = 0$
2	1.28	1.295	1.30	$y_2' = -0.82$	$y_2 = -0.23$	$x_2 = -0.46$	$x_2 = -0.83$
3	1.64	1.655	1.66	$\mu' = 2.115$	$y_1 = 0.41$	$y_1 = 0$	$y_1 = 0$
4	0.85	0.835	0.83		$y_2 = -0.41$	$y_2 = -0.82$	$y_2 = -0.47$
$\Sigma[f_k(\text{calcd}) - f_k(\text{obsd})]^2$		9×10^{-4}	12×10^{-4}		$\mu = 1.475$	$\mu = 2.115$	$\mu = 2.13$

x_2, y_1, and y_2, the normal equations corresponding to eq 24 are

$$\begin{aligned} 4\mu + 2x_1 + 2x_2 + 2y_1 + 2y_2 &= 5.90 \\ 2\mu + 2x_1 + y_1 + y_2 &= 3.41 \\ 2\mu + 2x_2 + y_1 + y_2 &= 2.49 \\ 2\mu + x_1 + x_2 + 2y_1 &= 3.77 \\ 2\mu + x_1 + x_2 + 2y_2 &= 2.13 \end{aligned} \quad (48)$$

Two linear dependences are obvious in eq 48: the sum of the second and third equations equals the first, as does the sum of the fourth and fifth. Hence eq 48 cannot be solved uniquely for the five parameters.

The conditions on the independent variables are

$$\begin{aligned} n_{k1} + n_{k2} &= 1 \\ m_{k1} + m_{k2} &= 1 \end{aligned} \quad (49)$$

This corresponds to case 10 of Table I and leads to

$$\begin{aligned} f_k(\text{calcd}) &= (\mu + x_1 + x_2) + (x_2 - x_1)n_{k2} + (y_2 - y_1)m_{k2} \\ &= \mu' + x_2'n_{k2} + y_2'm_{k2} \end{aligned} \quad (50)$$

where the variables n_{k2} and m_{k2} are independent. Minimizing eq 12 with respect to μ', x_2', and y_2' gives the normal equations

$$\begin{aligned} K\mu' + (\Sigma n_{k2})x_2' + (\Sigma m_{k2})y_2' &= \Sigma f_k \\ (\Sigma n_{k2})\mu' + (\Sigma n_{k2}^2)x_2' + (\Sigma n_{k2}m_{k2})y_2' &= \Sigma n_{k2}f_k \\ (\Sigma m_{k2})\mu' + (\Sigma m_{k2}n_{k2})x_2' + (\Sigma m_{k2}^2)y_2' &= \Sigma m_{k2}f_k \end{aligned}$$

or

$$\begin{aligned} 4\mu' + 2x_2' + 2y_2' &= 5.90 \\ 2\mu' + 2x_2' + y_2' &= 2.49 \\ 2\mu' + x_2' + 2y_2' &= 2.13 \end{aligned} \quad (51)$$

whose solution gives the results in Table III.

If instead of the natural conditions eq 49 on the independent variables, the Free–Wilson conditions

$$\begin{aligned} 2x_1 + 2x_2 &= 0 \\ 2y_1 + 2y_2 &= 0 \end{aligned} \quad (52)$$

are applied to the parameters as in case 11, Table I

$$f_k(\text{calcd}) = \mu + (n_{k2} - n_{k1})x_2 + (m_{k2} - m_{k1})y_2 \quad (53)$$

Then minimizing with respect to μ, x_1, and y_1 gives the normal equations

$$\begin{aligned} K\mu + [\Sigma(n_{k2} - n_{k1})]x_2 + [\Sigma(m_{k2} - m_{k1})]y_2 &= \Sigma f_k \\ [\Sigma(n_{k2} - n_{k1})]\mu + [\Sigma(n_{k2} - n_{k1})^2]x_2 + [\Sigma(n_{k2} - n_{k1}) \times (m_{k2} - m_{k1})]y_2 &= \Sigma(n_{k2} - n_{k1})f_k \\ [\Sigma(m_{k2} - m_{k1})]\mu + [\Sigma(m_{k2} - m_{k1})(n_{k2} - n_{k1})]x_2 + [\Sigma(m_{k2} - m_{k1})^2]y_2 &= \Sigma(m_{k2} - m_{k1})f_k \end{aligned}$$

or

$$\begin{aligned} 4\mu &= 5.90 \\ 4x_2 &= -0.92 \\ 4y_2 &= -1.64 \end{aligned} \quad (54)$$

Results are in Table III where it is seen, in accord with the general discussion above, that the calculated activity of each compound is the same using eq 49 as using eq 52.

The unsubstituted compound with X = Y = H is not included in the set studied. Nevertheless, as Kubinyi and Kehrhahn[10] point out, the Fujita–Ban method can be applied using any compound as reference. Choosing compound 1 of Table II gives for the Fujita–Ban conditions

$$x_1 = 0, y_1 = 0 \quad (55)$$

This is again case 11 of Table I and gives

$$f_k(\text{calcd}) = \mu + n_{k2}x_2 + m_{k2}y_2 \quad (56)$$

with normal equations identical to eq 51, except that the primes on the parameters are dropped. Results in Table III again show the same values of f_k(calcd), as expected.

With the Cammarata conditions, corresponding to case 5 of Table I

$$\mu = f_{\text{ref}}(\text{obsd}) = 2.13 \quad (57)$$

is imposed in addition to eq 55. The normal equations are then

$$\begin{aligned} (\Sigma n_{k2}^2)x_2 + (\Sigma n_{k2}m_{k2})y_2 &= \Sigma n_{k2}(f_k - 2.13) \\ (\Sigma m_{k2}n_{k2})x_2 + (\Sigma m_{k2}^2)y_2 &= \Sigma m_{k2}(f_k - 2.13) \end{aligned}$$

or

$$\begin{aligned} 2x_2 + y_2 &= -1.77 \\ x_2 + 2y_2 &= -2.13 \end{aligned} \quad (58)$$

Table III shows that this gives, as expected, values of f_k(calcd) different from the other methods and a poorer fit.

Free and Wilson's example is an interesting one since, if the two substituent sites in compound I are recognized as equivalent, the two conditions of eq 49 must be replaced by the single condition

$$n_{k1} + n_{k2} + m_{k1} + m_{k2} = 2 \quad (59)$$

which is not sufficient to remove both linear dependences of eq 48. The remaining linear dependence is now viewed as being due to a poor choice of experimental compounds rather than as inherent in the model. This would not happen in the more usual case where the two sets of substituents contain common members since eq 48 is then

replaced by normal equations with fewer unknowns.

Kubinyi and Kehrhahn give a number of more typical examples, and in particular their Table VI (ref 10, p 1045) shows results parallel to our Table III.

References and Notes

(1) These include at least one of our own, and we are grateful to Drs. H. Kubinyi and O. T. Kehrhahn of Knoll, AG, Ludwigshafen, for pointing out that the Free-Wilson conditions are eq 41 and 42 of the present paper rather than eq 4 of ref 2.
(2) L. J. Schaad, R. H. Werner, L. Dillon, L. Field, and C. E. Tate, *J. Med. Chem.*, 18, 344 (1975).
(3) F. B. Hildebrand, "Methods of Applied Mathematics", Prentice-Hall, New York, N.Y., 1952, Chapter 1.
(4) G. E. Shilov, "Linear Algebra", Prentice-Hall, Englewood Cliffs, N.J., 1971.
(5) C. Hansch, *Acc. Chem. Res.*, 2, 232 (1969), and earlier references cited in this review.
(6) R. F. Gould, Ed., "Biological Correlations–The Hansch Approach", American Chemical Society, Washington D.C., 1972.
(7) S. M. Free, Jr., and J. W. Wilson, *J. Med. Chem.*, 7, 395 (1964).
(8) W. P. Purcell, G. E. Bass, and J. M. Clayton, "Strategy of Drug Design", Wiley, New York, N.Y., 1973.
(9) P. N. Craig, Chapter 8 in ref 6.
(10) H. Kubinyi and O. H. Kehrhahn, *J. Med. Chem.*, 19, 578, 1040 (1976).
(11) H. Kubinyi, *J. Med. Chem.*, 19, 587 (1976).
(12) R. Courant, "Differential and Integral Calculus", Vol. II, Interscience, New York, N.Y., 1936, p 27.
(13) T. Fujita and T. Ban, *J. Med. Chem.*, 14, 148 (1971).
(14) A. Cammarata and S. J. Yau, *J. Med. Chem.*, 13, 93 (1970).

Quantitative Structure-Activity Relationships. 7.[1] The Bilinear Model, a New Model for Nonlinear Dependence of Biological Activity on Hydrophobic Character

Hugo Kubinyi

Research Institute of Knoll AG, D 6700 Ludwigshafen/Rhein, Federal Republic of Germany. *Received October 5, 1976*

The bilinear model, $\log 1/C = a \log P - b \log (\beta P + 1) + c$, a new model for nonlinear dependence of biological activity on hydrophobic character, is applied to 57 data sets of biological activity values in homologous series. From a comparison of the statistical parameters and the residuals obtained with the bilinear model and the parabolic model, the superiority of the bilinear model for a precise quantitative description of both linear and nonlinear parts of structure–activity relationships can be derived; the bilinear model explains the particular effect that in homologous series the relationship between biological activity and hydrophobic character is strictly linear for the lower members, while for higher members this relationship is nonlinear.

The biological response elicited by a drug is determined by its intrinsic activity and by its ability to reach a definite receptor site. While the intrinsic activity of a drug molecule depends on various physicochemical properties and the geometry of the molecule, drug permeation is considered to be—in most cases—a passive transport process which is influenced only by the lipophilicity of the molecule. Thus, in homologous series, where the intrinsic activity of all members can be assumed to be identical biological activity should be a simple function of lipophilicity.

Indeed linear relationships between biological activity and hydrophobic character (eq 1)[2-8] are obtained for a large number of homologous series. However, this linear relationship cannot go on infinitely, otherwise compounds with infinite activity would exist. A cut-off point[9] is reached in each homologous series: biological activity increases with increasing lipophilicity, reaches a maximum, and then decreases with further increase of hydrophobic character.

$$\log 1/C = a \log P + b \quad (1)$$

In 1964 Hansch[5-7] proposed a parabolic model (eq 2) for the dependence of biological activity on hydrophobic character on the basis of a "random walk" process; on the way from the outer phase, where the drug is applicated, to their receptor sites the drug molecules have to penetrate a number of lipophilic and hydrophilic barriers. While hydrophilic molecules tend to remain in the aqueous phases and lipophilic molecules tend to go into the lipid (membrane) phases, molecules with an optimal hydrophilic–lipophilic balance will have the best chance to penetrate all barriers and to reach the receptor sites.

$$\log 1/C = a(\log P)^2 + b \log P + c \quad (2)$$

Although the parabolic model has been supported by consideration of a kinetic model[7,10,11] and its suitability for practical purposes has been proven with some hundred examples, there remains a discrepancy between the linear model and the parabolic model; from the linear model at least the left side of the "parabola" should be strictly linear, while from the parabolic model both sides of the structure–activity relationship should be more or less curved.

Besides the parabolic model several other models for nonlinear dependence of biological activity on hydrophobic character have been presented in the last years.[12-16] Among these models the most interesting seems to be the Mc Farland model:[12] Mc Farland used a simple hypothetical system, made up of alternating aqueous phases and lipid (membrane) phases of equal volumes and considered the probability of a drug molecule to cross a definite number of barriers. The Mc Farland model can be represented by eq 3[17] (corresponds to eq 15 of ref 12). Symmetrical curves with linear ascending and descending sides and a parabolic part within the range of $\log P = 0$ are obtained from eq 3. Although Mc Farland recognized systematic deviations between his model and the parabolic model (the same deviations can be recognized between the computer plot from the kinetic model[7,10,11] and the parabola fitted to this plot), he considered the parabolic model to be sufficient for all practical purposes.

$$\log 1/C = a \log P - 2a \log (P + 1) + c \quad (3)$$

To adapt the Mc Farland model to biological reality, the model was reconsidered using a slightly modified system (Figure 1).[17] Only four relevant phases are regarded, e.g., as a model of a simple bacterial cell or an isolated tissue