Multivariate Quantitative Structure-Activity Relationships (QSAR): Conditions for Their Applicability

SVANTE WOLD*† and WILLIAM J. DUNN III‡

Research Group for Chemometrics, Institute of Chemistry, Umeå University, S-901 87 Umeå, Sweden, and
Department of Medicinal Chemistry, College of Pharmacy, University of Illinois at the Medical Center,
Chicago, Illinois 60680

Received May 26, 1982

For a QSAR, as well as any other scientific model, to have predictive relevance, the number of estimated parameters (P) must be substantially smaller than the effective number of degrees of freedom (DOF) in the data. Conditions related to the rule $P << \mathrm{DOF}$ are reviewed and exemplified for various multivariate data analytic methods commonly applied in QSAR: multiple regression (MR), factor analysis (FA), principal components analysis (PCA), and pattern recognition methods such as linear discriminant analysis (LDA), linear learning machine (LLM), Bayesean methods (Bayes), K nearest-neighbor rules (KNN), and the SIMCA method.

INTRODUCTION

Quantitative structure-activity relationships (QSAR) are quantitative models which relate the variation in measures of biological activity (BA) in a series of chemical compounds to the variation in chemical structure (CS) between the compounds in the series. Reviews of various aspects of QSAR are given in the literature.[1-3]

While the BA is usually obtained in a rather straightforward way, the variation in CS must be quantified before a QSAR can be formulated. This quantification is an interesting problem in itself but will not be much discussed in the present paper. Thus we assume that the variation in CS in a set of compounds for which a QSAR is to be developed has been translated to the variation of a number of structure-indicator variables as described, for instance, in ref 1-5.

The models used in QSAR range from simple additivity models, in the present context called Free-Wilson models,[6] to complicated models of pattern recognition relating multivariate BA data to multivariate CS data.[7-10] The BA of chemical compounds is usually (but not always) a result of complicated processes. It is rare that a single structural indicator is sufficient to explain the variation in the BA within a set of compounds. Hence, multivariate models are usually needed in QSAR.

These models give quantitative BA predictions (MR), qualitative BA predictions (pattern recognition), or both (generalized pattern recognition), as indicated in Table I. The qualitative predictions are given in the form of a class assignment for each compound, e.g., active vs. nonactive or agonist vs. antagonist vs. nonactive. This class assignment is for some methods of pattern recognition of a "fuzzy" nature; i.e., a compound is given a probability for its belonging to each of the classes.

For a QSAR to be of predictive relevance, i.e., allow the prediction of the BA of new compounds not included in the "training set" (see below), certain conditions must be fulfilled. These conditions of mainly statistical character do, of course, apply also in other problem areas where multivariate data analysis is applied. In the present paper we review these conditions which must be fulfilled in order for the BA predictions to be better than chance, i.e., nonrandom. In this context we discuss the data-analytic methods most commonly used in QSAR (see Table I).

PREMISES

We have a training set of N chemical compounds, all assumed to have known CS. For each of these compounds, one or several measures of the BA have been collected. The BA may be quantitative, e.g., log $1/C$, or qualitative, say, agonist vs. antagonist or active vs. inactive. One may also have a combination of qualitative and quantitative measures, for instance active vs. inactive *plus* perhaps the measured potency for some of the active compounds.

In some way we have managed to translate the variation in CS between the N compounds to the variation of M structure indicator variables. These may include measured or calculated parameters related to lipophilicity, e.g., log P,[1-4] or polarizability, say molecular refractivity.[1-4] Solubility, vapor pressure, and quantum mechanical and molecular mechanics "indices" such as bond orders and charges at certain atoms are often useful indicators. In addition, substituent parameters characteristic for parts of the molecules such as Hammett's σ[11,12] Hansch's π,[1-4] Taft's E_s,[12,13] and Verloop's sterimol parameters[14] are used to describe region-specific interactions. Thus we arrive at the data table shown in Figure 1.

Though the selection of compounds in the training set, the design and collection of the biological measurements, and the quantitative description of the structural variation are the three most fundamental and crucial parts of a QSAR study, we shall not discuss these problems here, but instead we shall concentrate on the statistical, data-analytic part of the problem.

SCOPE OF QSAR

The primary objective of QSAR is to predict the BA for new compounds in the test set (see Figure 1). In case such predictions are found to be substantially better than random, the QSAR also provides some understanding of the cause of the BA in question. The prediction objective corresponds to two conditions that must be fulfilled by the model and the data: (a) The model should reproduce the BA (Y) as well as possible for the compounds in the training set. (b) This should be done with few parameters (P) compared with the number of degrees of freedom (DOF) in the data.

Condition b is included because of the well-known property of empirical models to be able to exactly reproduce a given data set when the number of parameters, P, equals or exceeds the number of data elements.[15] Thus the primary condition for applicability of QSAR is formulated as in expression 1.

$$P \ll \mathrm{DOF} \qquad (1)$$

We shall review below various multivariate data-analytic models with respect to their number of adjustable parameters, P, and the way they utilize the data in terms of degrees of freedom, DOF.

We define the level of triviality (LOT) of a study as the point when $P = \mathrm{DOF}$. At this point and beyond ($P > \mathrm{DOF}$),

[†] Umeå University.
[‡] University of Illinois at the Medical Center.

0095-2338/83/1623-0006$01.50/0

Table I. Multivariate Data Analytic Methods Commonly Used in QSAR[a]

method and scope	TR	SV	SG	DG	DA
multiple regression (MR); quantitative linear relation between BA (Y) and CS (X)	QUANT	Y	Y	plots	Y
linear discriminant analysis (LDA) and the linear learning machine (LLM); separate the classes on the basis of CS (X)	CLASS	Y	Y	Y	Y
K nearest neighbors (KNN); find *K* (usually 3) nearest neighbors to compound in X space	CLASS	N, S	N, S	Y	N
Bayesean methods; describe the multivariate distribution of X for each class	CLASS FUZZY	Y	I, S	Y	N
SIMCA; model similarities between compounds in the same class by means of separate PC models for each class; on levels 3 and 4 also quantitative prediction of BA[9]	CLASS FUZZY QUANT	N, S	I, S	N	N
principal components analysis (PCA) and factor analysis (FA); model relations between variables x_i and compounds by a single model for the whole data set	CLASS QUANT	N	N	N	not appl

[a] The columns display the following information (Y = yes, I = intermediate, N = no, S = "provided the scaling of the variables x_i is *not* based on their class separation ability"): TR = type of results; QUANT = quantitative BA prediction; CLASS = class assignment, FUZZY = fuzzy class assignment (probability level for each class); SV = sensitive to many variables per compound and to multicollineariteis in CS matrix X; SG = sensitive to strong grouping in the data; DG = difficulty to detect strong groups in the data; DA = difficulty to handle the asymmetric situation with one diffuse class and one well defined class.[48] In MR this corresponds to inhomogeneity of the variance of y.

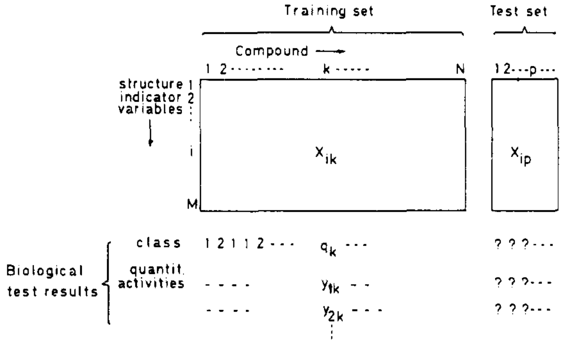

Figure 1. Data in a QSAR study are divided into a training set on which the quantitative model is "calibrated" and a test set for which the biological activities (BA) are predicted by means of the previously developed model. In the training set, data are defined both for the structure indicators, x_i, and for the biological activities, class q and test values y_j.

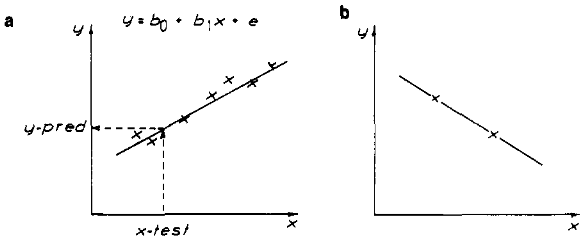

Figure 2. In the simplest case of multiple regression (MR), the biological activity, y, is modeled as a linear relation of the single structure indicator x. The regression coefficients b_0 and b_1 are calculated to make the model fit the training set data as well as possible. The value for the biological activity of a compound in the test set, y-pred, is predicted by inserting the x value of the compound, x-test, into the model with $e = 0$. (b) Level of triviality (LOT) in MR: $N = M + 1$. A line can always be drawn to fit two points exactly.

the predictions of the model with the calculated parameters are not better than random, even if the fit of the model to the training set data looks good. Thus, such results are spurious, trivial, and fortuitous.

SELECTING THE TYPE OF MODEL

In the case when the BA is qualitatitive, i.e., BA = pharmacological class, some method of pattern recognition (PaRC) is used to achieve a relation between the class assignment and the indicators X.[7,16,17] In statistics such analysis is often called discriminant analysis.[18] If, in addition, the pharmacological potency of the compounds or some other BA has been quantitatively measured for some classes (we denote this BA with Y), a regression type model (see below) is used for each of these classes to achieve a quantitative relation between X and Y.[1-3]

We note that all statistical models discussed here (Table I), with KNN as a possible exception, are based on the fact that complicated functional relationships between CS and BA can be locally linearized, i.e., approximated by simple mathematical models or rules in limited intervals:

$$X \pm \Delta \text{ and } Y \pm D \qquad (2)$$

Therefore, a QSAR study ranging over a wider variety of structure and/or activity must necessarily involve multiple models. In such cases classification methods (PaRC) are needed. Thus, one must first be able to predict the structural or pharmacological *type* (class) before a quantitative relationship between CS and a quantitatively measured BA can be achieved. In consequence, generalized methods of PaRC[9,10] are important in QSAR since they can be used both to classify and to quantitatively model the relation between CS and BA.

COMMON MODELS USED IN QSAR

We shall below shortly review the multivariate models commonly used in QSAR. Thereby we shall use graphical illustrations based on the *simplest possible* case, hoping that the generalizations to more realistic cases are obvious.

(1) Multiple Regression (MR). This model relates *one* measure of BA, denoted y, to one or several structure indicator variables, x_i, by the linear model in eq 3 (here b denotes parameters and regression coefficients and e deviations and residuals, see Figure 1 for further notation).

$$y_k = b_0 + \sum_i b_i x_{ik} + e_k \qquad (3)$$

This model corresponds to fitting an M-dimensional plane to the observation points of the training set in the (M + 1)-dimensional space formed by y and the M x variables. In the simplest case with one x variable, the regression model is a straight line as shown in Figure 2a.

The number of degrees of freedom (DOF) in the data is N, since one y value is used per observation. The number (P) of adjustable parameters b is M + 1, one for each x_i plus one for the constant term. Hence expression (1) for multiple regression becomes

$$M + 1 \ll N \text{ or } M \ll N - 1 \qquad (4)$$

Consequently, the level of triviality (LOT) for MR is when M + 1 equals N, i.e., when the number of structure-indicator variables equals the number of compounds minus one. For the simplest case of a straight line, the LOT is shown in Figure 2b.

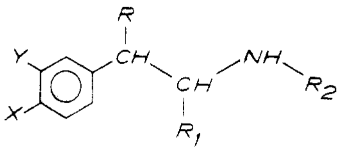

Figure 3. Free–Wilson model for the agonist activity of the set of phenethylamines from Lefkowitz et al.,[19] SIMCA analyzed by Dunn et al.,[20] would require 14 zero–one descriptor variables. This is because the site Y has two, R has two, R_1 has three, and R_2 has eleven different substituents, respectively. This make 14 parameters plus the constant term, i.e., 15. Hence the number of parameters equals the number of available compounds in the training set ($n_1 = 15$). Consequently, the parameters cannot be independently estimated from the available data.

As discussed briefly in the Remedy section below, MR models may be applied and estimated also in situations with more variables than observations without using variable selection. One then cannot get information about the influence on y of each individual x variable. Still, good predictions of y from X are obtained, which is one essential objective in QSAR. Also, some limited information is obtained about the relevance of each x variable, but no longer in the strict MR way.

The Free–Wilson model used in QSAR[6] is a special case of the MR model. Here J substitutents at G different sites in a compound are varied. An additive model with one parameter for each substituent (except one; see ref 2) and site is formulated as eq 5 (notation is analogous to that above).

$$y_k = b_0 + \sum_{i}^{G(J-1)} b_i x_{ik} + e_k \qquad (5)$$

Each variable x_i is specific for one site and one substituent, having the value 1 when the substituent at site g is in state j, otherwise it is zero. The coefficients b describe the additive influence of the ith combination of site and substituent (see example in Figure 3). Consequently, for Free–Wilson models, the LOT is reached when $G \times (J - 1) = N$.

(2) Linear Discriminant Analysis (LDA) and Linear Learning Machine (LLM). These PaRC methods[16-18] are often used when the BA is a qualitative variable describing the class of a compound, say, agonist or antagonist or active or nonactive. Usually the problem is formulated as one or several binary decisions, i.e., two-class problems. For such a two-class problem, the methods work by assigning weights c_i to the variables x_i so that the weighted sum S is positive as far as possible for members of class 1 and negative as far as possible for members of class 2 (eq 6).

$$S = c_0 + \sum_i c_i x_{ik} \qquad (6)$$

In the M-dimensional space formed by the x variables, this corresponds to the calculation of an ($M - 1$)-dimensional plane which separates the two classes (the training set points) as well as possible. Figure 4a shows the simplest case with two x variables. As with MR, one y value (the class number in the present case) per compound is used in the estimation. Hence expression 1 is the same as for MR, as well as the level of triviality (Figure 4b). LDA and LLM are called regression-like methods since they have similar mathematical properties and can be formulated in a way similar to MR.[18] Thus, for LDA and LLM, rule 1 is the same as for MR (eq 4 above) and LOT is $M = N - 1$.

As with regression analysis, generalized inverse solutions to LDA can be used when the number of variables is large compared with N (Remedy section).

(3) *K* Nearest Neighbor (KNN). This method of PaRC (see ref 17 and 21–23 for details) does not form any explicit model for the training set classes. Instead, for new compounds in the test set, the distances d_{kj} (see Figure 5) to all compounds in the training set are calculated. A test set compound is classified as belonging to the class of the majority of the K nearest neighbors in the training set. K values of 1 or 3 are mostly used.

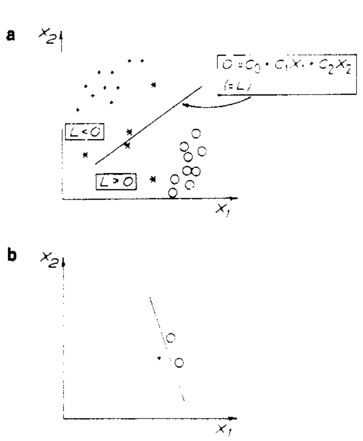

Figure 4. (a) ($M - 1$)-dimensional hyperplane (line when $M = 2$) of linear discriminant analysis (LDA) and the linear learning machine (LLM) separating two classes (+ and O). The equation of the hyperplane is defined by the M parameters c_0, c_1, ..., c. The value of the discriminating function L is zero on the hyperplane, positive in the region of one class, and negative in the region of the second. Test set compounds are classified by inserting their structure indicator values x_i into the discriminating function. This corresponds to finding out on which side of the hyperplane their observation points (asterisks) are situated. (b) LOT for LDA and LLM: $M + 1 = N$.

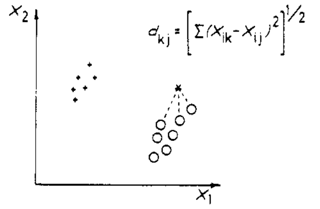

Figure 5. In the K nearest-neighbor (KNN) method, test set compounds (asterisk) are classified according to the majority of its K nearest neighbors in the training set. Usually K is 3 or 1. The "nearness" is computed by the euclidean distance d_{kj} or other more sophisticated distance measures such as the Mahalanobis distance.[21-23]

The KNN method involves no adjustable parameters, and therefore the level of triviality corresponds to one compound per class. However, since distances in M space depend on the scaling of the variables, the KNN results are heavily scale dependent. If the scaling of the variables is now made to improve the class separation in the training set, then parameters in the form of scaling factors are manipulated, and LOT changes. Often so-called Fisher weighting[24,25] is used. Then, each variable is given a scaling factor which is proportional to the ratio of its between-class variance and its within-class variance. In such cases M parameters (scaling factors) are calculated, and LOT is $M = N$. Hence, for KNN (with two classes) the LOT is $N = 2$ when the scaling is not conditioned on class separation and $N = M$ with Fisher weighting or other scaling based on class separation.

(4) Principal Components, Factor Analysis, and SIMCA. In principal components analysis (PCA) and factor analysis (FA) one fits a single model to the whole training set.[27-29] Hence for PCA and FA, $n_q = N$ in the formulas below. This analysis corresponds to the projection of the data set down on an A-dimensional hyperplane. With $A = 1$ and $A = 2$ this plane can be directly visualized, which gives a display of the relative positions of the training set compounds. New compounds are thereafter projected down on the same plane and classified according to into which region they fall in the projection. The LOT of FA and PCA is treated below together with SIMCA.

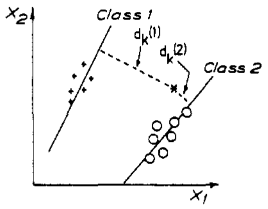

Figure 6. In the "calibration" phase of the SIMCA method (soft independent modeling of class analogy), separate principal component (PC) models are least-squares fitted to the n_q observations of each class training set. The number of terms in each class model is determined by cross-validation.[30-32] In the simplest case, the PC models correspond to straight lines in the measurement space as shown above. A compound in the test set is thereafter classified according to its distances to the different class models $d_k^{(q)}$. These distances are calculated by least-squares fitting the test compound data vector (x) to each of the "calibrated" class models.

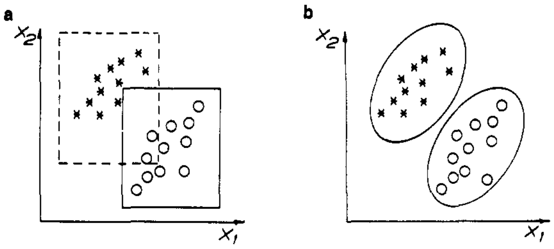

Figure 7. (a) Bayesean class domains for a given probability level (usually 0.95), computed from the mean and standard deviation of each variable in each class training set. (b) More realistic Bayesean class domains based, in addition, on the covariances $s(x_i,x_j)$.

In the SIMCA method, one constructs instead an independent hyperplane (PC model) for each class in the training set.[7,9,30-32] In the simplest case with two x variables this corresponds to fitting a line to the points of each class in the training set (Figure 6). New compounds are classified according to their distances $d_k^{(q)}$ to the class models q. These distances are determined by using multiple linear regression.

The number of estimated parameters is $(M + (M + n_q - 1)A_q)$ for each class q. PCA and FA can be seen as one-class problems and can therefore be incorporated in the same discussion. Here M is, as usual, the number of variables, n_q the number of compounds in class q, and A_q the number of components used in the PC model of class q. The estimation is based on $M \times n_q$ data elements. The minimum value of A_q is 1. Usually the scaling of the variables in PCA and SIMCA is not conditioned on the class separation. FA is scale independent. Hence LOT for SIMCA, PCA and FA is as shown in eq 7 (with $A_q = 1$).

$$2M + n_q - 1 = Mn_q \qquad (7)$$

When $M \gg n_q$, this reduces to $n_q = 2$. We realize that like KNN (but unlike LDA and LLM), SIMCA, FA, and PCA can be used also in cases when the number of variables, M, greatly exceeds the number of compounds, N. In fact, the classification stability of SIMCA increases with the number of variables as long as they are relevant to the given problem. This is because the distances $d_k^{(q)}$ in Figure 6 are based on M residuals e_{ik}, making $d_k^{(q)}$ more stable with a rate proportional to the square root of M.

(5) Bayes. With Bayesean methods of PaRC,[22-26] one computes for each class a probability density function (PDF). From these PDFs one then computes "confidence domains" for each class, inside which new compounds should fall in order to be assigned to the corresponding class. Usually, however, one assigns a new compound to the class of highest probability.

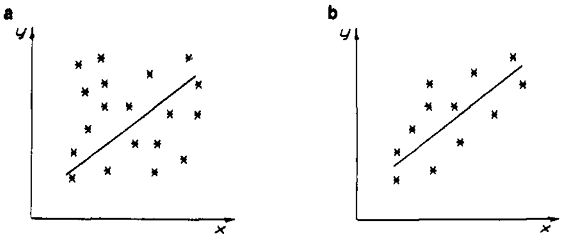

Figure 8. By deleting points that "do not fit" in part a, an apparent agreement between the data and the regression model is reached in part b.

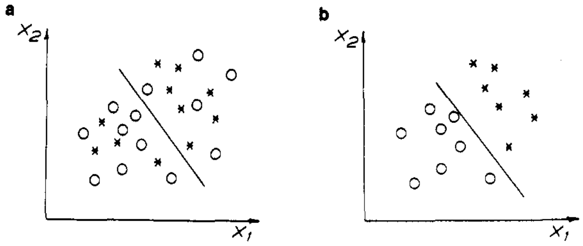

Figure 9. By deleting points that "do not fit" in part a, an apparent agreement between the data and the discriminant model is reached in part b.

In the simplest case this PDF involves only the mean and standard deviation (SD) of each variable (Figure 7a). This involves $2M$ parameters per class estimated from the $M \times n_q$ data of the class. When the variables are nonorthogonal, this PDF is too simplistic, and the covariances between the variables should be included. The number of parameters then increases to $(2M + M(M - 1)/2)$ per class (Figure 7b). Most Bayesean methods are independent of scaling (with the exception of ALLOC; see ref 26) but not of the selection of variables, as discussed below. Hence, depending on the sophistication of the Bayesean PDFs, the LOT is

$$\text{simplest case:} \quad 2M = Mn_q \text{ or } 2 = n_q$$

$$\text{realistic case:} \quad 2M + M(M - 1)/2 = Mn_q \text{ or } 2 + (M - 1)/2 = n_q$$

SOME WAYS TO REACH THE APPARENT FULFILLMENT OF $P \ll$ DOF

Though the rule $P \ll$ DOF seems to be straightforward, man's desire to see results consistent with his preconcieved views causes many data analyses to be made in such a way that the rule is not fulfilled. The results then, of course, have little value. In particular, predictions based on the analysis are no better than those based on mere guessing. The results are then random, trivial, spurious, or fortuitous. In multivariate data analysis there are some ways to reach an apparent fulfillment of the $P \ll$ DOF rule, while the rule is actually far from being fulfilled. This makes the results particularly misleading since their triviality is not easily seen. We shall therefore discuss these ways below in some detail.

(a) Selection of Compounds That Fit a Preconcieved Model. This is the simplest, probably most common, and definitely most difficult-to-check way to "cheat" the $P \ll$ DOF rule. Consider Figures 8 and 9. By selecting points that "fit" and explaining away the others or just simply forgetting them, it is very simple to create data sets that behave nicely but which actually are just cases of self-deception.

To editors of chemical journals who hesitate to publish reports of the analysis of data sets from the literature it can be pointed out that in such reports one can at least check that all compounds from the original literature sources are included. This is not possible with articles which report both the "data collection", e.g., biological testing, and the data analysis.

(b) (1) Selection of Variables. A popular way to increase the chance of good fit of a model to a data set is to (i) include very many variables x_i (M), (ii) select a small number of these (m) on the basis that they make the model fit (MR) or the classes separate (PaRC), and (iii) report the results and compute "confidence levels" as if the initial number of variables actually was m and not M.

This procedure has even been given some credibility by applied statisticians who call it stepwise multiple regression or stepwise discriminant analysis. Though the statisticians seldom fail to point out the dangers with the procedure and the inflated levels of "confidence", the availability of black box computer programs to users who have not read the warnings has produced many reports of apparently significant but actually spurious results.

The reason this type of variable selection is dangerous (if one wants results of predictive relevance) is that there are very many ways to select a combination of m variables out of M. In fact the number of ways is given by eq 8 (where ! denotes

$$\binom{M}{m} = M!/((m!(M-m)!) \tag{8}$$

factorial). For example, 5 variables can be selected from 100 in (100 × 99 × 98 × 97 × 96/2 × 3 × 4 × 5) different ways (approximately 10^8 ways). The statistical significance levels in MR and PaRC are based on the assumption that the selection of variables is arrived at by insight and understanding of the actual problem or as a consequence of the actual physical situation, i.e., *independently* of the analysis.

Thus by selecting variables by a procedure conditioned on the fit of the model to the data, one actually has a larger number of parameters P than one realizes or admits.

$$\text{real } P > \text{apparent } P$$

If now the apparent P is smaller than DOF, one believes that the rule $P \ll$ DOF is fulfilled while actually it might be far from being so.

In the QSAR field, Topliss et al. have made a commendable analysis of the variable-selection problem with MR analysis.[33] They also show examples of spurious results achieved by having too many variables without really realizing this. The results and discussion are valid also for the application of LDA and LLM since they are similar in nature to MR. The problem is the same whenever variable selection is conditioned on class separability in PaRC.

(2) Remedy. Interestingly, there is a straightforward solution to the problem of variable selection. This is to use generalized regression models and PaRC methods which tolerate any number of variables M. These methods are usually based on the representation of the M-dimensional X matrix by a smaller m-dimensional matrix. This representation is calculated in a way which does not utilize the dependent variable y in an incorrect manner.

Generalized inverses,[34,35] principal components regression,[36,46] ridge regression,[37,46] and path modeling with latent variables[38] provide such alternatives to MR. These methods also solve the multicollinearity problem in MR.[46] The SIMCA method is a PaRC method based on the same principle. KNN is also independent of the number of variables, but unlike SIMCA, KNN provides no direct measures of the relevance of the x variables.

(c) Dependence Among the Observations (Compounds, Objects). In the case above (selection of variables), the apparent fulfillment of the rule $P \ll$ DOF was reached by having an apparent P smaller than the actual P. The other popular way is to have a larger apparent N, and thereby a larger DOF, than the actual value.

Special cases of this situation are (a) when the BA data (y) in a MR model have very little variation or (b) when the large majority of compounds in a pattern-recognition study belongs to one class, for instance, active compounds, and only a few compounds belong to a second class, for instance, inactive compounds.

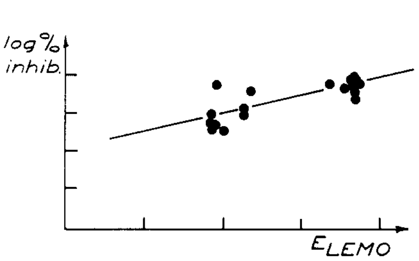

Figure 10. Plot of log (percent inhibition) of the compounds A–X against the calculated E-LEMO: left cluster, compounds A–Ph–Z; right cluster, compounds A–Z.

In case a it is then easy to construct a model by using any independent set of CS variables X, a model that "fits" the BA data (y) well because there is no variation in y to describe. A simple analysis of variance,[46] however, reveals the statistical nonsignificance of such modeling.

In case b it is in the same way simple to construct a pattern-recognition rule that correctly classifies almost all of the compounds in the large class and some of the compounds in the small class. This can be done, for instance, by assigning a compound to the larger class unless certain rare features are present in the compound, features selected from some of the compounds in the smaller class. Consequently, one should test for significant proportions of correct predictions in *both* classes separately.

A more subtle way to achieve apparent success is to have *several* strong dependencies among the compounds in the training set. Consider the following example from the QSAR literature: The inhibition of an in vitro system was measured for $N = 23$ compounds A–X. The logarithm of the percent inhibition (y) was related to the calculated energy difference between the lowest unoccupied molecular orbital (E-LUMO) by a linear model. The resulting regression is apparently highly significant ($R = 0.705$, $p < 0.01$).

However, statistical significance levels are based on the assumption that the observations are *independent*. Figure 10 shows a plot of the data and the linear model. One realizes immediately that one does not have $N = 23$ independent compounds, but rather closer to $N = 2$, which is the LOT for the MR model.

By having two subsets of compounds in the study which differ substantially but trivially in E-LUMO, one creates the impression of a real correlation which actually is trivial. In this case one subgroup consists of compounds where the "substituents" X are ordinary small groups such as methyl and chloro and a second subgroup where X is phenyl–Z with Z being various ordinary substituents.

In the MR situation it is fairly easy to check for this situation of strong subgroups in the data by making simple plots. The relation which "explains" the whole data set must, to be real, be applicable also within each subgroup. In Figure 10 we see that within the subgroups there is no simple relation between y and E-LUMO.

In PaRC it is often more difficult to see strong subgroups in plots. Hence dependencies between observations are sometimes not detected. Consider, as an example, the following PaRC analysis from the QSAR literature. $N = 219$ compounds with either a tranquilizer (T) or a sedative (S) effect were described by $M = 69$ variables and subjected to a PaRC analysis with the LLM method. This data set is notorious in the QSAR literature. No less than four analyses have been reported, all actually being trivial as described below.

The training set is separated by LLM to 100%. This perfect separation should catch ones attention in itself. A validation using an artificial test set gave 90% correct classification, an apparently impressive result. However, when the chemical

Table II. Chemical Subclasses of the $N = 219$ Compounds Classified as Tranquilizers (TRANQ) or Sedatives (SEDAT)[a]

subgroup		distribution	
		TRANQ	SEDAT
phenothiazines and analogues		*88*	3
indoles		*20*	1
other heterocycles	benzodiazepines	*21*	2
	barbiturates	0	*27*
	butyrophenones	*3*	1
	other	4	*23*
aromatic	diphenylmethanes	*2*	0
	benzoic acids	0	*1*
	other	2	*3*
aliphatic	glycols	0	*2*
	carbamates	0	*4*
	carbinols	0	*3*
	amides	0	*7*
	other	0	*2*

[a] With the use of 69 structure indicators, LLM classifies the training set ($N = 219$) correctly while artificial test sets, obtained by deleting ten compounds from the training set at the time, were classified about 90% correctly. The majority compound percentages are italic.

structures of the 219 compounds are inspected (Table II), one realizes that the data set is strongly grouped into chemical subclasses. Moreover, the distribution within each subclass is highly uneven so that the tendency is strong for each subclass to contain *either* mainly tranquilizers (subgroups 1, 2, 3, 5, and 7) *or* mainly sedatives (subgroups 4, 6, 10, ...). If we add up the number of minorities in all subgroups, we arrive at the surprisingly small number of 13.

Hence, if we manage to recognize merely the chemical subclass of a compound we shall correctly classify all but 13 of the 219 compounds. This goes also for the validation phase where compounds are left out of the training set and then treated as an artificial test set. This is because the uneven distribution of activity type within the subclasses is not disturbed by leaving one out except in the very small subclasses.

If we now use a PaRC method which maximizes class separation, such as LLM, we need at most 14 variables of the 69 to recognize the subclasses of all the compounds. The remaining 55 can be used to "classify" the 13 minority compounds which will give 100% "success" since we have more parameters than observations.

In the validation phase, the 13 minority compounds are, of course, classified randomly. This gives an expected rate of about 200 correct of 219 (90%) if the results were *completely trivial.* This expected rate is computed by including the 13 minority compounds plus all subclasses smaller than subclass 7 in the "randomly" classified ensemble. The concordance between the LOT and the results actually reported is striking.

We note that we have different expected rates of classification corresponding to LOT for different PaRC methods. Hence, in cases when more than one PaRC method has been used, we can compare several expected LOT rates with the actual outcome. Thus the confidence of the reality or spuriousness of the results is increased.

In the validation phase (leaving compounds out of the training set), LLM will classify a left out compound correctly if it is a "majority compound" and has a 50% chance of classifying it correctly if it is a "minority compound". The other PaRC methods will classify only the "majority compounds" correctly, since their probability density functions or class models are based on these majority compounds. Hence LLM will give apparently better results than the other PaRC methods in the case of strongly grouped data sets.

Sampling Artifacts. An interesting variant of dependence between compounds is the situation called the "fallacy of probabilities" by Unger[43b] (in statistics called sampling artifacts). Examples in the QSAR field are referred to in ref 39 and 43b.

The situation arrives when one attempts to construct a predictive QSAR scheme based on the occurrence of substructural fragments in compounds of diverse structures tested for some type of BA. Consider, for example, carcinogenicity. A large number of compounds, say N, have been screened by the government agency Z with the result that n of the compounds are noncarcinogenic and $N - n$ are carcinogenic, according to some rule of cut off. A computer analyst from the budget department in Z is hired to try to make some kind of probabilistic model. He proceeds as follows.

A number of substructural fragments, e.g., methyl, ethyl, phenyl, cyclooctyl, secondary amine, etc. are collected into a "basis set" which adequately describes the N compounds. The frequency of occurrence of each fragment in each compound is recorded. Subsequently, the total number of occurrences of each fragment in the "class" of carcinogens is calculated and divided by $N - n$ to give the "probability of carcinogenicity" (p_+) for each fragment. In the same way, the "probability of noncarcinogenicity" (p_-) is computed from the fragment occurrences in the n noncarcinogens.

It is now claimed that the activity of new compounds can be predicted by noting which fragments occur in their structure and then calculating a "carcinogenicity score" by multiplying together their p_+ scores and dividing by their p_- scores.

For the medicinal chemist it is clear that a scheme of this type cannot work particularly well, since carcinogenicity (or any other BA) is not caused by substructural fragments but by whole molecules where it matters very much both which combination of fragments is present and at which relative positions they are situated. How come, then, the scheme seems to work when the statistics are calculated? The reason is an interesting interplay between the screening agency (Z) and the scientific community providing the compounds.

We note that the calculated probabilities depend on the frequency of certain fragments in the "training set" which, in turn, depend on the interest of the scientific community in constructing various types of compounds. When, for instance, it was found in the 1970's that some nitrosoamines are carcinogenic, a large number of these compounds were synthesized and "screened". Since many of these indeed were carcinogenic, the fragment N—N=O gets a high p_+ and a moderate p_-. When one then attempts to validate the predictive scheme, it will show "statistically significant" predictive rates, since nitrosoamines often are carcinogenic in the way people make them today. If, however, in 1985, somebody notes that some nitrosoamines without α-hydrogens (mainly noncarcinogens) are potent antiviral agents, medicinal chemists will make large numbers of such compounds. These will usually turn out to be noncarcinogens in the screens in Z.

This dependency of a model on the current interest of medicinal and synthetic chemists is not very desirable. In fact, the predictive results of the model are *not* statistically significant, when correctly evaluated. The apparent predictive power is just a sampling artifact. It is remarkable that such "models" presently are used in U.S. government agencies and also are marketed in the U.S. as predictors of various kinds of "environmental impact" of compounds, including carcinogenicity.

The simplest way to test the *real* predictive power of such a model is to evaluate its behavior on a structurally homogeneous class of compounds, say nitroso amines. If the predictions within the class are no better than the sampling rate of active vs. inactive, then the model actually does no better than recognizing the structural class of the compound. This result is trivial to the chemist, who already knows that polycyclic hydrocarbons and nitroso amines are more often carcinogenic than disaccarides and aliphatic hydrocarbons.

To be of any use in QSAR, a model must be better than the level of triviality, not only statistically but also pharmacologically.

Literature Survey. When going through the QSAR literature, approximately 40 papers reporting PaRC analyses of a structure–activity type were found. This survey is reported in a separate paper.[39] A few of the 40 papers could not be checked because the primary data were lacking. Nevertheless, of the ones that could be checked, 19 papers were actually reporting results that were completely trivial because the rule $P \ll$ DOF was not fulfilled in reality. Fourteen of these 19 "spurious" papers were due to the presence of strong subgroups in the data set plus a very uneven distribution of the type of activity in the subgroups. Ten of the 19 papers reported analyses where one initially had very many variables and then selected subsets of variables using their class separation ability as the selection criterion. Hence five of the papers are "doubly trivial" since they use both approaches to achieve the apparent fulfillment of the $P \ll$ DOF rule.

In all cases where an active class contained compounds with "diverse structures" the results were trivial (above LOT). This supports our view that empirical models used in PaRC have a foundation as locally valid linearizations of complicated functional relationships. Since the linearizations are valid only for sets of similar objects (compounds), a class in a PaRC training set can neither be modeled nor be separated from other classes if it contains compounds with diverse (nonsimilar) structures.[40]

CONCLUSIONS

Methods of multivariate data analysis, here taken to include multiple regresson (MR), pattern recognition (PaRC), and principal components and factor analysis (PCA and FA), provide powerful data-analytic tools for QSAR. However, certain conditions must be fulfilled in the data set to make the results of the data analysis be better than trivial.

The most important of these rules are based on the simple fact that for a model to have predictive value, the number of estimated parameters, P, must be appreciably smaller than the number of degrees of freedom in the data set, DOF. Several authors have pointed out the risks involved with variable selection[33] and with the lack of independence among the observations.[41–44] Heilbronner[45] gives an illustrative example of the risk of using the same data both for calibrating and evaluating a model if an appropriate validation[32,47] of the results is not made.

Since data-analytic methods exist that can handle both the regression problem and the pattern-recognition problem when the number of variables, M, exceeds the number of compounds in the training set, N (see Remedy section), there is no need to restrict oneself to the use of a small set of variables nor to use risky methods of variable selection. It is clear that the complicated relations involved in QSAR often demand the characterization of the compounds by a large number of variables. One must make certain that the data analysis is made in a way that minimizes the risk for spurious correlations; i.e., one should use a good statistical method for the data analysis.

This should not be read as to mean that one can resort to sloppy fishing expeditions, throwing all possible and impossible variables into the X matrix and then waiting to se what comes out. On the contrary, irrelevant x variables always introduce noise, which disturbs any data analysis. Hence, the need for insight is as large as ever, but today methods exist that can handle more realistic data sets than before.

It is rather discouraging to see that a high percentage of the QSAR papers using PaRC methodology, approximately 50%, report results which can be shown to be no better than trivial because the fundamental statistical conditions have not been fulfilled.

When the major rules are obeyed, still there are more subtle points not discussed in detail here, points that must be considered to make the results of the multivariate analysis interpretable and useful of making predictions. These points include the following (see also Table I): (a) Homogeneity in the variance of the biological effect variable, y, over the experimental domain (correctable by weighting)[15,46] should be maintained. (b) Independence of the variables x_i[46] should be considered. This is a problem particularly with MR, LDA, and LLM and can be solved by using generalized regression instead of MR and by using KNN or SIMCA instead of LDA and LLM. (c) The classes in a PaRC problem have widely different variance–covariance matrices. This "asymmetry problem"[48] seems to be common in QSAR; the class of "inactive" compounds often lacks systematic structure. This can be handled by KNN, SIMCA, and quadratic discrimination[18] but not by LDA or LLM.

The reader might at this point conclude that the use of statistical methods in QSAR is risky because of the problems with the number of degrees of freedom. He might, consequently, in the future avoid the use of such methods and turn instead to methods which apparently are not plagued by these risks. We feel that this would be a gross misunderstanding. With statistical methods, these problems are pointed out *explicitly*. They can therefore be checked and controlled. In other "nonstatistical" QSAR approaches there still are "parameters" which can be manipulated to make the model fit a given data set. Since these parameters often are not explicitly identified, one might run a still larger risk with these methods since the DOF problem, while still there, is only implicit and therefore not recognized.

In summary, we would like to emphasize the value of empirical multivariate models in the analysis of QSAR data. Such models can be seen as linearizations of complicated functional relationships and therefore have only local validity. Only pharmacologically and chemically similar compounds can be incorporated in each model. Hence data sets incorporating compounds of diverse structure and activity should be modeled by a number of disjoint models, both in regression and classification problems.

Since it is often difficult to specify the precise meaning of pharmacological and chemical similarity in a given application and since the selection of variables and compounds demands understanding and insight to be consistent with the questions put by the researcher, multivariate data analysis provides no simplistic panacea which makes QSAR an automatic success. However, applied with care and common sense and with the fulfilling of basic statistical conditions, multivariate methods of data analysis provide tools which have the right properties to handle the complexities of relations between chemical structure and biological activity.

ACKNOWLEDGMENT

Support from the Swedish Council for Planning and Coordination of Research (FRN) and the Swedish Natural Science Research Council (NFR) is gratefully acknowledged. Helpful comments have been made by a referee.

REFERENCES AND NOTES

(1) Hansch, C. "A quantiative Approach to Biochemical Structure–Activity Relationships". *Acc. Chem. Res.* **1969**, *2*, 232–239.

(2) Martin, Y. C. "Quantitative Drug Design. A Critical Introduction"; Marcel Dekker: New York, 1978.

(3) Seydel, J. K.; Schaper, K.-J. "Chemische Struktur und Biologische Aktivität von Wirkstoffen"; Verlag Chemie: Weinheim/Bergstr., Germany, 1979.

(4) Hansch, C.; Leo, A. J. "Substituent Constants for Correlation Analysis in Chemistry and Biology"; Wiley: New York, 1979.

(5) Cammarata, A.; Menon, G. K. "Pattern Recognition. Classification of Therapeutic Agents According to Pharmacophores". *J. Med. Chem.* **1976**, *19*, 739–748.

(6) Free, S. M.; Wilson, J. W. "A Mathematical Contribution to Structure–Activity Studies". *J. Med. Chem.* **1964**, *7*, 395–399.

(7) Dunn, W. J., III; Wold, S. "Relationships between Chemical Structure and Biological Activity Modelled by SIMCA Pattern Recognition". *Bioorg. Chem.* **1980**, *9*, 505–523.

(8) Mager, P. P. In "Drug Design"; Ariens, E. J., Ed.; Academic Press: New York, 1980; Vol. 9, pp 188–236.

(9) Albano, C.; Dunn, W. J.; Wold, S.; et al. "Four Levels of Pattern Recognition". *Anal. Chim. Acta Comput. Tech. Optim.* **1978**, *103*, 429–443.

(10) Dunn, W. J., III; Wold S.; Edlund, U.; Hellberg, S. "QSAR between Data from a Battery of Biological Tests and an Ensemble of Chemical Descriptors. Mutagenicity Data for Seven Halogenated Aliphatic Hydrocarbons". *Tech. Rep. R. G. Chemometrics, Umeå Univ.* **1982**, *2*, 1–17.

(11) Hammett, L. P. "Physical Organic Chemistry", 2nd ed.; McGraw-Hill: New York, 1970.

(12) Exner, O. "A Critical Compilation of Substituent Constants. Correlation Analysis in Chemistry"; Chapman, N. B., Shorter, J., Eds.; Plenum: London, 1978; pp 437–540.

(13) Taft, R. W. "Steric Effects in Organic Chemistry"; Newman, M. S., Ed; Wiley: New York, 1956.

(14) Verloop, A.; Hoogenstraaten, W.; Tipker, J. "Drug Design"; Ariens, E. J., Ed.; Academic Press: New York, 1976; Vol. 7.

(15) Box, G. E. P.; Hunter, W. G.; Hunter, J. S. "Statistics for Experiments"; Wiley: New York, 1978.

(16) Stuper, A. J.; Brugger, W. E.; Jurs, P. C. "Computer Assisted Studies of Chemical Structure and Biological Function"; Wiley: New York, 1979.

(17) Varmuza, K. "Pattern Recognition in Chemistry"; Springer Verlag: West Berlin, 1980.

(18) Lachenbruch, P. A. "Discriminant Analysis"; Hafner Press: New York, 1975.

(19) Mukherjee, C.; Caron, M. C.; Mulliken, D.; Lefkowitz, R. J. *Mol. Pharmacol.* **1976**, *12*, 16.

(20) Dunn, W. J.; Wold, S.; Martin, Y. C. "Structure-Activity Study of β-Adrenergic Agents Using the SIMCA Method of Pattern Recognition". *J. Med. Chem.* **1978**, *21*, 922–930.

(21) Fix, E.; Hodges, J. L. "Discriminatory Analysis, Non-Parametric Discrimination"; USAF School of Aviation Medicine: Randolph Field, TX, 1951; Project 21-49-004, Report No. 4.

(22) Fukunaga, K. "Introduction to Statistical Pattern Recognition"; Academic Press: New York, 1972.

(23) Andrews, H. C. "Introduction to Mathematical Techniques in Pattern Recognition"; Wiley: New York, 1972.

(24) Kowalski, B. R.; Bender, C. F. "Pattern Recognition. A Powerful Approach to Interpreting Chemical Data". *J. Am. Chem. Soc.* **1972**, *94*, 5632–5639.

(25) Kanal, L. "Patterns in Pattern Recognition: 1968–1974". *IEEE Trans. Inf. Theory* **1974**, *20*, 697–722.

(26) Coomans, D.; Massart, D. L.; Brockaert, I.; Tassin, A. "Potential Methods in Pattern Recognition. Part 1. Classification Aspects of the Supervised Method ALLOC". *Anal. Chim. Acta Comp. Tech. Optim.* **1981**, *133*, 215–224.

(27) Weiner, P. H.; Malinowski, E. R.; Levinstone, A. R. "Factor Analysis of Solvent Shifts in Proton Magnetic Resonance". *J. Phys. Chem.* **1970**, *74*, 4537–4542.

(28) Malinowski, E. R.; Howery, D. G. "Factor Analysis in Chemistry"; Wiley: New York, 1980.

(29) Jöreskog, K. G.; Klovan, J. E.; Reyment, R. A. "Geological Factor Analysis"; Elsevier: Amsterdam, 1976.

(30) Wold, S. "Pattern Recognition by Means of Disjoint Principal Components Models". *Pattern Recognition* **1976**, *8*, 127–139.

(31) Wold, S.; Sjöström, M. "SIMCA a Method for Analyzing Chemical Data in Terms of Similarity and Analogy. Chemometrics, Theory and Application". *ACS Symp. Ser.* **1977**, *No. 52*, 243–282.

(32) Wold, S. "Cross Validatory Estimation of the Number of Components in Factor and Principal Components Models". *Technometrics* **1978**, *20*, 397–406.

(33) Topliss, J. G.; Edwards, R. P. "Chance Factors In Studies of Quantitative Structure–Activity Relationships". *J. Med. Chem.* **1979**, *22*, 1238–1244.

(34) Golub, G.; Kahan, W. "Calculating the Singular Values and Pseudo-Inverse of a Matrix". *J. SIAM Numer. Anal., Ser. B* **1965**, *2*, 205–224.

(35) Marquardt, D. W. "Generalized Inverses, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation". *Technometrics* **1970**, *12*, 591–612.

(36) Hawkins, D. M. "On the Investigation of Alternative Regressions by Principal Components Analysis". *Appl. Stat.* **1973**, *22*, 275–286.

(37) Hoerl, A. E.; Kennard, R. W. "Ridge Regression: Biased Estimation for Nonorthogonal Problems". *Technometrics* **1970**, *12*, 55–67.

(38) Wold, S.; Wold, H.; Dunn, W. J., III; Ruhe, A. "The Collinearity Problem in Linear and Nonlinear Regression. The Partial Least Squares (PLS) Approach to Generalized Inverses". Report 83; Department of Information Science, Umeå University: Umeå, Sweden, 1980; pp 1–33.

(39) Wold, S.; Dunn, W. J., III; Hellberg, S. "Survey of Applications of Pattern Recognition to Structure–Activity Problems. *Tech. Rep. R. G. Chemometrics, Umeå Univ.* **1982**, *1*, 1–12.

(40) Wold, S.; Sjöström, M. In "Correlation Analysis in Chemistry: Recent Advances"; Chapman, N. B., and Shorter, J., Eds; Plenum: New York, 1978; pp 1–54.

(41) Clerc, J. T.; Nägeli, P.; Seibl, J. *Chimia* **1973**, *27*, 639.

(42) Perrin, C. L. "Testing of Computer Assisted Methods for Classification of Pharmacological Activity". *Science (Washington, DC)* **1974**, *183*, 551–552.

(43) (a) Unger, S. H. *Cancer Chemother. Rep.* **1974**, *4*, 45. (b) Unger, S. H. In "Drug Design"; Ariens, E. J., Ed.; Academic Press: New York, 1980; Vol. 9, pp 47–119.

(44) Mathews, R. J. "A Comment on Structure–Activity Correlations obtained Using Pattern Recognition Methods". *J. Am. Chem. Soc.* **1975**, *97*, 935–936.

(45) Heilbronner, E.; Schmelzer, A. "Some Comments on Matching Model Calculations with Experiment Through Linear Regression, or, Your Theory May Be Worse Than You Think". *Nouv. J. Chim.* **1980**, *4*, 23–28.

(46) Draper, N. R.; Smith, H. "Applied Regression Analysis", 2nd ed.; Wiley: New York, 1981.

(47) Stone, M. "Cross-Validatory Choice and Assessment of Statistical Predictions". *J. R. Stat. Soc. B* **1974**, *36*, 111–133.

(48) Dunn, W. J., III; Wold, S. "Structure–Activity Analyzed by Pattern Recognition: The Asymmetric Case". *J. Med. Chem.* **1980**, *23*, 595–599.