BIOORGANIC CHEMISTRY 9, 505–523 (1980)

Relationships between Chemical Structure and Biological Activity Modeled by SIMCA Pattern Recognition

WILLIAM J. DUNN III AND SVANTE WOLD*

*College of Pharmacy, Department of Medicinal Chemistry, University of Illinois at the Medical Center, 833 South Wood Street, Chicago, Illinois 60612, and *Research Group for Chemometrics, Institute of Chemistry, Umea University, S-901 87 Umea, Sweden*

Received June 6, 1979

For a class of chemically and pharmacologically similar compounds, one can formulate a quantitative relation between the variation in chemical structure and the variation in measured biological activity. This relation is based on (1) the translation of the compound structures into quantitative variables by means of *either* substituent parameters derived from the influence of the substituents on chemical model reactions *or* theoretical variables derived from quantum chemical or other theoretical calculations, (2) multivariate statistical models extracting the common "pattern" of the structural descriptor variables for the compounds in the class, and (3) relations between parameters emerging from the statistical models and the measured biological activities of the compounds. In cases when several classes of compounds are studied, the data analysis also involves a classification, and the total analysis becomes one of pattern recognition or discriminant analysis. The methodology is illustrated by means of four examples: the classification and activity prediction of some β-adrenergic compounds, the prediction of the carcinogenicity of some 4-nitroquinoline-1-oxides, the prediction of the carcinogenicity of some polycyclic aromatic hydrocarbons, and the prediction of the glycemic activity of some *o*-toluenesulfonyl (thio)ureas.

INTRODUCTION

The study of relations between the chemical structure and biological activity of chemical compounds is an area of interest both from the basic and from the applied research points of view. Knowledge about the mechanisms of interactions between organic or inorganic molecules on the one hand, and, on the other hand, biological macromolecules is of importance for the fundamental understanding of biochemistry, pharmacology, etc. Applied areas such as environmental chemistry, chemical carcinogenesis, and drug research depend on evaluations of the toxicity, mutagenicity, carcinogenicity, pharmacological effects, and other biological activities of chemical compounds.

A major part of these studies is presently the development of methods which can be used to predict quantitatively the outcome of an interaction of a chemical compound with a living orgnism. A number of such methods have been developed and are being used. One, the Hansch method, was the subject of a recent review

0045-2068/80/040505-19$02.00/0

(*1*). The origins of these methods are in physical organic chemistry where the counterparts are the Brønsted equation (*2*) and the Hammett (*3*) structure-reactivity relationship. Due to the very complex nature of compound–organism interactions, as is also the nature of chemical reactions, the expression of such interactions must be in the form of simple (*4*) or "soft" models (*5*).

The Hansch, Brønsted, Hammett and similar extrathermodynamic relationships can be seen as the quantification of the analogy principle, i.e., "that like substances react similarly and that similar changes in structure produce similar changes in reactivity" (Ref. (*3*), p. 348). Thus, parameters derived from the influence of substituents on chemical model reactions are used to mathematically and statistically model the influence of the same substituents on the chemical or biological "reaction" which is studied.

In the same way, the medicinal chemist attempts to predict the behavior of a biological system upon its interaction with an untried or untested compound from the similarity of this compound with other compounds which already have been studied with respect to this interaction. Thus similarity and analogy form the basis for solution of this and a number of structure-biological activity problems. For example, from the levels of antibacterial response measured on members of a series of similar sulfonamides, the level of antibacterial activity of an unknown sulfonamide can be estimated using the Hansch method. Apart from this, it may be of interest to predict whether an unknown structure will have antibacterial or diuretic activity. So formulated, this is a classification problem, and traditionally such classification problems are approached by comparing the structure of the unknown with those of sulfonamides known to have antibacterial or diuretic activity. As such, classification is also based on similarity and analogy.

Problems of structure–biological activity often involve compounds which are modified at several positions, i.e., have several substituents varied. This makes the mathematical-statistical problem multivariate; the biological activity of the compounds must be modeled in terms of influences from several variables. When, in addition, several models of interaction are possible between substituents and the biological "receptor"—steric effects, electronic effects of different kinds, solvation effects, etc.—the problem becomes even more multivariate; the number of possible variables in the quantitative models rises rapidly.

When part of the problem is formulated as one of classification, the multivariate mathematical-statistical methods are called methods of pattern recognition (PaRC), and PaRC methods have recently been applied to a number of such chemical-biological classification problems (*6*). Of the PaRC methods available for structure biological activity studies, the SIMCA (SImple Modeling of Chemical Analogy) method, which has only recently been developed (*7*), has performed well in such classification studies (*8*). It is this method, its philosophical and mathematical bases, and its utility in structure-biological activity studies that is the subject of this review.

To illustrate its utility, the method will be applied to four examples: three previously reported by us; and one taken from the literature, to allow the comparison of the SIMCA method with a more traditional method of data analysis.

BASICS OF PATTERN RECOGNITION (PaRC)

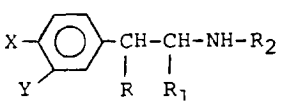
PaRC has the primary objective of producing quantitative rules for classifying objects (here chemical compounds) into one of a number of given classes. These rules are derived in a first phase from objects "known" to belong to either of the classes and quantitative data measured or observed (the training set, learning set, or reference set). This first phase is analogous to the chemical practice of "learning" the regularities of classes of compounds from "model" compounds—for instance learning the "typical" C13-nmr spectrum of phenols from a representative set of spectra of "normal" phenols. In the second phase of PaRC, the derived rules are used to obtain a classification of new, unassigned, objects by applying the rules to the same type of data observed on these objects (the test set).

In the present type of studies, the objects are chemical compounds, the data are derived from their structure (see below), and the classes are related to their type of biological activity. Usually one has only two classes—active or nonactive, carcinogenic or noncarcinogenic, etc.—but sometimes, as in the phenethylamine example below, three or more classes are defined, in this example agonist, antagonist, or neither. In this latter example, the inactive class is unrepresented in the training set, but must still be taken into account.

In addition to rules and results of a classification, other types of information are often desired in the analysis. In structure–activity studies one often has measured not only the type of response (active or inactive, agonist or antagonist), but also the *level* of activity for active compounds in the training set. This level might even exist for a multitude of test systems, so that one has a vector of "effect" variables available. It is then desirable to derive rules relating the descriptor variables of the compounds in the training set to their effect variables. These rules then in the second phase of an analysis can be used to predict the levels of the effect variables of the compounds in the test set. We call this type of analysis Level III or IV of PaRC.

DESCRIBING A COMPOUND'S CHEMICAL STRUCTURE BY MEANS OF A DATA VECTOR

In applying quantitative structure–activity models to a given problem, the compound structures must first be translated into vectors of numbers, data vectors where the variables are structural descriptors. In PaRC applications common practice has been to use zero–one variables describing the occurence–nonoccurence of structural fragments, "substituents." In the phenethylamine example, the 37 compounds had the basic structure I and 22 different substituents occur in the positions X, Y, R, R_1, and R_2, ranging from H, CH_3, and OH to $CH(CH_3)CH_2CH_2C_6H_4$-OH-*p*. Making the doubtful assumption that a substi-



I

tuent has the same influence on biological activity regardless of its position in the molecule, one would get a vector of 21 zero–one elements for each molecule. Taking also the position into account, one would get 27-dimensional zero–one vectors.

The continuity assumptions underlying most methods of PaRC, including SIMCA, are not well fulfilled with the use of zero–one variables. Also, predictions beyond the range of substituents used in the training set are impossible. One can not utilize the chemical knowledge that, for instance, a nitro group in some ways resembles a cyano group.

We have, therefore, preferred to use a description of each substituent which is based on analogy. We use for each position (five positions in structure I) several parameter scales derived from the substituent's influence on model systems or derived from theoretical calculations of the substituent's spatial requirements. Thus we have used the lipophilic constants of Hansch (*11*) (π) or Rekker (*11*) (f), Hammett's scale (*11*) (σ), Taft's electronic (σ^*) scale (*11*) and "steric" scale (E_s) (*11*), and Verloop's size descriptors (*11*) and other analogous scales. Describing the compounds I by means of such variables plus the experimentally derived receptor binding constant, we translate each compound into a 13-dimensional data vector. This type of description makes SIMCA a multivariate generalization of the few variable extrathermodynamic relationships (ETRs) such as the Hansch and Hammett model. This is reassuring with regard to the well-established applicability of the latter to structure-active correlations involving a single substituent.

QUANTITATIVE SIMILARITY AS A BASIS OF CLASSIFICATION

Consider a training set of N compounds divided into Q classes on which data y_{ik} were obtained either by measurement or theoretical arguments. These data form a matrix as in Fig. 1 below.

Each object data vector can be represented as a point in an M-dimensional space (one axis per variable), abbreviated M-space. In the trivial case when each object within a class is identical, the class can within experimental error be represented as a single point in M-space. If the identity requirement is relaxed so that the objects within the class are similar, the class will appear as a cluster of points in M-space. This is illustrated graphically (Fig. 2a) below in a 3-space for convenience.

If another class of objects is introduced on which the same M variables have been obtained a possible graphical result is shown in Fig. 2b. With this representation, PaRC can be considered as the methodology for determining into which of the class clusters, if any, new unclassified object points will fall.

A number of PaRC methods are available. The hyperplane methods, such as the linear learning machine (LLM) and linear discriminant analysis (LDA) derive the equation for an $(M - 1)$-dimensional hyperplane that, when inserted between the class clusters, optimizes their separation. Classification of new objects is then based on which side of the plane the object point falls. The philosophy of these methods is based on searching for *differences* between the classes.

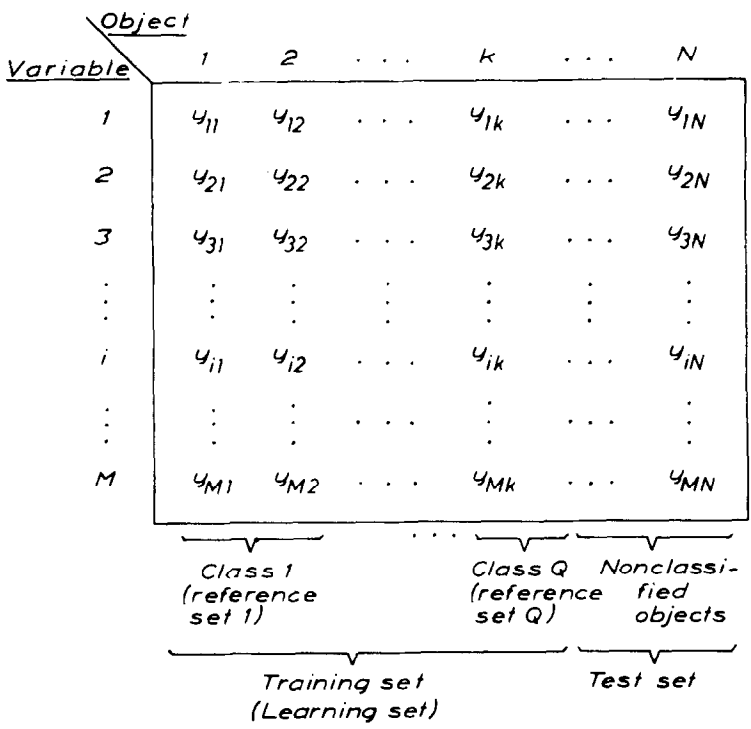


FIG. 1. Data matrix for a Q class pattern recognition problem.

The SIMCA approach to PaRC is the search for the *similarities* within each class. This is done by describing each class by a mathematical model. It has been shown (*7*) that provided some assumptions are fulfilled (see below), the data of a single class (say q) are well approximated by the model in Eq. [1] with few product terms, A. In statistical terminology this is the principal components (PC) or factor model (*12*).

$$y_{ik} = m_i^{(q)} + \sum_{a=1}^{Aq} b_{ia}^{(q)} u_{ak}^{(q)} + e_{ik}^{(q)} \qquad [1]$$

Here i and k are indices referring to variable and object, respectively, m is the mean of the variable, b the loading of the variable, u the object specific component, and e the residuals. For Q classes it may be possible to derive Q different PC models, one for each class. Fitting a PC model to the data of a class consists of approximating the data for that class by an A-dimensional hyperplane (a line when $A = 1$ and a plane when $A = 2$). By calculating the standard deviation (SD) of the residuals of each class, a confidence interval about each class model can be obtained. This encloses the class clusters in mathematical structures as shown in Fig. 3.

The classification of objects is based on determining if object points fall within the mathematical structure of a class; i.e., how *similar* the object is to the class.

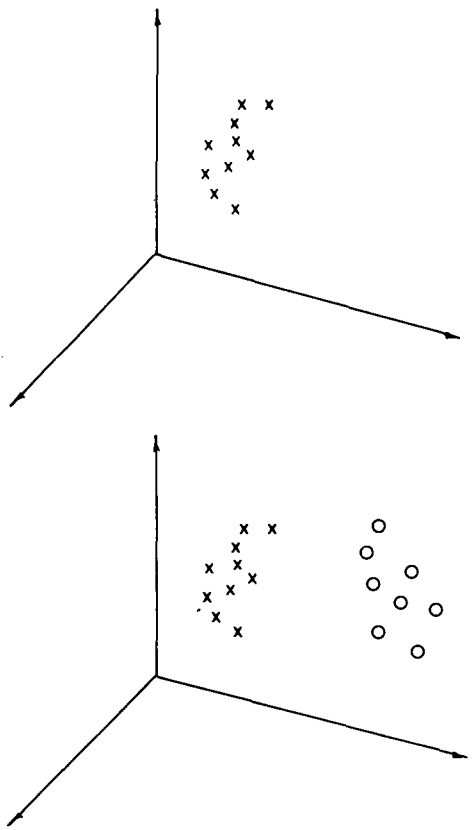


FIG. 2. (a) Cluster of a single class in 3-dimensional space; (b) clustering of two classes in 3-dimensional space.

Mathematically, this corresponds to a simple multiple regression for each class model q and object (with data y_i^*).

$$y_i^* - m_i^{(q)} = \sum_{a=1}^{A_q} b_{ia}^{(q)} v_a^* + e_i^{(q)*} \qquad [2]$$

The regression coefficients v_a^* are determined to minimize the residual (e_i^*) sum of squares. The distance between the object point and the class model q is measured simply by the residual SD.

We note that the SIMCA analysis divides the data y_{ik} into one part, m_i and b_{ia}, which relates only to the variables and one part, u_{ak}, which relates only to the objects. The latter parameters describe the position of each object k within the class. They can be used in a Level III or IV analysis to investigate whether objects (compounds) with similar level of biological activity lie close to each other in the class. If such a systematic behavior is found, the corresponding parameters v_a^* for new objects can be used to get predictions of the level of the biological activity for the new objects.

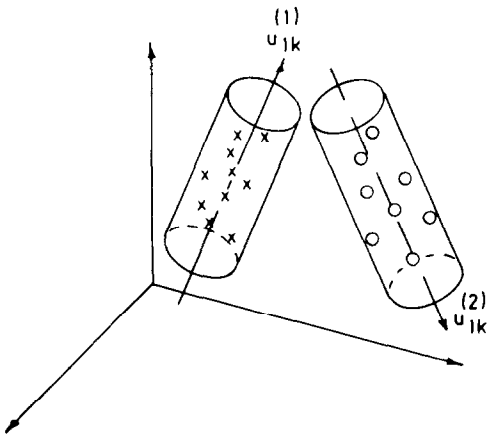


FIG. 3. SIMCA enclosure of the class clusters by mathematical structure.

There are two assumptions on which the derivation of Eq. [1] is based. These are (1) the data y_{ik}, for a class, are generated by a smooth function, and (2) the compounds to which the data apply are "similar" within a class. With measured or calculated chemical data the first assumption is easily justified since the postulates of quantum theory are that any observable property on a chemical system is an eigenvalue of an operator equation. With regard to the second assumption, two types of similarity are important. The first is pharmacological similarity which implies that all objects in a class exert their biological effect by the same biological mechanism. The other type of similarity is structural similarity in the sense that the objects within each class can be described by their data in some way to be chemically similar. Of the two types of similarity, pharmacological similarity presupposes structural similarity.

Provided that these two assumptions above are fulfilled, Eq. [1] has the same approximation properties as a polynomial expansion has of bivariate data $y = f(x)$. The number of product terms, A in Eq. [1], corresponds to the degree of the polynomial. The more complex the variation of the data, the more terms A are needed to describe the class adequately.

DATA STRUCTURE IN M-SPACE

The representation of the data, y_{ik}, in M-space can result in two types of configurations as shown in Fig. 4. The first is characterized by a systematic structure of only one of the classes which can be well described by a PC model. This case we call the *asymmetric* structure (Fig. 4a). The second case, called the *symmetric* case, is characterized by all classes having systematic structure; all classes are well described by separate PC models.

Asymmetric data structures often result when the biological testing results are reported as active vs nonactive. The asymmetry is rather natural, considering the rigid "specificiations" of an active compound structure while inactive compounds

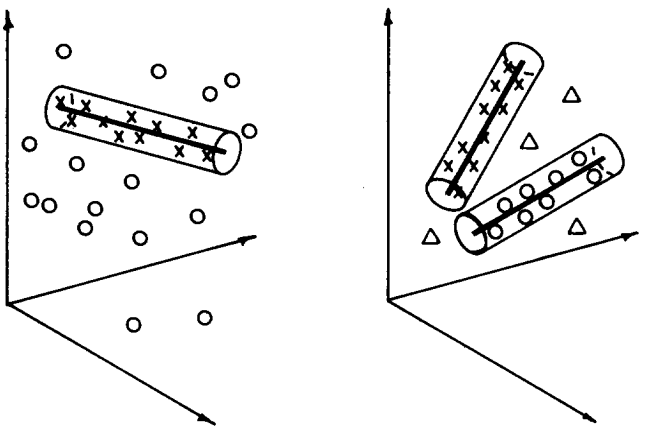


FIG. 4. Possible data structure: (a) asymmetric structure; (b) symmetric structure.

can have almost any structure. This corresponds to the "active" class being well defined as a small cluster in M-space while the "inactive" class is spread out more or less randomly in M-space; the inactive class is mathematically ill defined.

In the study of the carcinogenicity of 4-nitroquinoline-1-oxides we have an example of this asymmetric data structure. When active (carcinogenic) and inactive (noncarcinogenic) compounds are described by 43 variables of the analogy type discussed above, the "active" class was well described by a PC model with $A = 4$, while the inactive class had no appreciable structure.

Using the PC model of the single modeled class, classification can still be obtained on the basis of the position of object points inside or outside the class confidence interval (see Fig. 4). The application of hyperplane methods to asymmetric data will, however, fail. As realized from Fig. 4, it is impossible to construct a good discriminant plane unless both classes have structure.

Symmetric data structures when all classes are well defined are common when the biological testing results are nonbinary. Results are reported, for example, as agonist vs antagonist (see example below) or enzyme substrate vs enzyme inhibitor. Because all classes are described by PC models, symmetric data structures yield more information than asymmetric.

LEVELS OF CLASSIFICATION

The representation of the classification problem in graphical terms (Fig. 2) leads at once to the realization of levels of classification. We formulate these as Levels I–IV (*9*) which depend on the scope of the problem and the information which can be extracted from the data by the classification method.

Level I

Level I classification constitutes a mere classification into one of a number of classes. In the case of symmetric data structure, the objective of the application of

a PaRC method would be to classify an unknown as being a member of one of the defined classes. Most methods of PaRC, such as linear discriminant analysis and the linear learning machine, are designed to operate at this level.

Level II

Classification at this level is the same as at Level I with the additional possibility that an object might be a member of none of the well-defined classes. Since SIMCA operates by describing a class in terms of similarity and enclosing it in a mathematical structure, it always works at least at this level. The asymmetric case (Fig. 4a) can be treated only in this level.

Level III

At Level III, classification into a well-defined class, as at Level II, results. In addition the position of the objects in the classes are related to their level of activity. This constitutes classification and the derivation of a structure–activity relationship for the class. It is based on the assumption, that *within* each class, objects of *similar activity* will cluster. In some cases classification at Levels I and II may be trivial to one trained in structure–biological activity relationships of specific compound types. Level III classification is certainly nontrivial.

Level IV

At this level of classification, in addition to the descriptor matrix for the classes, a *matrix* of effect variables exists for some classes. This matrix can be comprised of biological measurements made on each of the members of the classes. In the case of carcinogens and noncarcinogens, for example, a matrix of their response in the various prescreens may be available. With such data, classification at Level IV results in relating the position of the members of a class in its descriptor matrix to its position in its effect matrix. Thus data analysis consist of two phases: (1) description of the systematic class structure in both matrices and (2) correlation of the position of each object in one matrix with its position in the other matrix. Mathematically the analysis at Level IV represents deriving path models in latent variables (*10*). As yet no examples of classification at this level have been reported.

SIZE LIMITATIONS OF THE DATA SET

In many data analytic methods, the number of variables is limited. This is the case with multiple regression (MR) as well as PaRC methods maximizing the separation between the classes, i.e., LDA and LLM. If totally N objects have been included in the training set, approximately $N/3$ is the largest number of variables allowed (*13*). In methods which are not based on class separation directly, this limitation is not necessary. With SIMCA the only limitation is that the final number of product terms in Eq. [1] (A) is much smaller than the number of objects in the modeled class, say $A \leq n_q/3$. Hence, SIMCA has no

limitations on the number of variables entering the analysis. In fact, the stability of the SIMCA classification increases monotonically with the number of relevant variables.

THE SIMCA METHODOLOGY

The details of SIMCA PaRC analysis have recently been reviewed (*5*). We therefore give only a short summary of the steps in the analysis.

1. Define the classes pertinent to the problem and select a representative training set for each class. Then describe each compound (object) as a M-dimensional data vector (see above).
2. Normalize the data to zero mean and unit variance for each variable over the whole data set. This gives each variable equal initial weight in the analysis.
3. Fit a separate PC model to each class q. The dimensionality of the models (A_q) is determined by cross-validation. In the asymmetric case, the cross-validation will indicate that some classes lack structure, i.e., $A_q = 0$ for some q.
4. Delete irrelevant variables, i.e., such variables not participating in the class models (low modeling power) and not differentiating between the classes (low discriminating power). Also, delete obvious outliers among objects in the training set.
5. Fit new PC models to the possibly reduced class matrices.
6. Calculate the typical residual SD for each class. These form the basis for the confidence intervals around the class together with the distribution of the u values in each class.
7. Classify the objects in the training set, i.e., fit all class models to all training set data vectors.
8. Validate the classification of the training set (step 7) by deleting parts of the training set and calculating new PC class models from the reduced class matrices. Fit the resulting class models to the deleted object vectors to obtain their class assignment. Rotate the deletion pattern until each object in the training set has been deleted once and only once. The classification rate calculated from assignment of the deleted objects gives a conservative estimate of the "correct" rate.
9. Classify the objects in the test set by fitting each class model from step 5 to their data vectors.
10. In case of a Level III or IV analysis, search for relations between the parameters u_{ak} (the position of object k in the class) and the measured level of activity of the objects as shown below in the examples. This analysis can be done graphically or using multiple regression (linear or nonlinear).

EXAMPLES

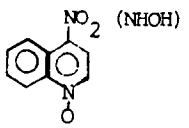
Prediction of Carcinogenic Potential

The prediction of whether an unknown or untested substance will be carcinogenic or not is an important problem to which the application of PaRC methods,

theoretically, could have fruitful results. A number of classes of organic compounds have been shown to have the potential to induce tumors in experimental animals. Among these are the 4-nitroquinoline-1-oxides (**II**) and the polycyclic aromatic hydrocarbons. There is sufficient, reliable testing data on both types of compounds to justify their analysis by PaRC and applications of the SIMCA method to such data have recently been reported (*8b,c*).

The pharmacological response data for the 4-nitroquinoline-1-oxides were taken from the literature and this resulted in 33 analogs of which training sets of the inactive and active compounds were constructed. The substances, all mono- and disubstituted, were described by substituent constants and a total of 43 such variables were used. Thirty-five were eventually required for classification and eight were found irrelevant.

After the deletion of the 3-substituted compounds, which showed ambiguous behavior in the analysis, a similarity model was derived from the training set of active compounds. A model could not be derived for the inactive compounds, indicating an asymmetric data structure for this problem.



II

Classification, on the basis of the derived similarity model, led to 83% correct classification with two false negatives and three false positives. At Level III classification the ability of the 6-substituted compounds to initiate unscheduled DNA repair (*14*), was related to class position as shown in Fig. 5, and from this figure correlation between the effect variables is apparent. In retrospect, an explanation for the anomoulous behavior of the 3-substituted compounds is probably the participation of the 3-position in the formation of adducts with DNA (*14*). The ability of this PaRC Method to detect such outliers is significant.

Probably receiving more theoretical interest as carcinogens than any other compounds are the polycyclic aromatic compounds (*15*) and most attempts to treat structure activity data for these substances have relied on correlating single

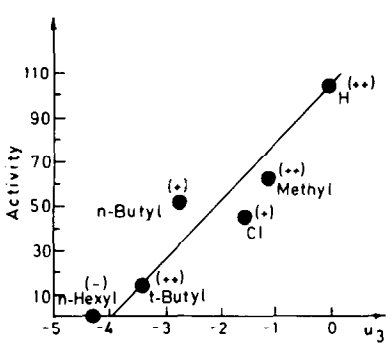


FIG. 5. Level III classification of 6-substituted analogs of (**II**). (++) Highly active carcinogens; (+) weakly active; (−) inactive.

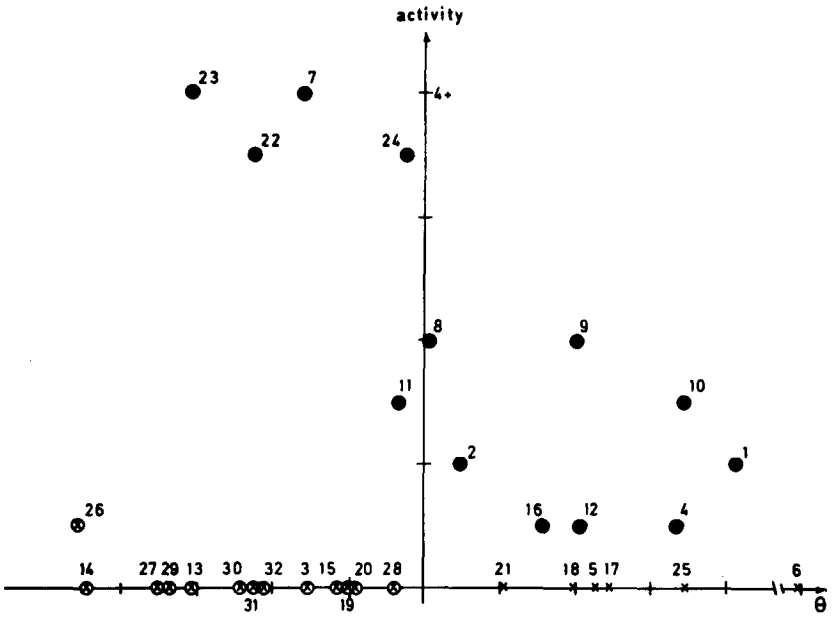


FIG. 6. Level III classification of polycyclic aromatic hydrocarbons. ●, Active; ⊗, inactive but initially predicted to be active; × correctly predicted to be inactive.

theoretical or measured variables with carcinogenicity. Unlike correlation techniques, PaRC is designed to handle a large number of variables. This, for one, is a reason that PaRC methods are better suited for such analysis.

Using 15 theoretical (mainly derived from MO calculations) and 8 measured variables for 13 carcinogenic and 19 noncarcinogenic polycyclic aromatic hydrocarbons, interesting results occurred on application of SIMCA to these data. In the initial phase of the analysis of these data, a large number of compounds appeared as false positives. By placing these compounds in a separate class a correct classification resulted. This indicates that the false positives in some way are deactivated and this deactivation mechanism is related to variables included in the descriptors for the class. In addition a Level III classification was significant with relative activity being related to class position (Fig. 6).

These two examples illustrate the point regarding chemical and pharmacological similarity. Both data sets can be assumed to be pharmacologically similar. The quinolines are structurally similar in that each analog is derived from the parent by replacement of hydrogen with a substituent. These substances can, therefore, be described by substituent constants. The hydrocarbons, on the other hand, are structurally diverse but can be described as being similar by choice of appropriate molecular descriptors.

CLASSIFICATION OF β-ADRENERGIC AGONISTS AND ANTAGONISTS

A number of compounds of general structure have β-adrenergic activity as

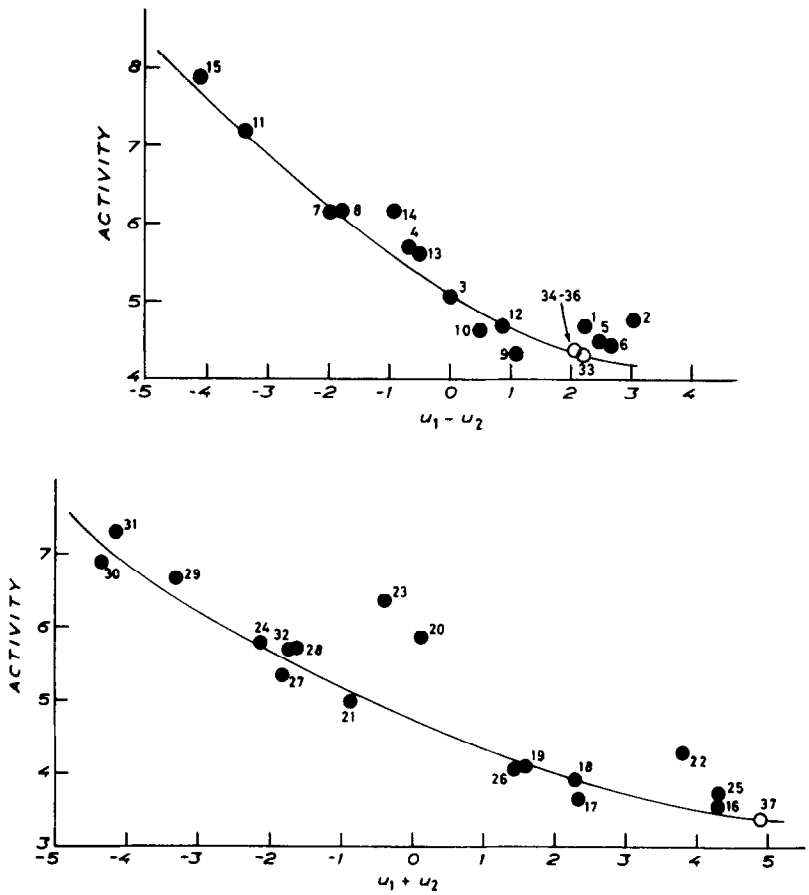


FIG. 7. Level III classification of (a) agonist analogs of (I) and (b) antagonist analogs of (I).

either agonists or antagonists (*16*) and their activities have been studied in detail at the receptor level (*16*). Very subtle structural changes can reverse receptor activity within the series and the ability to predict such behavior has obvious advantages. In an attempt to predict such behavior SIMCA was applied to data for 37 compounds of which 17 were antagonists, 15 were agonists, and 5 were weakly active or inactive as neither. Training sets were constructed using 13 variables for the description of the objects. Both classes were described by three-component models in terms of nine variables. Since the data structure is symmetric, the variables were deleted on the basis of low modeling and discrimination power. Classification of the training sets resulted in 94% correct classification.

Classification at Level III, with level of activity as agonist or antagonist as the effect variable, was of interest and the results are shown in Fig. 7. The test set compounds which were not well classified were correctly predicted to have low activity.

This last example involves a symmetric data structure in which classification information can be obtained about both classes, and in this example, classification

into one class or another is not sufficient to establish the identity of a test object. Only at Level III can sufficient information be obtained.

CLASSIFICATION OF *o*-TOLUENESULFONYL (THIO)UREAS AS HYPER- OR HYPOGLYCEMIC

The data in the examples above were not possible to analyze with traditional methods, i.e., LDA or LLM for the classification analysis and multiple regression (MR) for the quantitative relation between the descriptor variables and the measured level of biological activity. The reason for the inapplicability of these methods is that the number of variables (M) is too large in relation to the number of cases in the training set (N).

To allow a direct comparison between SIMCA and LDA, we have analyzed one additional data set from the literature. This set is such that LDA (but not MR) is appropriate and SIMCA therefore has no a priori advantage. The example concerns the effect of 22 *o*-toluenesulfonyl ureas and thioureas on the blood sugar level in rat, recently published by Franke (*17*).

The variation in structure between the 22 compounds was described using the parameters π, MR, and σ_p for the substituent on the aromatic ring and π and MR for the substituent on the urea moeity (see Ref. (*17*) for explanation). An indicator variable (−1, 0, 1) was used to describe the electron-releasing properties of the aromatic substituent and another indicator variable (0, 1) to differentiate between ureas and thioureas. Finally a variable R_M was used to describe the lipophilicity of the whole molecule. Thus, totally eight variables were used by Franke to describe the 22 compounds and the same data were analyzed by us. According to the measurements of Franke, compounds 1–9 had a hyperglycemic effect (class 1), Nos. 10–17 had no effect (class 2) and Nos. 18–22 a hypoglycemic effect (class 3). The compound numbering corresponds to the degree of hyperglycemic activity, No. 1 being the most hyperglycemic and No. 22 the least (the most hypoglycemic).

The LDA analysis of Franke was made in two ways. First a total analysis was made attempting a simultaneous separation of the three classes. This leads to the misclassification of No. 6 to class 3, No. 8 to class 2 and No. 14 to class 1. Second, disjoint analyses were made separating class 1 from class 2 and class 2 from class 3, respectively. This leads to the misclassification of Nos. 7 and 9 to class 2, No. 14 to class 1, and No. 21 to class 2.

An eigenvector projection of the data set (Fig. 8) indicates that the three classes are not well separated but that there is a continuous change from hyperglycemic *via* nonactive to hypoglycemic compounds. The most hyperglycemic (Nos. 1 and 2) are clearly at one extreme of the data structure and the most hypoglycemic (Nos. 20, 21, and 22) at the other extreme.

The SIMCA analysis describes classes 1 and 2 by three-component models and class 3 by a two-component model. These models classify all compounds in class 1 correctly, only three of eight in class 2 correctly, and three of five in class 3 correctly. However, SIMCA also finds a direction in classes 1 and 3 which are

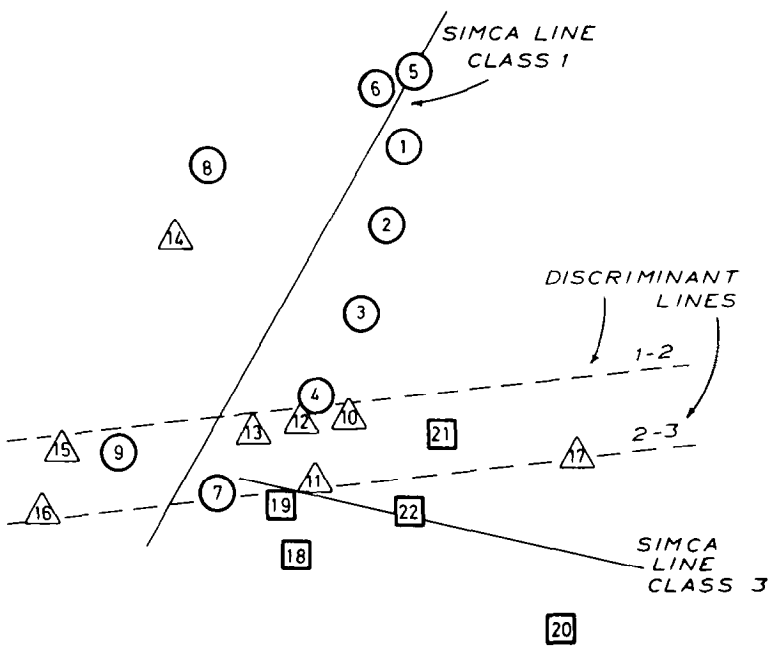


FIG. 8. Eigenvector projection of the 8-dimensional space containing the 22 o-toluenesulfonyl(thio)ureas. The projection is slightly distorted to enable a linear separation of the classes by the discriminant lines corresponding to the separation in the original 8-space. The two solid lines correspond to the major directions of the SIMCA models for classes 1 and 3 (hyper- and hypoglycemic) and the two dashed lines to the discriminant planes separating class 1 from class 2 and class 2 from class 3, respectively. Compounds of class 1, circles; class 2, triangles; and class 3, squares.

related to the degree of hyper- and hypoglycemic activity, respectively (see Fig. 8). The compounds in class 2 which were misclassified were described as close to both class 1 and class 2 and, if belonging to class 1, predicted to have low activity as indicated by their position in Fig. 8. Analogously, the two compounds of class 3 which were misclassified (Nos. 18 and 19 to class 1) were predicted to have a very low hyperglycemic activity.

Thus SIMCA gives results in accordance with a continuous change between the classes which, in our view, is a more fair view of the data set. LDA and SIMCA also give different variable combinations as those important for the biological activity. In LDA the only important variable discriminating between classes 1 and 2 is MR (molecular refractivity) of the urea substituent. Franke interprets this as small size and low lipophilicity of this substituent leading to hyperglycemic activity. SIMCA indicates hyperglycemic activity to be correlated with a high R_M (total lipohilicity) in combination with low values of MR and π of the urea substituent. For hypoglycemic activity the LDA indicates the most important variables being MR_2, I_1, and R_M indicating that a large substituent at the urea moeity in combination with a thioamide group and a high lipohilicity is important for hypoglycemic activity. SIMCA finds the direction indicated in Fig. 8 (the class 3 line) to be strongly correlated to σ_p and the electron-releasing property of the

aromatic substituent in combination with high lipophilicity. This corresponds to the interpretation that hypoglycemic activity is related to (i) the aromatic substituent *not* having electron donating properties and (ii) the presence of a thioamide moeity.

In conclusion, SIMCA has two main advantages over LDA in the present example. It finds directions in both classes 1 and 3 which are related to the level of activity of the compounds in the class. Second, SIMCA indicates the presence of a fuzzy region containing compounds of low hyper- and hypoglycemic activity together with inactive compounds. In this region classification is difficult or impossible according to SIMCA while LDA classifies compounds in this region with the same apparent certainty as it classifies active compounds.

DISCUSSION

Quantitative relationships between chemical structure and biological activity are, in the empirical way discussed here, applicable only within series of pharmacologically and chemically similar compounds. This limitation is set both by our lack of knowledge of how to describe structurally diverse molecules and by the philosophical foundations of empirical mathematical models as linearizations of complex relations valid only in limited intervals.

When the series of compounds under study involves the modification at several substituent positions and when each such position can influence the biological activity in several possible ways, the quantification becomes a multivariate one. When, in addition, the biological activity divides the compounds into several classes—often active and inactive—the problem, in addition, involves problems of classification which necessitate the use of pattern recognition methodology.

Correctly used, PaRC methods extract most of the information inherent in multivariate data. When less than perfect results are obtained, which is typical for the state of the art in the area of structure–activity relationships, this indicates that our knowledge of how to describe chemical structures, i.e., how to select the data, is less than perfect. This lack of knowledge is particularly obvious with respect to the following points:

1. Substituent descriptors. The existing substituent parameters are all derived from the substituents influence of model reactions involving simple organic molecules. With this in mind, their applicability to biological activity studies is remarkable. A systematic development of *biological* model systems and the systematic study of the influence of substituted organic molecules on these biological systems would provide the possibility of defining substituent scales directly relevant to the study of biological activity. This would, in our view, be the most obvious and promising way to dramatically increase the precision of structure–activity relationships.

2. Substituents as isolated entities. The description of molecules in terms of a structural backbone plus substituents requires that the latter modify the biological activity only moderately and that the substituents interact only weakly. This follows from the theoretical derivation of extrathermodynamic relationships and

their PaRC generalization (*18*). In many cases these assumptions might be dubious in structure–activity applications which make the simple models break down.

3. Difficulties to identify the biologically active form. Often, compounds are metabolized after their entry into the biological system and often it is the metabolites that exert the biological activity. Searching for relationships between the compounds preceeding the active form and the measured activity, at best, can give a weak correlation in such case. Analogously, often the biological activity is produced by a particular conformer or complex of the molecule about which little knowledge exists. The description of a set of molecules whose active conformer is unknown in terms of "backbone" and "substituents" is almost impossible at this time. This is probably the reason for the limited success of structure–activity relationships for nonrigid molecules.

4. Inhomogeneities in some classes. When a class contains strong subgroups, the assumption of similarity between the compounds in the class is not well fulfilled and most PaRC methods become very inefficient. Such subgrouping may be caused by letting a class contain compounds of different chemical types and/or compounds being biologically active according to several different mechanisms. Usually the subgrouping is difficult to detect in the data analysis and the poor results of the analysis are difficult to rationalize.

In addition, there are of course problems beyond the direct data analysis which often are even more serious. These include difficulties to obtain appropriate measures of the biological activity which have the same meaning over the whole data set, the interpretation of negative biological test results as true negative or weak positive under the detection level, lack of knowledge about whether the studied compounds or their metabolites constitute the active form, and other potentially difficult problems.

In contrast to separation methods such as linear discriminant analysis and the linear learning machine (LDA and LLM) (*6*), SIMCA can handle the asymmetric situation which seems to be rather common in structure–activity studies. The "class" of inactive compounds is often heterogeneous, i.e., the inactivity is caused by a multitude of factors. This leads to the situation that the inactive class can be neither modeled, nor separated from the active class(es).

Finally, we would like to comment on the choice of PaRC method and the relation between SIMCA and the more conventional multiple regression (MR) techniques. It has recently been argued (*19*) that PaRC cannot be used as a direct path to the construction of compounds with a specific activity. From the present treatment it should be clear that this statement is valid only for the separating methods such as LLM and LDA and other "Level I" methods which do not enclose each class in closed mathematical structures. With SIMCA it is possible not only to get the structural "profile" for active compounds, but also to get the profile for compounds of high predicted activity if the analysis is made on Level III as done in the present examples.

Hence, SIMCA can combine classification with the prediction of the level of activity for compounds in the active class(es). This is accomplished by performing the analysis in two phases. In the first phase, information about the similarity

(chemical and pharmacological) between compounds in the same class is used to model the correlations between the structural descriptor variables within the class. This first phase also gives u values for each compound in the class defined as linear combinations of the original variables with the coefficients b.

In the second phase, the measured biological effect of each compound in the class (y) is related to the u values by means of a linear model. SIMCA and multiple regression (MR) models are in this phase closely related. The difference is that MR is, philosophically, based on the assumption that all included "independent" variables, i.e., structural descriptors, also are statistically independent and in addition that they are relevant to the problem. MR then combines these variables as to maximize their joint correlation with the measured activity. This maximization severely limits the number of variables that possibly can be included in the analysis. With N compounds in a class, MR becomes highly unstable if the number of variables exceeds ca. $N/4$. In the applications described above, this limitation makes MR useless. SIMCA can be seen as a pretreatment of each class data matrix to extract a few statistically independent and relevant variables, the u values, which then in a Level III analysis are used as independent variables in an ordinary MR analysis. Since very few u vectors are extracted, the stability of the MR phase of the data analysis is assured.

In cases when M is much smaller than N, however, MR is appropriate provided that the data set is homogeneous. Since most MR applications reported in the literature have this condition ($M < N$) fulfilled, a comparison between SIMCA and MR is difficult. Rather, one should see SIMCA and MR as complimentary methods, the former being one possibility when MR is not applicable due to a large number of variables and/or inhomogeneities in the data set warranting a classification analysis.

REFERENCES

1. C. HANSCH, *Accounts Chem. Res.* **2,** 232 (1969).
2. H. N. BRØNSTED, *Chem. Rev.* **5,** 322 (1928).
3. L. P. HAMMETT, "Physical Organic Chemistry." McGraw-Hill, New York, 1940.
4. J. E. LEFFLER AND E. GRUNWALD, "Rates and Equilibria of Organic Reactions," Chap. 6. Wiley, New York, 1963.
5. S. WOLD AND M. SJOSTROM, "Chemometrics, Theory and Application" (ACS Symposium Series No. 52), (B. R. Kowalski, Ed.). American Chemical Society, Washington, D.C., 1977.
6. (a) K. H. TING, R. C. T. LEE, G. W. A. MILNE, H. SHAPIRO, AND A. M. GAURINO, *Science* **180,** 417 (1973). (b) B. R. KOWALSKI AND C. F. BENDER, *J. Amer. Chem. Soc.* **96,** 916 (1974). (c) A. J. STUPER AND P. C. JURS, *J. Amer. Chem. Soc.* **97,** 182 (1975). (d) K. C. CHU, R. J. FELDMAN, M. B. SHAPIRO, G. F. HAZARD AND R. I. GERAN, *J. Med. Chem.* **18,** 539 (1975). (e) L. J. SOLTZBERG AND C. L. WILKINS, *J. Amer. Chem. Soc.* **99,** 439 (1977). (f) A. CAMMARATA AND G. K. MENON, *J. Med. Chem.* **19,** 739 (1976).
7. S. WOLD, *Pattern Recognition* **8,** 127 (1976).
8. (a) W. J. DUNN III, S. WOLD, AND Y. C. MARTIN, *J. Med. Chem.* **21,** 922 (1978). (b) W. J. DUNN III, S. WOLD, *J. Med. Chem.* **21,** 1001 (1978). (c) B. NORDEN, U. EDLUND, AND S. WOLD, *Acta Chem. Scand. B* **32,** 602 (1979).
9. C. ALBANO, W. J. DUNN, III, U. EDLUND, E. JOHANSSON, B. NORDEN, M. SJOSTROM, AND S. WOLD, *Anal. Chim. Acta* **103,** 429 (1978).

10. H. WOLD, "Essays in Honor of Oskar Morgenstern" (R. Henn and O. Moeschlin, Eds.) Springer-Verlag, Berlin, 1977.
11. (a) C. HANSCH, P. O. MALONEY, T. FUJITA, AND R. M. MUIR *Nature (London)* **194,** 180 (1962). (b) R. W. TAFT, "Steric Effects in Organic Chemistry" (M. S. Newman, Ed). Wiley, New York, 1956. (c) R. F. REKKER, "The Hydrophobic Fragment Constant," Elsevier, Amsterdam, 1977. (d) A. VERLOOP, W. HOOGENSTRAATEN, AND J. TIPKER, "Drug Design" (E. J. Ariens, Ed.), Vol. 7. Academic Press, New York, 1976.
12. H. WOLD, "Festschrift for J. Neyman" (F. N. David, Ed.), p. 411. Wiley, New York, 1966.
13. C. P. WEISEL AND J. L. FASCHING, Anal. Chem., **49,** 2114 (1977).
14. H. F. STITCH, R. C. H. SAN, AND Y. KOWAZOE, *Nature (London)* **229,** 416 (1971).
15. P. Pott, *Nat. Cancer Inst.* **10,** 7 (1963). (Monograph)
16. C. MUKERGEE, M. C. CARON, D. MULLIKEN, AND R. J. LEFKOWITZ, *Mol. Pharmacol.* **12,** 16 (1976).
17. S. DOVE, R. FRANKE, O. L. MNDSHOJAN, W. A. SCHKULJEV, AND L. W. CHASHAKJAN, *J. Med. Chem.* **22,** 90 (1979).
18. S. WOLD AND M. SJOSTROM, "Correlation Analysis in Chemistry" (N. B. Chapman and J. Shorter, Eds.). Plenum, London, 1978.
19. A. J. STUPER AND P. C. JURS, *J. Pharm. Sci.* **67,** 745 (1978).