Journal of Medicinal Chemistry

Volume 19, Number 6 June 1976

Pattern Recognition. Classification of Therapeutic Agents According to Pharmacophores

Arthur Cammarata* and Govind K. Menon

Temple University, School of Pharmacy, Philadelphia, Pennsylvania 19140. *Received September 29, 1975*

One of the preprocessing methods used in pattern recognition—factor analysis—is shown to be well suited to the derivation of structure–activity relationships. Applications of the procedure developed are illustrated using sets of compounds which are of accepted therapeutic utility.

Within the past few years a branch of artificial intelligence known as pattern recognition has been introduced into the literature of chemistry.[1,2] The techniques associated with pattern recognition[3-6] are finding interest in the design, assay, and development of biologically active substances as they offer the promise of rapidly identifying, from stores of accumulated information, substances which seem worthy of more detailed study. That is, the results obtained by the application of pattern recognition methods may be envisaged as providing a basis for establishing lead substances, for recognizing structural uniqueness in relation to a particular biological effect, for the setting of priorities in conducting biological assays, and for the identification of pharmacophoric patterns of molecular substitution.

At present there are very few published articles which attempt to represent structures in a coded manner and which are directed toward deriving structure–activity relationships using pattern recognition methods.[7-10] Two of these are effectively tests of whether the techniques enable classifications of pharmacological activities, e.g., sedatives and tranquilizers,[8,10] while another two attempt to use the methods in efforts to predict the anticancer activity of various agents.[7,9] The results have been encouraging and suggest that, with suitable interaction between user and computer,[11] many of the envisaged potentials of pattern recognition in structure–activity studies will be realized.

Criticism[12-14] has been directed at some studies[7,15] which have made use of pattern recognition methods in arriving at structure–activity relationships. In particular, the choice of compounds included in the data sets and the molecular structure representations were viewed as inappropriate. These criticisms are overcome by the approach which is to be described.

A key to the successful application of pattern recognition in deriving structure–activity relationships lies with the coding of molecular structural features. Biological effects are readily coded and compounds giving rise to particular actions can easily be sorted out.[16,17] However, the identification of a particular pharmacophoric pattern of substitution in relation to a specified biological effect is a more difficult problem to do by computer and requires suitable molecular-descriptor codings. Among the molecular descriptors which have been used are augmented atom,[7-10] heteropath,[9,10] and substructural[10,18] representations. These choices may have been dictated because chemical information is often catalogued in computer data stores using such codings,[19] or they may have been suggested by an early effort to relate the mass spectral fragmentation patterns to the pharmacological actions of drugs.[15] A variety of disadvantages attend the use of such codings, however, especially when seeking to establish structure–activity relationships: (a) a large number of descriptors usually must be used to represent a given chemical compound; (b) information concerning the relative arrangement of groups or substituents within molecules frequently may be lost; (c) redundant codings may have to be employed to properly discriminate between compounds; and (d) it is extremely difficult to interpret the results of a pattern recognition solution in terms of structural prototypes when these codings are used.

One technique encompassed by the term pattern recognition—factor analysis[20,21]—is particularly well suited for arriving at readily interpretable structure–activity relationships. Factor analysis has been used previously in structure–activity studies[22] but not as a preprocessing method intended to separate compounds into classes based on their molecular-descriptor codings. As applied in this report, the method necessitates that one view molecular structures in a manner consistent with the way medicinal chemists have done when proposing candidate compounds. The codings adopted are intended to discriminate between atoms and groups, in particular atoms and groups which are bioisosteric.[23,24] It is shown that only a small number of molecular features, corresponding to pharmacophoric patterns of substitution, need be considered to properly resolve sets of compounds into classes and subsequently to identify these classes as therapeutically distinct.

Basis of Approach. Pattern recognition may be generalized as consisting operationally of two distinct steps: (a) preprocessing, where a data matrix is analyzed by a procedure which could enable a reduction in the number of variables or the elimination of redundancies arising

because of interrelationships between the variables; and (b) classification, where the results of the preprocessing step are transformed, or otherwise operated upon, so as to discriminate between the classes that may have been revealed by the preprocessing. The preprocessing technique used in this article is principal component analysis, which is but one of many methods in factor analysis that enables a reduction in the dimensionality, i.e., in the number of variables, of a data matrix.[25,26] This preprocessing method has been used successfully in chemical applications of pattern recognition,[1] but its relationship to factor analysis was not stated explicitly. Classification is done graphically in this article, rather than by a mathematical technique such as cluster analysis,[27] discriminate analysis,[28] or "learning machines",[5,10] to provide interested persons with a conceptual foundation and to point out the parallelism between the classical and the pattern recognition approaches for arriving at structure–activity relationships. A detailed summary of factor analysis and pattern recognition is beyond the scope of this work. Here it is intended only to present a basis for more detailed study. The necessary computer programs are generally available[26,29] and include procedures which are much more sophisticated than the one which is to be presented.

A set of M molecules whose structures are coded so as to represent N features gives rise to an $M \times N$ data matrix. This data matrix may be used to construct an $N \times N$ interfeature or an $M \times M$ "intercompound" correlation matrix. When factor analysis is applied as a preprocessing method to the $N \times N$ interfeature correlation matrix, so as to attempt to isolate discriminating molecular features, the method is known as R factor analysis. On the other hand, if factor analysis is applied to the $M \times M$ "intercompound" correlation matrix so as to attempt to isolate similar compounds, the method is known as Q factor analysis. R factor analysis is usually the more important in structure–activity studies, since one is usually most interested in identifying distinctive molecular features leading to a particular biological response. R factor analysis also offers the advantage of having to deal usually with a much smaller correlation matrix than is required when making a Q factor analysis. Operationally these two preprocessing methods are identical, the designation R or Q factor analysis simply serving to specify the manner by which a correlation matrix is to be viewed.

A correlation matrix may be operated upon in a variety of ways so as to extract factors, i.e., reduce the dimensionality. The two most generally employed techniques are principal-component analysis and general-factor analysis. These differ in the assumptions used in their derivation. With principal-component analysis no particular assumption is made about the possible structuring within the data. One simply seeks the "best" linear combinations of the variables in a data matrix to account for the variance within the data. All of the variance within a data matrix can be accounted for by an appropriate *set* of linear combinations of the variables. Many times, however, a relatively few linear combinations (in comparison with the number of variables in a data matrix) can account for a high proportion of the variance in the data. These may then be used as representations of the data and, since the linear combinations are orthogonal to one another, a reduction of the dimensionality of the data matrix will have been achieved. In contrast, general-factor analysis makes the assumption that the variables in a data matrix may be influenced to varying degrees by properties that may or may not be held in common by the respective variables. In other words, each variable used to represent a portion of a molecular structure is viewed as reflecting a common set of physical attributes for some compounds in the data set and a differing set of physical attributes for others. What is sought are those variables which seem to reflect common physical attributes for the majority of compounds in the data set, all other compounds being designated as unique. This technique offers the promise of providing a degree of mechanistic insight into factors leading to alternative pharmacological actions. Some progress in adapting the methodology to physicochemical pursuits[30] and in investigating physiological processes such as olfaction[31] has been made, but in the absence of firm guidelines for the application of general-factor analysis to structure–activity studies it is more prudent to make use of the less physicochemically satisfying technique of principal-component analysis.

The basis for a principal-component analysis may be placed in perspective by considering an attempt to transform a multiple regression problem into a simple linear one. In multiple regression it is presumed that the biological activities A for M compounds can be estimated from a linear combination of N variables X.

$$A_i = b_1X_{1i} + b_2X_{2i} + \ldots + b_NX_{Ni} + \mu \qquad (1)$$
$$i = 1, 2, \ldots, M$$

However, in evaluating the least-squares estimates of the coefficients b, the values of the coefficients necessarily are a function of the covariance between the independent variables X as well as of the covariance between the observed biological activities and the independent variables. One may thus seek to account for the covariance between the independent variables X prior to attempting a regression analysis.

In making this attempt consider defining a transformed variable T which is a linear combination of the independent variables X.

$$T_i = m_1X_{1i} + m_2X_{2i} + \ldots + m_NX_{Ni} \qquad (2)$$
$$i = 1, 2, \ldots, M$$

By analogy with the least-squares procedure used in multiple regression, one can write a set of normal equations which would enable an estimate of the coefficients m if the respective values of T were known. Equation 3 represents

$$\sum_{i=1}^{M}(T_i - \bar{T})(X_{ki} - \bar{X}_k) = \sum_{i=1}^{M}\sum_{j=1}^{N} m_j(X_{ji} - X_j)(X_{ki} - \bar{X}_k) \qquad (3)$$
$$k = 1, 2, \ldots, N$$

a set of N simultaneous equations in N unknowns. A bar over a variable designates that the average for the variable is to be used.

Since the "best" estimates of T are not known, a procedure must be devised to gain such estimates. This procedure has as its basis the observation, in matrix notation, that premultiplication of eq 3 by the transpose of the column vector for the coefficients $\mathbf{m}^T$ leads to the relation

$$\mathbf{T}^T\mathbf{T} = \mathbf{m}^T\mathbf{C}\mathbf{m} \qquad (4)$$

Now since $\mathbf{T}^T\mathbf{T} = \Sigma(T_i - \bar{T})^2$ the "best" estimates of T are those where the sum of squares of T taken with respect to each coefficient m is a minimum. This solution may be made under the restriction that the coefficients m are

orthonormal, in which case eq 4 becomes

$$\mathbf{T}^T\mathbf{T} = \mathbf{m}^T\mathbf{C}\mathbf{m} - \lambda(\mathbf{m}^T\mathbf{m} - 1) \quad (5)$$

where λ is a Lagrangian multiplier. With eq 5 as a basis, the minimization leads to the set of simultaneous equations

$$\sum_{j=1}^{N}\sum_{i=1}^{M}[m_j(X_{ji} - \overline{X}_j)(X_{ki} - \overline{X}_k) - \delta_{jk}\lambda] = 0 \quad (6)$$

$$k = 1, 2, \ldots, N$$

where δ_{rs}, the Kronecker delta function, has the value 1 when $r = s$ and the value 0 when $r \neq s$. A nontrivial solution to such a set of simultaneous equations, i.e., a solution where the coefficients m are not all 0, is obtained by evaluating the roots λ of the characteristic determinant

$$\sum_{j=1}^{N}\sum_{i=1}^{M}[(X_{ji} - \overline{X}_j)(X_{ki} - \overline{X}_k) - \delta_{jk}\lambda]| = 0 \quad (7)$$

$$k = 1, 2, \ldots, N$$

and substituting these, in turn, into the simultaneous equations represented by eq 6 to arrive at values for the coefficients m.

In general, when the data matrix contains N variables the characteristic equation will have N roots and consequently there will be obtained N eigenvectors T. The reproduction of the data matrix in terms of these eigenvectors (the "explained" variance) is measured by the relation

$$\text{fraction "explained" variance} = \sum_{j=1}^{N}\lambda_j/N \quad (8)$$

All of the eigenvectors necessarily represent the data exactly, i.e., the fraction of "explained" variance is 1. Many times, however, a fewer number of eigenvectors T than of variables X can be used to represent, or to approximate, the original data matrix. Thus, one may elect to use those high-value eigenvectors which "explain" a fraction greater than 0.95 of the data, such as is done in physicochemical applications,[30] or one may choose to work with an approximated form of the data matrix, selecting only those eigenvectors whose associated eigenvalues are 1 or greater,[26] and thereby arrive at eigenvectors which "explain" a fraction on the order of 0.80 or greater of the data. The latter is the approach taken in this article since a fewer number of eigenvectors are required to represent the data in graphic form.

It may be noted that principle-component analysis presents a means by which optimization procedures involving regression methods, such as Fujita–Hansch analyses,[32,33] can be made less prone to error due to intercorrelations between independent variables. Correlation analysis which invokes principle-component analysis as a preprocessing method would thus be based on the multiple regression model

$$A_i = \sum_{t=1}^{K} c_t T_{ti} + \mu \qquad i = 1, 2, \ldots, M \quad (9)$$

wherein the K *orthogonal* variables T "explain" the major proportion of the variance in the original data.[34] An interface thus exists between pattern-recognition techniques and the other more commonly employed quantitative structure–activity methods.[34]

A technical problem is presented when seeking to solve a set of normal equations in multiple regression analysis or the characteristic equations in principal-component analysis if the elements to the equations are expressed as covariances, i.e., as sums of cross-products taken between variables. This problem arises in consequence of the relative magnitudes of the values used to represent the variables. Those variables which are numerically greater in a data matrix will lead to covariances of numerically high magnitude and hence these variables will be designated, artificially, as the more important contributors in a given linear combination. The potential for an artifact of this type in Fujita–Hansch analyses has been pointed out.[35] Autoscaling of variables[1] is a means of avoiding this problem. What this entails is the use of a correlation matrix **R** in place of a covariance matrix **C** when establishing normal equations such as is represented by eq 3. The elements to a correlation matrix r_{rs} are defined by the relation

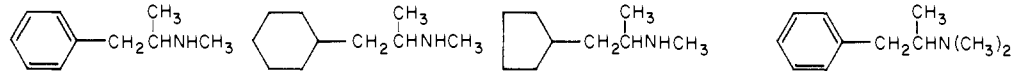
$$r_{rs} = \sum_{i=1}^{M}(X_{ri} - \overline{X}_r)(X_{si} - \overline{X}_s)/[\sum_{i=1}^{M}(X_{ri} - \overline{X}_r)^2 \cdot \sum_{i=1}^{M}(X_{si} - \overline{X}_s)^2]^{1/2} \quad (10)$$

where r and s are the variables whose intercorrelation is being evaluated. By this definition any one element in a correlation matrix **R** can vary only between 1 and 0; hence, no element to the matrix can be inordinately weighted. The coefficients to a linear combination evaluated by the use of a correlation matrix **R** will differ from those evaluated by the use of a covariance matrix **C**. In a principal-component analysis it is only necessary to premultiply an eigenvector by the square root of its eigenvalue to convert the solution from a correlation coefficient to a covariance basis.[26] All of the applications reported in this article make use of autoscaled variables in the principal-component analysis.

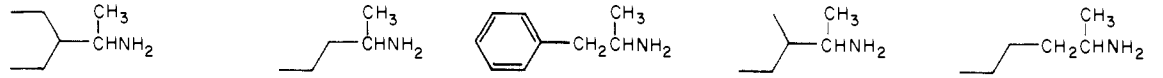
Once a set of data has been preprocessed by a factor analytic technique, the derived eigenvectors which "best" represent the data may then be used as a basis for seeking regularities in the data. The simplest procedure, and one which is not always possible to employ, is to construct a graph taking for each axis one of the derived eigenvectors T. A point on this graph is found by introducing the descriptors X for a compound into each of the linear combinations representing the eigenvectors T so as to evaluate the length of the component to take along each of the axes in plotting the point. With only two or three axes to consider an entire set of data can be displayed as a two- or three-dimensional graph. Regularities in the data may be manifested as discrete "clusters" of points. These "clusters" serve to classify the data. In structure–activity studies each point in this type of graph represents a compound, and each compound may or may not elicit a differing biological response. When a particular type of biological response can be identified with one of the "clusters" of points (compounds), a structure–activity relationship will have been recognized. If no such identification can be made it does not mean that a structure–activity relationship does not exist. Rather, what is signified is either that the descriptors used to represent the compounds bear no relationship to the biological activities under consideration or that too general a criterion of biological activity has been selected in seeking classifications; i.e., the biological response of interest may arise from a variety of alternative physiological–pharmacological processes.

As the application of pattern recognition to drug development is a new field of endeavor, work in these laboratories has been proceeding at a naive level of mathematical sophistication so as to gain a better understanding of the problems and pitfalls that may arise when applying this technique. Ultimately, it is proposed, the development

Chart I. A Set of Compounds Which Includes Weak and Strong Pressor Agents



1 (methamphetamine) 2 (propylhexedrine) 3 (cyclopentamine) 4 (*N*,*N*-dimethylamphetamine)



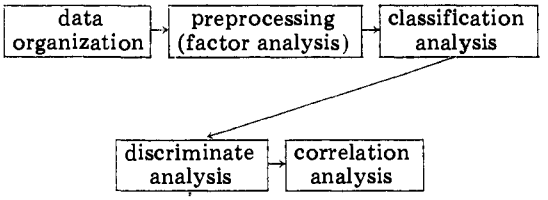
5 (2-amino-3-ethylpentane) 6 (2-aminopentane) 7 (amphetamine) 8 (methylhexanamine) 9 (2-aminohexane)

10 (β-phenethylamine) 11 (α-phenethylamine) 12 (*N*-methylbenzylamine) 13 (benzylamine)

Table I. A Descriptor Set to Distinguish Aromatic and Aliphatic Moieties

Feature	Nature	Descriptor
1-6	Aromatic atom	2
	Aliphatic atom	1
	No atom	0
7	CH_2 present	1
	CH_2 absent	0
8	$CHCH_3$	1
	CH_2	0
9, 10	CH_3 present	1
	CH_3 absent	0

of quantitative structure–activity relationships from large bodies of information will follow a flow which can be characterized under the heading "multivariate analysis". This flow is represented by the diagram

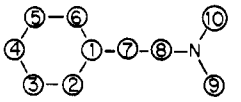


According to this diagram structure–activity data will be organized in a form suitable to apply a factor analysis technique. The results of the factor analysis are then made use of in seeking classifications that exist within the data, with the aid of techniques like graphic projection,[2] cluster analysis,[27] or "learning machines.[5,10] Once the classifications are specified a discriminate analysis procedure can be employed so as to mathematically characterize each classification. Compounds within a given classification may then be investigated by a multiple regression method so as to establish more quantitative structure–activity relationships.

Applications. Two examples will be presented to illustrate the utility of pattern recognition in establishing structure–activity relationships. The sets of compounds chosen in each example are generally of known therapeutic utility. The method is applied to these to demonstrate that the mathematical approach leads to results which are consistent with known structure–activity relationships. This work is thus serving to standardize the approach before venturing into areas where structure–activity relationships are less defined. Each example discussed poses a differing technical problem. The resolution of each problem that is presented is not necessarily the best possible or the only solution that may be proposed. A generally direction is indicated, however, which may be fruitful to follow when attempting to develop structure–activity relationships with the aid of pattern recognition.

Example 1. Consider the series of 13 compounds shown in Chart I. These may be recognized to consist principally of weak and strong pressor agents. The objective of this example is to propose a means of structural representation which is consistent with the mathematical procedure used for preprocessing.

All of the compounds in Chart I can be superimposed on the reference diagram shown below.



The numbering in this diagram is arbitrary and serves only to establish the column in which a particular descriptor value is to appear in a molecule–feature data matrix. For this particular case one may choose to express each feature in terms of the descriptor values shown in Table I. Use of these descriptors leads to the molecule–feature data matrix presented in Table II. It may be noted that in this instance the molecule–feature data matrix is constructed

Table II. Molecule–Feature Data Matrix for a Series Including Pressor Agents

	Feature									
Compd	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
1	2	2	2	2	2	2	1	1	0	1
2	1	1	1	1	1	1	1	1	0	1
3	1	1	1	0	1	1	1	1	0	1
4	2	2	2	2	2	2	1	1	1	1
5	1	1	1	0	1	1	0	1	0	0
6	1	1	1	0	0	0	0	1	0	0
7	2	2	2	2	2	2	1	1	0	0
8	1	1	1	0	0	1	0	1	0	0
9	1	1	1	0	0	0	1	1	0	0
10	2	2	2	2	2	2	1	0	0	0
11	2	2	2	2	2	2	0	1	0	1
12	2	2	2	2	2	2	0	0	0	1
13	2	2	2	2	2	2	0	0	0	0

Table III. Correlation Matrix for R Factor Analysis of Data of Table II

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
1.000	1.000	1.000	0.963	0.910	0.900	0.071	−0.507	0.267	0.097
	1.000	1.000	0.963	0.910	0.900	0.071	−0.507	0.267	0.097
		1.000	0.963	0.910	0.900	0.071	−0.507	0.267	0.097
			1.000	0.926	0.904	0.150	−0.488	0.257	0.205
				1.000	0.947	0.158	−0.461	0.243	0.281
					1.000	0.064	−0.456	0.240	0.230
						1.000	0.225	0.267	0.414
							1.000	0.158	0.057
								1.000	0.365
									1.000

Table IV. Accepted Principal-Component Solution for the Data of Table II

Factor no.	Eigenvalue, λ	Coefficients in the linear combination T									
		M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8	M_9	M_{10}
1	6.10	0.40	0.40	0.40	0.40	0.38	0.38	0.05	−0.22	0.13	0.08
2	1.78	−0.06	−0.06	−0.06	0	0.04	0	0.56	0.38	0.48	0.55

in a manner similar to the procedure used in arriving at a Free–Wilson data matrix.[36] The distinction here, however, is that the type of biological activity and the relative potencies for the compounds are not specified.

Using the molecule–feature data matrix, a 10 × 10 correlation matrix is constructed so as to determine the degree of intercorrelation between the represented features. This correlation matrix is shown in Table III. The conversion is essentially the same as would be made if one were to follow the Fujita–Ban[37] multiple regression solution of a Free–Wilson data matrix. With the correlation matrix as a basis, preprocessing is accomplished by determining the eigenvalues, λ_k, and associated eigenvectors, T_k, that characterize the correlation matrix. For the problem under consideration, a 10 × 10 correlation matrix leads to ten eigenvalues and associated sets of eigenvectors. However, not all of these eigenvectors are needed to represent the data matrix.

The relative ranking of the eigenvalues, largest to smallest, provides a measure of the information content (proportion of the variance within the molecule–feature data matrix) accounted for by the respective eigenvectors. Usually the higher eigenvalues are associated with those eigenvectors which account for the most frequently occurring sets of molecular descriptors within a molecule–feature data matrix. As a rule, all eigenvectors whose associated eigenvalue is less than 1 can be neglected,[26] since the information content (proportion of "explained" variance) provided by these is small or negligible relative to the others. For the example under discussion, there are only two eigenvalues of acceptable magnitude (≥1). These, and their associated eigenvectors, are presented in Table IV. The two solutions account for a

$$\text{fraction "explained" variance} = (6.10 + 1.78)/10 = 0.788$$

The eight solutions which have been neglected account for a fraction of "explained" variance of 0.212. Hence, the two eigenvectors given in Table IV contain 78.8% of the information content provided by the original molecule–feature data matrix. To a good first approximation, then, these two eigenvectors accurately represent the original molecule–feature data matrix.

Inspection of the absolute magnitudes of the coefficients for the eigenvectors shown in Table IV provides an indication of the meaning of each eigenvector. It will be noted that features X_1–X_6 are for a ring or a pseudoring structural fragment, while features X_7–X_{10} are for a side chain. The coefficients in the linear combination T_1 for factor 1 are all of appreciable magnitude, excepting those coefficients associated with features X_7 and X_{10} which are near 0. The linear combination T_1 may thus be viewed as characterizing ring or pseudoring features, side chains one carbon in length, and single substitutions on nitrogen (feature X_9). In contrast, the coefficients in the linear combination T_2 for factor 2 are of appreciable magnitude only when associated with features X_7–X_{10}, those associated with features X_1–X_6 being near 0. The linear combination T_2 may thus be viewed as characterizing side chains two carbons in length and double substitutions on nitrogen (features X_9 and X_{10}). The two linear combinations T_1 and T_2 thus account for all of the distinguishing molecular characteristics for the compounds in Chart I.

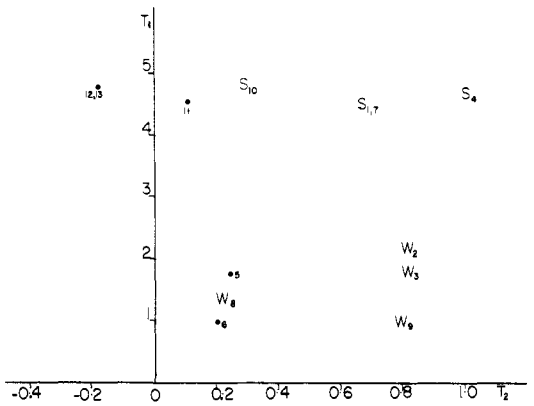


Figure 1. Factor space showing the resolution of weak (W) and strong (S) pressor agents. Points having no designation are not identified explicitly as pressor agents.

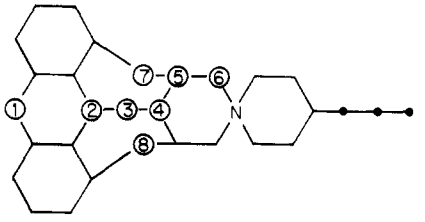
The use of a molecular superposition as a reference for the construction of a molecule–feature data matrix thus appears justified when preprocessing structural representations by principal-component analysis. The factors which are derived provide a concise summary of distinguishing molecular characteristics and their interpretation is facilitated because of the identification made between defined molecular features X and true molecular variations.

A graphical representation of the solution is provided by introducing the descriptors for each compound so as to evaluate the linear combinations T_1 and T_2. Considering these as scalar quantities, factor axes for T_1 and T_2 can be established and the points for each compound can be plotted with reference to these axes. Figure 1 depicts this graph. It will be noted that there is a clear distinction between the aromatic and the aliphatic compounds, the

aromatic compounds tending to lie higher on factor axis T_1 than the aliphatic compounds. The compounds having a two-carbon side chain separating the amino group from a ring or pseudoring moiety are also found to be higher on factor axis T_2 than the compounds having a side chain which is only one carbon in length.

By labeling the points in Figure 1 as S for a "strong" and W for a "weak" pressor substance,[38] it can readily be recognized that (a) compounds having aromatic rings and a two-carbon side chain are the more potent and (b) the length of the side chain with the aliphatic compounds does not seem as critical as with the aromatic compounds. This classification has led to a structure–activity relationship, since a separation of structurally distinct compounds, achieved using a factor analytic method, when matched against differing biological responses indicates the structure and activity distinctions to coincide.

Example 2. A more complex problem is represented by the 43 compounds shown in Chart II. These may be recognized as consisting of antihistamines, anticholinergics, analgesics, and antidepressants, antipsychotics, and anti-Parkinsonian agents patterned after chlorpromazine. These compounds are structurally similar, but they are not congeneric. In contrast to the previous example, this entire set of compounds is not superimposable on any structure which bears a resemblance to a real compound. However, a reference figure which is the simplest possible superposition of the structures can be constructed. This structure is shown below.



The interconnections between the points in this "superstructure" are arbitrary and are included simply to give a form to this geometric representation. It will be noted that any compound in Chart II can have its structure replicated by suitably interconnecting the points in the reference structure. Not all points have to be made use of when designating a compound in this way.

If the "superstructure" representing the simplest superposition of compounds in a data set is used as a basis for the construction of a molecule–feature data matrix, principal-component analysis should reveal the most frequently occurring atoms, bonds, or features *which distinguish* the compounds in the data set. Such distinctions have meaning, in a classification analysis, when they can be associated with differing pharmacological responses, e.g., antihistaminic as opposed to anticholinergic, since pharmacophoric requirements will thus have been recognized.

The coding for the structures, in this instance, bears a relationship to that used in the first example in the sense that the presence or absence of an aliphatic carbon atom is recognized by assigning a value of 1 or 0, respectively. A major distinction between the compounds of Chart II, however, is the bioisosteric group replacement that can be referenced with respect to the positions numbered 1, 2, 3, and 7 of their superposed structures. A principal problem in the coding of these structures, then, lies in the means by which bioisosteric groups might be coded. Most recently a concept termed "physical bioisosterism" has been presented,[27] which is based on the premise that structurally similar compounds having similar physicochemical properties should behave in a biologically similar manner. A variety of group parameters designating an electronic, σ, lipophilic, π, or steric, E_s, property of a group could be employed as a coding, either taken individually or in differing combinations, but at this stage in the application of pattern recognition to structure–activity studies too many descriptors (codings) can lead to problems of interpretation. Therefore, for the sake of simplicity, the definition of isosterism[39]—"Atoms, ions or molecules in which the peripheral layers of electrons can be considered to be identical"—shall here be taken to signify that molar refraction, MR, can be used as a measure of isosterism and bioisosterism. The molar refractivity for a substance is a function of the "looseness" of outer shell electrons[40], has been viewed as a measure of molecular volume,[35,41] and, it can be noted, is in general similar in value for groups which have classically been considered bioisosteric, e.g., –S–, MR = 7.92, and –CH=CH–, MR = 7.52. The latter point, which has been previously unrecognized, has been used as a guide to establish codings for the compounds of Chart II.

Table V. A Coding Incorporating Bioisosteric Atoms

Feature 1	Descriptor	Features 2–8	Descriptor
CH_2CH_2	1.5	S	1.3
S	1.3	CH_3	1.2
CH=CH	1.0	CH_2, CH, C	1.0
CH_2O	0.9	C=O	0.8
O	0.3	N	0.6
No atom/group	0	C=	0.4
		O	0.3
		No atom/group	0

Two sets of coding values were calculated using atom–bond molar refractivities. The first set, which applies to the feature labeled 1 in the reference diagram, is for groups which are substituted on an aromatic ring. These group refractivities have been arbitrarily taken relative to the group refractivity of –CH=CH–. The second set, which applies to all other features labeled in the reference diagram, is for atoms or groups which constitute side-chain modifications. These are arbitrarily taken relative to the atom refractivity for C. The values for each set are given in Table V. It must be noted that these values must be considered tentative as no effort has been made to develop a uniform or a generally applicable coding for bioisosteres. The results which are gained, however, suggest molar refractivity may be a useful index to pursue in developing codings for at least the classical bioisosteres.

The compounds of Chart II, when viewed in terms of the features labeled 1–8 of the "superstructure", lead to the pharmacophore–feature matrix shown in Table VI. Using this matrix as a basis it may be hoped that a pattern recognition solution would lead to a correct classification of the compound's pharmacological actions. Success would designate that distinguishing pharmacophores have been identified, while incorrect classifications could imply that additional features must be considered to gain a correct classification. The features dealt with in this example are a reasonable first guess of the pharmacophores which distinguish the actions of the compounds in Chart II. It should be noted that at least three feature variations occur in each column of the data matrix. This may be considered desirable because correlation coefficients derived using this matrix would be expected to have greater significance when based on three or more composite points than when based only on two. That is, intrafeature distinctions, as measured by correlation coefficients, would be expected to be rep-

Table VI. Pharmacophore-Feature Data Matrix for the Compounds of Chart II

Compd	Feature X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
1	0	1.0	0	0.3	1.0	1.0	0	0
2	0	1.0	0	0.3	1.0	1.0	1.2	0
3	0	0.6	0	1.0	1.0	0	1.0	0
4	0	0.6	0	1.0	1.0	0	1.0	0
5	1.0	0.4	0	0.4	1.0	1.0	0	0
6	0	0.4	0	0.4	1.0	1.0	0	0
7	1.3	0.6	0	1.0	1.0	1.0	0	0
8	1.5	0.6	0	1.0	1.0	1.0	0	0
9	0	1.0	0	0.3	1.0	1.0	1.2	0
10	1.5	0.4	0	0.4	1.0	1.0	0	0
11	0	1.0	0	0.6	1.0	1.0	0	0
12	1.0	1.0	0	1.0	1.0	1.0	0	0
13	1.5	0.6	0	1.0	1.0	1.0	0	0
14	1.5	0.4	0	0.4	1.0	1.0	0	0
15	0.9	0.4	0	0.4	1.0	1.0	0	0
16	0	1.0	0.3	1.0	1.0	0.8	0	0
17	0	1.0	1.0	0.6	1.0	1.0	0.6	0
18	0	1.0	0	1.0	1.0	0	0.6	0
19	0	1.0	0	1.0	1.0	1.0	0.6	0
20	0	1.0	0	1.0	1.0	0	0.6	0
21	0	1.0	0	1.0	1.0	0	0.6	0
22	0	1.0	0	1.0	1.0	0	0.6	0
23	0	0	0	0	1.0	1.0	1.0	1.0
24	0.3	1.0	0.8	0.3	1.0	1.0	0	0
25	0	1.0	0.8	0.3	1.0	1.0	0.6	0
26	0	1.0	0.8	1.3	1.0	1.0	0	0
27	0	1.0	0.8	0.3	1.0	1.0	0	0
28	1.0	0.6	0.8	0.7	0	0	0	0
29	0	1.0	0.8	1.6	1.0	1.0	0	0
30	0	0.4	0.4	1.0	1.0	0.8	0	0
31	1.3	0.6	0	1.0	0.8	0	0	0
32	1.3	0.6	0	1.0	0.8	0	0	0
33	1.3	0.6	0	1.0	0.8	1.0	0	0
34	1.3	0.6	0	1.0	1.0	1.0	0	0
35	0	0.4	0	0.4	1.0	1.0	1.0	1.0
36	0	1.0	0	0.8	1.0	0	1.0	0.3
37	0	1.0	0	1.0	0.8	0	0	0.8
38	0	1.0	0.8	0.3	1.0	0.8	0.6	0
39	1.3	1.0	0	1.0	1.0	0	0	0
40	0	1.0	0.8	0.3	1.0	0	0	0
41	0	1.0	0.8	0.3	1.0	0	0.6	0
42	0	1.0	0.8	0.3	1.0	0	0.6	0
43	0	1.0	0.8	0.3	1.0	0	0.6	0

resented more reliably. Table VII gives the correlation matrix for the data of Table VI.

Upon solving for the characteristic roots of the pharmacophore-feature correlation matrix there are obtained four solutions of acceptable magnitude ($\lambda \geq 1$). These, it will be noted, account for $(2.13 + 1.71 + 1.33 + 1.14)/8$ or 0.788 of the variance contained within the original data matrix. Table VIII presents the accepted eigenvalues and associated eigenvectors.

Inspection of the absolute magnitudes of the eigenvector coefficients shown in Table VIII gives an indication of the significance of each eigenvector solution. The first solution, T_1, has a near zero coefficient for feature 8, thus indicating it serves to characterize compounds in which features 1–7 are present. The second solution, T_2, has a near zero coefficient for feature 1, thus indicating this solution to characterize compounds in which features 2–8 are present. These two eigenvectors therefore seem to discriminate principally between compounds which have joined as opposed to unjoined (feature 1) ring moieties and compounds which have a group occupying feature position 8 (analgesics) as opposed to compounds which have no such group present. In contrast, the third solution, T_3, and fourth solution, T_4, when compared to T_2, appear to characterize differing side-chain lengths. T_2 characterizes a side chain spanning features 2–6, T_3 characterizes a side chain spanning features 3–6, and T_4 characterizes a side chain spanning features 2–5. The four eigenvector solutions thus appear to account for all distinguishing characteristics referenced as features 1–8 for the compounds of Chart II.

Contrary to the previous example, a graphic representation of these results cannot be given since four factor axes are involved. Graphic projection methods may be employed so as to display the results in two or three dimensions,[2] but these can be difficult to interpret. As an alternative, one may note that the solutions T_1, T_2, and T_4 characterize the more important structural variants. These solutions may then be accepted as adequate for the definition of factor axes for the construction of a three-dimensional graph. Figure 2 depicts this graph. As can be noted, the anticholinergics, S, the antihistamines, H, and the analgesics, A, tend to associate in identifiable clusters. Alvarine (compound **23**), an anticholinergic, is recognized as distinct from the other anticholinergics, as the point for this compound is remote from the cluster of points for the other anticholinergics. The antidepressants, D, and the antipsychotics, P, also tend to form identifiable clusters. However, there is a degree of overlap between these clusters because of the close bioisosteric similarity in their structures. The anti-Parkinsonian agents, AP, form no identifiable cluster. These compounds are structurally similar to antipsychotics, antihistamines, or anticholinergics but have found therapeutic utility as anti-Parkinsonian agents. For these, it is clear, the

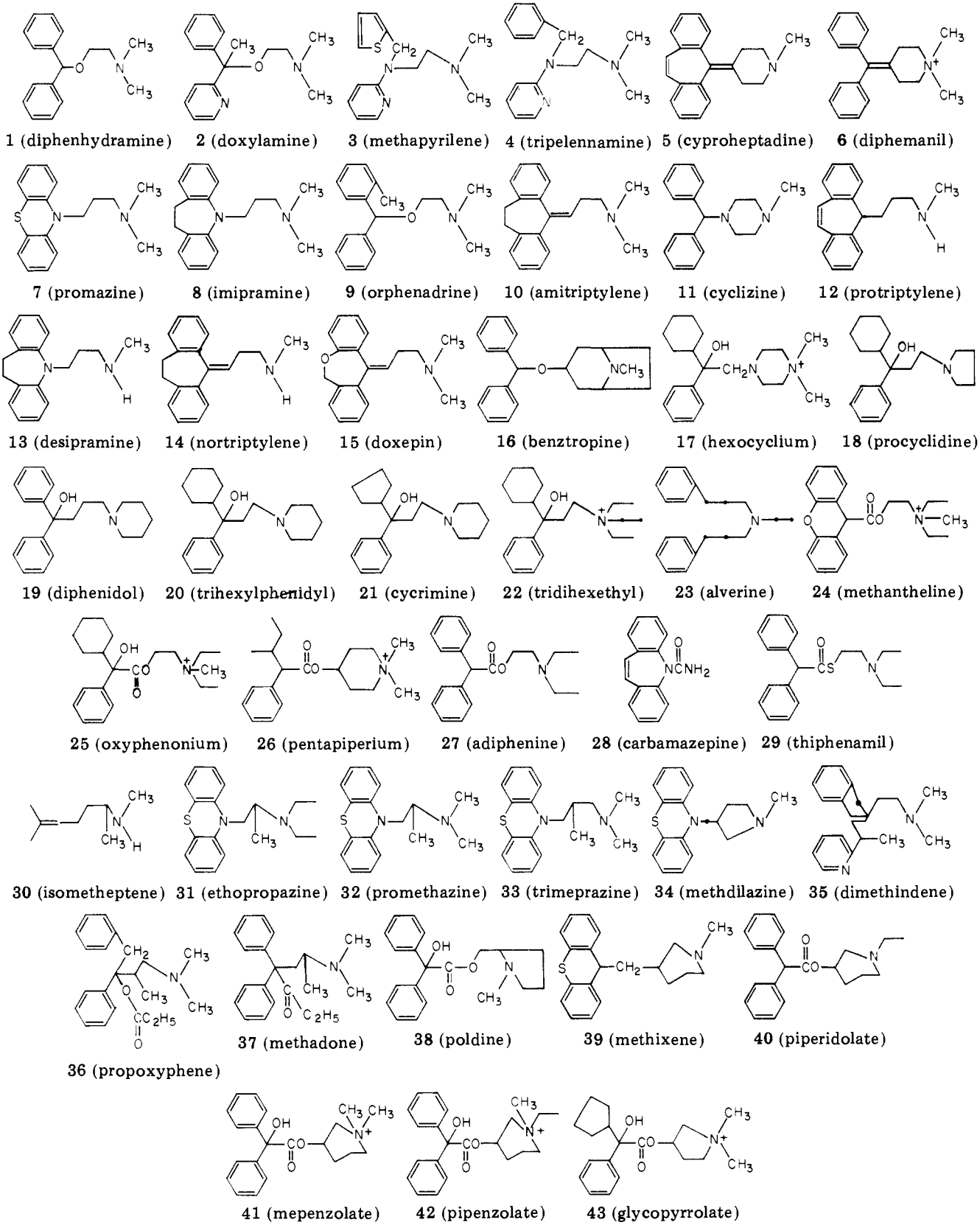
Table VII. Correlation Matrix for R Factor Analysis of Data of Table VI

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
1.000	−0.476	−0.341	0.212	−0.252	0.152	−0.532	−0.202
	1.000	0.380	0.110	0.141	−0.240	0.108	−0.335
		1.000	−0.206	−0.147	−0.044	−0.052	−0.194
			1.000	−0.098	−0.212	−0.236	−0.186
				1.000	0.281	0.202	−0.020
					1.000	−0.238	0.013
						1.000	0.305
							1.000

Table VIII. Accepted Principal-Component Solution for the Data of Table VI

Factor no.	Eigenvalue, λ	Coefficients in the linear combination T M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8
1	2.13	−0.61	0.42	0.31	−0.22	0.20	−0.19	0.46	0.08
2	1.71	0.05	0.45	0.35	0.34	−0.23	−0.26	−0.33	−0.55
3	1.33	0	0.09	0.23	−0.31	0.45	0.67	−0.27	−0.30
4	1.14	−0.02	0.20	−0.53	0.50	0.60	−0.03	0.14	−0.17

Chart II. A Set of Therapeutically Useful Antihistamines, Anticholinergics, Antipychotics, Antidepressants, Analgesics, and Anticonvulsants



clinically desirable effect can arise from a variety of alternative mechanistic modes.

Limitations of the Approach. The foregoing examples demonstrate the utility of pattern recognition in arriving at structure–activity relationships. As presently formulated, however, the method is not without its limitations. First among these is the necessity for recognizing, prior to analysis, the simplest possible molecular superposition to serve as a reference diagram. This limitation is not overly restricting when dealing with small sets of compounds, but upon expanding a set to include many hundreds of differing substances the limitation quickly

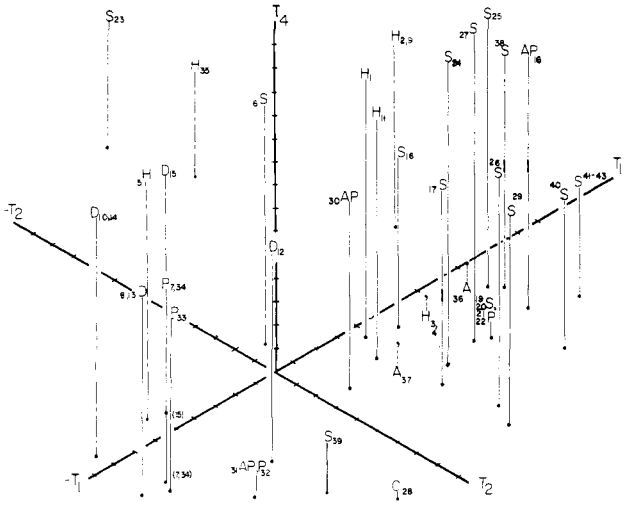


Figure 2. Partial factor space (three of the four factor axes) showing the resolution of compounds which are antihistamines (H), anticholinergics (S), antipsychotics (P), antidepressants (D), analgesics (A), and anti-Parkinsonics (AP).

becomes realized. A second limitation concerns the coding. Molar refraction can be used to define a coding for bioisosteric groups, but it is doubtful that this index will prove generally useful. The classical definition of bioisosterism, when extended to incorporate the concept of "physical bioisosterism",[27] recognizes bioisosteric groups as structural features which have similar values for one or more physical properties. No single physical parameter can thus provide a reliable basis of coding.

These limitations, while profound, are not so restricting that they could not be obviated. Suitably referenced connectivity matrixes could be compared using a computer and in this way a composite that would serve as a reference diagram could be constructed. The problem of codings might also be avoided by making use of indexes having a basis in graph theory,[42-45] as these may encompass structural as well as physical molecular attributes.

Conclusions

An elementary form of pattern recognition has been applied in arriving at structure–activity relationships. The approach taken differs from earlier attempts in many important respects. (a) All molecules in a set have their structures referenced relative to a composite-structure geometric diagram. This serves to superimpose the many ways lead substances may have been viewed in arriving at candidate agents and also formalizes the problem in a manner which is appropriate to the mathematics that is employed. (b) The coding adopted, which makes use of molar refractivities, is consistent with the classical definition of bioisosterism and necessarily leads to "clustering" of bioisosteric substances. (c) Either entire molecules or certain of their features, corresponding to pharmacophoric patterns of substitution, can be dealt with when attempting to arrive at structure–activity relationships. (d) The mathematical solutions arrived at can readily be interpreted in terms of molecular features, thus enabling intuitive extrapolations to be made from the results of a classification.

Acknowledgment. The support provided G.K.M. through a University Fellowship is gratefully acknowledged.

References and Notes

(1) B. R. Kowalski and C. F. Bender, *J. Am. Chem. Soc.*, **94**, 5632 (1972).
(2) B. R. Kowalski and C. F. Bender, *J. Am. Chem. Soc.*, **95**, 686 (1973).
(3) H. C. Andrews, "Introduction to Mathematical Techniques in Pattern Recognition", Wiley-Interscience, New York, N.Y., 1972.
(4) W. S. Meisel, "Computer-Oriented Approaches to Pattern Recognition", Academic Press, New York, N.Y., 1972.
(5) P. C. Jurs and T. L. Isenhour, "Chemical Applications of Pattern Recognition", Wiley-Interscience, New York, N.Y., 1975.
(6) T. L. Isenhour, B. R. Kowalski, and P. C. Jurs, *Crit. Rev. Anal. Chem.*, **4**, 1 (1974).
(7) B. R. Kowalski and C. F. Bender, *J. Am. Chem. Soc.*, **96**, 916 (1974).
(8) K. C. Chu, *Anal. Chem.*, **46**, 1181 (1974).
(9) K. C. Chu, R. J. Feldmann, M. B. Shapiro, G. F. Hazard, Jr., and R. I. Geran, *J. Med. Chem.*, **18**, 539 (1975).
(10) A. J. Stuper and P. C. Jurs, *J. Am. Chem. Soc.*, **97**, 182 (1975).
(11) R. J. Mathews, *J. Am. Chem. Soc.*, **97**, 935 (1975).
(12) J. T. Clerc, P. Naegeli, and J. Seible, *Chimia*, **27**, 639 (1973).
(13) C. L. Perrin, *Science*, **183**, 551 (1974).
(14) S. H. Unger, *Cancer Chemother. Rep., Part 2*, **4**, 45 (1974).
(15) K.-L. H. Ling, R. C. T. Lee, G. W. A. Milne, M. Shapiro, and A. M. Guarino, *Science*, **180**, 417 (1973).
(16) E. J. Lien and G. A. Gudauskas, *J. Pharm. Sci.*, **62**, 645 (1973).
(17) E. J. Lien and G. A. Gudauskas, *J. Pharm. Sci.*, **62**, 1968 (1973).
(18) R. D. Cramer III, G. Redl, and C. E. Berkoff, *J. Med. Chem.*, **17**, 533 (1974).
(19) W. T. Wipke, S. R. Heller, R. J. Feldmann, and E. Hyde, Ed., "Computer Representation and Manipulation of Chemical Information", Wiley, New York, N.Y., 1974.
(20) S. A. Mulaik, "The Foundations of Factor Analysis", McGraw-Hill, New York, N.Y., 1972.
(21) J. E. Overall and C. J. Klett, "Applied Multivariate Analysis", McGraw-Hill, New York, N.Y., 1972.
(22) M. L. Weiner and P. H. Weiner, *J. Med. Chem.*, **16**, 655 (1973).
(23) H. L. Friedman, *Natl. Res. Counc. (U.S.), Publ.*, **No. 306** (1951).
(24) A. Burger in "Medicinal Chemistry", Part I, 3rd ed, A. Burger, Ed., Wiley-Interscience, New York, N.Y., 1970, p 72.
(25) H. H. Harmon, "Modern Factor Analysis", University of Chicago Press, Chicago, Ill., 1967.
(26) N. H. Nie, D. H. Bent, and C. D. Hull, "Statistical Package for the Social Sciences (SPSS)", McGraw-Hill, New York, N.Y., 1970.
(27) C. Hansch, S. H. Unger, and A. B. Forsythe, *J. Med. Chem.*, **16**, 1217 (1973).
(28) Y. C. Martin, J. B. Holland, C. H. Jarboe, and N. Plotnikoff, *J. Med. Chem.*, **17**, 409 (1974).
(29) Program ARTHUR: Dr. B. Kowalski, Laboratory of Chemometrics, Department of Chemistry, University of Washington, Seattle, Wash.
(30) P. H. Weiner, E. R. Malinowski, and A. R. Levinstone, *J. Phys. Chem.*, **74**, 4537 (1970).
(31) S. S. Schiffman, *Science*, **185**, 112 (1974).
(32) C. Hansch in "Drug Design", Vol. I, E. J. Ariëns, Ed., Academic Press, New York, N.Y., 1971.
(33) A. Cammarata and K. S. Rogers in "Advances in Linear Free Energy Relationships", N. B. Chapman and J. Shorter, Ed., Plenum Press, New York, N.Y., 1973.
(34) A. Cammarata, unpublished results.
(35) A. Cammarata, *J. Med. Chem.*, **10**, 525 (1967).
(36) S. M. Free, Jr., and J. W. Wilson, *J. Med. Chem.*, **7**, 395 (1964).
(37) T. Fujita and T. Ban, *J. Med. Chem.*, **14**, 148 (1971).
(38) R. B. Barlow, "Introduction to Chemical Pharmacology", 2nd ed, Methuen, London, 1964, p 312.
(39) H. Erlenmeyer and N. Leo, *Helv. Chim. Acta*, **15**, 1171 (1932).

(40) Y. K. Syrkin and M. E. Dyatkina, "Structure of Molecules and the Chemical Bond", Dover Publications, New York, N.Y., 1964.
(41) C. Hansch and C. Silipo, J. Med. Chem., 17, 661 (1974).
(42) L. B. Kier, L. H. Hall, W. J. Murray, and M. Randic, J. Pharm. Sci., 64, 1971 (1975).
(43) L. H. Hall, L. B. Kier, and W. J. Murray, J. Pharm. Sci., 64, 1974 (1975).
(44) W. J. Murray, L. H. Hall, and L. B. Kier, J. Pharm. Sci., 64, 1978 (1975).
(45) L. B. Kier, W. J. Murray, and L. H. Hall, J. Med. Chem., 18, 1272 (1975).

Quantitative Structure–Activity Relationships among Steroids. Investigations of the Use of Steric Parameters

Robert A. Coburn and Alan J. Solo*

Department of Medicinal Chemistry, School of Pharmacy, State University of New York at Buffalo, Buffalo, New York 14214.
Received May 27, 1975

The importance of steric factors in quantitative structure–activity relationships involving steroid hormones is discussed. A variety of steric parameters, such as parachlor, molecular volume, van der Waals volume, and including difference and squared steric terms, is explored in an attempt to find preferred forms for such expressions. Improved correlations for 6-substituted 16-methylene-17α-acetoxy-4,6-pregnadiene-3,20-dione derivatives were found in which activity is related to π and a squared or difference steric factor. The activity of 9α-substituted cortisols correlates well with o_1 and a simple steric factor, provided that the 9α-hydroxylated compound is excluded from the series.

The use of extrathermodynamic linear free-energy relationships in the correlation of biological data from in vivo systems has resulted in growing experimental support.[1] The modification of this method by inclusion of physicochemical, theoretical (quantum chemical), and dummy parameters not derived from linear free-energy relationships represents a widespread stochastic approach to quantitative structure–activity relationships (QSAR).[2]

Relatively few QSAR studies have been reported for steroids. James has correlated the lipophilicity of testosterone and 19-nortestosterone esters with the prolongation of their biological effects.[3-5] However, if one excludes the, thus far, unpublished study mentioned by Ostrenga,[6] the only reports of quantitative correlations of steroid structure with activity are those of Wolff and Hansch on 9α-substituted cortisol derivatives[7] and on 6-substituted 16-methylene-17α-hydroxy-4,6-pregnadiene-3,20-dione acetate derivatives.[8] While these studies suggest that the multiparameter regression technique is of value in the study of steroids, the results to data have been less satisfying than applications in other areas. We discuss below previously ignored factors relating to the steric influence of substituents in QSAR studies and report methods leading to improved QSAR for steroids.

Following the submission of this article for review, a report by Topliss and Shapiro[9] appeared in which structure–activity relationships of 6-substituted 16-methylene-17α-hydroxy-4,6-pregnadiene-3,20-dione acetates were reappraised. In that report improved correlations were obtained by inclusion of a term involving the circumference of the 6-substituent. Although no evidence was offered relating circumference to a linear free-energy steric term, the finding suggests the importance of a steric factor in this correlation. In this investigation we confirm the importance of such a factor by obtaining further improvement in correlations employing a variety of more conventional steric terms. This study was then extended to a second group of compounds. This investigation was undertaken to contribute to an understanding of factors determining the type and optimal form of the steric term to be used in QSAR.

In view of the enormous number of steroid analogues which are known, the paucity of QSAR reported seems surprising. Part of this problem arises from the fact that relatively few large series of steroids, in which only a single substituent is systematically varied, have been prepared and assayed. A further complication arises from the variability of much of the in vivo assay data. For the steroid hormones an important additional complication stems from the interactions with a number of high-affinity relatively specific receptor sites including the hormonal receptors, the active sites of steroid metabolizing enzymes, and carrier proteins in the blood. In addition to these specific receptors, which all have limited capacity but high (and frequently similar) binding affinities, there are nonspecific binding sites which bind large quantities of steroid less tightly. While any of these factors may influence an assay, the classical in vivo assays which involve multiple dosing over a week or longer seem most likely to be influenced by factors which affect transport, rate of metabolism, and interaction with the hormonal receptor(s). This has been generally recognized in the past and was used in a qualitative sense in the design of steroid hormone analogues. To the extent that assays are affected by interaction of the test substance with secondary receptor sites, such as metabolizing enzymes, QSAR must reflect an optimization of transport and a balancing of those factors which maximize interaction with the primary receptor against those which minimize the interaction with the secondary receptors.

Competing processes, differently influenced by variation of substituent properties, would not in themselves be expected to lead to nonlinearity among free-energy correlations. Although the hydrophobic Hansch constant π is a linear free-energy parameter, biological activity is often better represented as a quadratic function of π or log P.[10] In dealing with steric interactions generally two types of parameters have been considered: Taft's constant E_s, or Hancock's modification E_s^c, and the physicochemical constant, molar refractivity (MR). It has been rationalized that either intra- or intermolecular steric interactions may have to be examined.[11] Since, as Hansch has cautioned,[12] the equivalence of molar refractivity with steric requirements can be misleading, we have chosen to investigate several other parameters which may be more directly related to substituent steric influence. [Most substituents found in this study contain π bonds or nonbonded electrons suggesting that polarizability contributes substan-