Pattern Recognition of Steroids Using Fragment Molecular Connectivity

DOUGLAS R. HENRY and JOHN H. BLOCKˣ

Received May 18, 1979, from the *School of Pharmacy, Oregon State University, Corvallis, OR 97331.* Accepted for publication July 6, 1979.

Abstract □ The use of pattern recognition methods to classify a set of steroids into five therapeutic categories was investigated. First-order fragment molecular connectivity values were determined for 10 positions on each molecule using a template-based method of position assignment. Learning set and test set classifications were performed. Although the numbers of compounds misclassified were comparable for all of the methods, the identities of the misclassified compounds varied depending on whether the classification method assumed a local or a global view of the data. The best classification results were comparable to those obtained by linear and quadratic discriminant analyses. For this set of compounds, it was concluded that pattern recognition methods offer no advantages over traditional discriminant analysis methods if classification alone is considered, especially since most discriminant analysis procedures utilize stepwise variable selection, which is not as common in pattern recognition analyses.

Keyphrases □ Steroids—pattern recognition using fragment molecular connectivity □ Fragment molecular connectivity—pattern recognition of steroids □ Pattern recognition methods—classification of steroids in therapeutic categories using fragment molecular connectivity

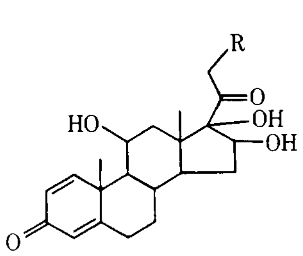
The use of fragment molecular connectivity values as position descriptors in a template-based discriminant analysis has been demonstrated as a means of classifying structures according to their type of pharmacological activity (1). The use of statistical discriminant analysis in such classification problems is optimal for multivariate, normal random variables (2). For observations that are not distributed normally with respect to the descriptor variables, a discriminant analysis may not be the optimal classification method; certain distribution-free pattern recognition techniques may be more suitable (3). Although pattern recognition methods are not used as commonly in medicinal chemistry as are regression analysis and analysis of variance (4), the application of these techniques is increasing (5–9).

Accordingly, a number of classification-type pattern recognition procedures were applied to a set of steroidal compounds studied previously using linear and quadratic discriminant analysis[1]. These compounds show markedly nonnormal distributions with respect to many of the position variables used to describe them (Table I). It was felt that better classification results and greater insight into the value of molecular connectivity in classification problems possibly could be gained with pattern recognition methods.

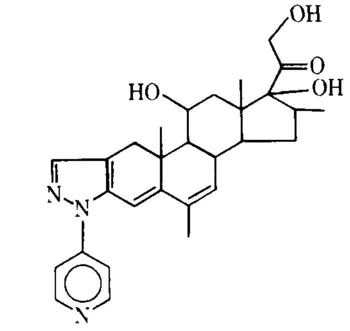
Four such methods were selected: Andrews function curves (10), *K* nearest neighbor analysis (11, 12), multicategory linear learning machine analysis (13), and statistical isolinear multicomponent analysis (SIMCA)[2] (14, 15). The use of the Andrews function for representing multivariate data in statistical research is well documented (16). The remaining methods were used as part of ARTHUR, a pattern recognition program (17, 18).

EXPERIMENTAL

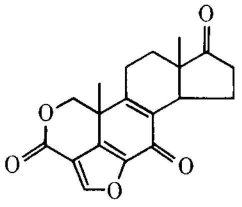
A learning set of 46 steroidal compounds, classified into five therapeutic categories, was used (Table I). A test set of nine steroids also was selected (I–IX) (19). The pharmacological categories of these compounds are estrogens (E or ESTR), progestogens (P or PROG), androgens (A or ANDR), corticosteroids (C or CORT), and cardiac steroids (H or CARD). A template structure (6) was designated, and 10 positions of interest were selected. Each compound in the study was superimposed on the template



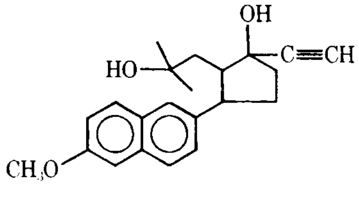
I: C



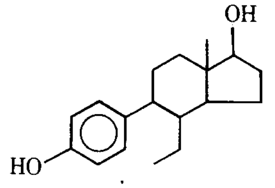
II: C



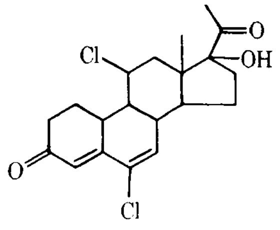
III: C



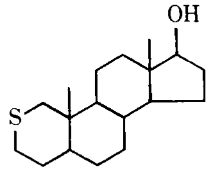
IV: E



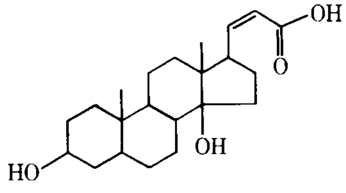
V: E



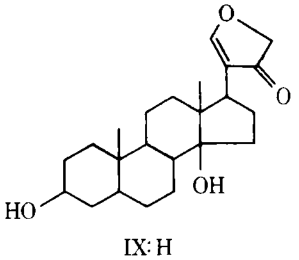
VI: P



VII: A



VIII: H



IX: H

[1] See D. R. Henry and J. H. Block, *Eur. J. Med. Chem.*, in press.

[2] The term SIMCA also can serve as an acronym for SIMilarity Components Analysis, SImple Modeling by Class Analogy, and other methods.

0022-3549/80/0900-1030$01.00/0

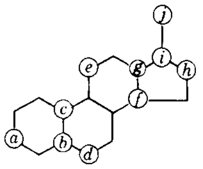


Table I—Raw First-Order Molecular Connectivity Values of Compounds in the Analyses

Compound[a]	Class[b]	Position									
		a	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*	*j*
1 Diethylstilbestrol	E	0.801	0.622	0.827	0.707	1.061	0.827	0.622	0.622	0.801	0.224
2 Dienestrol	E	0.801	0.622	0.827	0.577	0.204	0.827	0.622	0.622	0.801	0.224
3 Methallestril	E	0.781	0.827	0.827	0.622	0.707	1.105	1.115	1.000	0.854	0.781
4 Benzestrol	E	0.801	0.622	0.866	0.707	1.115	0.866	0.622	0.622	0.801	0.224
5 Promethestrol	E	0.742	0.577	0.866	0.707	1.115	0.866	0.622	0.622	0.781	0.408
6 Estradiol	E	0.801	0.892	0.827	0.854	0.908	1.030	1.431	0.908	0.955	0.258
7 Estrone	E	0.801	0.892	0.827	0.854	0.908	1.030	1.392	0.854	0.808	0.204
8 Estriol	E	0.801	0.892	0.827	0.854	0.908	1.030	1.431	0.999	0.880	0.258
9 Equilin	E	0.801	0.892	0.827	0.854	0.908	1.030	1.392	0.854	0.808	0.204
10 Equilinin	E	0.801	0.827	0.789	0.622	0.854	0.986	1392	0.854	0.808	0.204
11 Progesterone	P	0.846	0.892	1.392	0.854	0.908	1.030	1.392	0.854	1.077	0.954
12 17-α-Hydroxyprogesterone	P	0.846	0.892	1.392	0.854	0.908	1.030	1.392	0.854	1.077	0.954
13 Medroxyprogesterone	P	0.846	0.827	1.392	1.274	0.908	1.030	1.392	0.854	1.077	0.954
14 Dihydroprogesterone	P	0.846	0.827	1.392	0.622	0.908	1.030	1.431	0.908	0.986	0.933
15 Ethisterone	P	0.846	0.827	1.392	0.622	0.908	1.030	1.431	0.908	0.986	0.993
16 Norethindrone	P	0.846	0.892	1.392	0.854	0.908	1.030	1.392	0.854	1.077	0.224
17 Norethynodrel	P	0.846	0.931	1.030	0.854	0.908	1.030	1.392	0.854	1.077	0.224
18 Dimethisterone	P	0.846	0.931	1.030	0.854	0.908	1.030	1.246	0.854	1.077	0.224
19 Norgestrel	P	0.846	0.931	1.030	0.854	0.908	1.030	1.246	0.854	1.077	0.224
20 Ethynodiol	P	0.999	0.931	1.030	0.854	0.908	1.030	1.392	0.854	1.077	0.224
21 Testosterone	A	0.846	0.892	1.392	0.854	0.908	1.030	1.430	0.908	0.955	0.258
22 α-Methyltestosterone	A	0.846	0.892	1.392	0.854	0.908	1.030	1.392	0.854	1.327	0.224
23 Oxymetholone	A	0.808	1.105	1.430	0.908	0.908	1.030	1.392	0.854	1.327	0.224
24 Nandrolone	A	0.846	0.931	1.030	0.854	0.908	1.030	1.431	0.908	0.955	0.258
25 Dromostanolone	A	0.846	1.105	1.431	0.908	0.908	1.030	1.431	0.908	0.955	0.258
26 Stanozolol	A	0.827	1.105	1.431	0.908	0.908	1.030	1.392	0.854	1.327	0.224
27 Ethylestrenol	A	0.908	0.931	1.030	0.854	0.908	1.030	1.392	0.854	1.181	0.224
28 Methandrostenolone	A	0.781	0.892	1.327	0.854	0.908	1.030	1.392	0.854	1.327	0.224
29 Oxandrolone	A	0.762	1.105	1.431	0.908	0.908	1.030	1.392	0.854	1.327	0.224
30 Hydrocortisone	C	0.846	0.931	1.030	0.854	1.000	1.030	1.392	0.854	1.077	0.808
31 Prednisone	C	0.781	0.892	1.327	0.854	0.846	1.030	1.392	0.854	1.077	0.808
32 Methylprednisolone	C	0.781	0.827	1.327	1.274	1.000	1.030	1.392	0.854	1.077	0.808
33 Triamcinolone	C	0.781	0.892	1.289	0.854	0.955	1.030	1.392	0.955	1.012	0.808
34 Fluorandrenolone	C	0.781	0.827	1.327	0.567	1.000	1.030	1.392	0.955	1.012	0.808
35 Dexamethasone	C	0.781	0.892	1.288	0.854	0.955	1.030	1.392	1.274	1.012	0.808
36 Paramethasone	C	0.781	0.827	1.327	0.567	1.000	1.075	1.030	1.274	1.051	0.808
37 Flumethasone	C	0.781	0.827	1.288	0.567	0.955	1.030	1.392	1.274	1.012	0.808
38 Fluprednisolone	C	0.781	0.827	1.327	0.567	1.000	1.030	1.392	0.854	1.077	0.808
39 Halcinolone	C	0.846	0.827	1.354	0.854	0.955	1.030	1.392	0.955	1.012	0.808
40 Digitoxigenin	H	1.052	1.105	1.431	0.908	0.908	1.116	1.392	0.908	0.986	0.931
41 Digoxigenin	H	1.052	1.105	1.431	0.908	0.816	1.116	1.193	0.908	0.986	0.931
42 Gitoxigenin	H	1.052	1.105	1.431	0.908	0.908	1.116	1.392	0.999	0.911	0.931
43 Ouabain	H	1.052	1.181	1.181	0.854	0.999	1.116	1.392	0.908	0.986	0.931
44 Strophanthidin	H	1.052	1.181	1.181	0.854	0.908	1.116	1.392	0.908	0.986	0.931
45 Proscillaridin	H	0.927	0.892	1.392	0.854	0.908	1.116	1.392	0.908	0.986	0.781
46 Bufogenin B	H	1.075	1.105	1.431	0.908	0.908	1.116	1.392	0.977	0.911	0.866
Mean		0.853	0.908	1.196	0.835	0.912	1.029	1.310	0.892	1.016	0.538
SD		0.090	0.141	0.234	0.162	0.121	0.056	0.227	0.131	0.147	0.324

[a] Structures may be found in Refs. 25 and 26. [b] Key: E = estrogen, P = progestogen, A = androgen, C = corticosteroid, and H = cardiac steroid.

structure, and a value of the descriptor index was assigned to each of the 10 positions on the molecule.

The selected descriptor was first-order molecular connectivity ($^1\chi^v$), which was calculated as the sum of the first-order terms for each bond joining the position of interest (20). This descriptor was selected because this index, when compared to molar refraction and a number of other fragment molecular connectivity indexes, showed lower rates of misclassification when used in linear and quadratic discriminant analyses (1). Using the molecules in the learning set, each position variable was standardized to zero mean and unit variance prior to use. An observation to variable ratio of about 10 was sought. Accordingly, variable selection was performed using both the Fisher and variance weighting methods (18).

The top five variables from a variance-weighted standpoint were, in order of importance, positions *j*, *a*, *i*, *c*, and *b*. These variables were the same as those selected in the stepwise discriminant analysis of the data. Using Fisher weighting, the same variable subset was selected, except that position *f* replaced *b*, which assumed sixth place in significance. Since position *f* had a constant value for three of the groups in the analysis (estrogens, progestogens, and cardiac steroids), the five-variable subset selected by variance weighting was adopted as a basis for performing the pattern recognition analyses.

RESULTS AND DISCUSSION

Andrews Functions—Andrews (21) introduced a trigonometric series expansion with orthonormal coefficients, which is useful for representing multivariate data in two dimensions. This function takes the form:

$$F(x) = x_1\,0.707 + x_2 \sin t + x_3 \cos t + x_4 \sin 2t + x_5 \cos 2t + \cdots \quad \text{(Eq. 1)}$$

where the *x* values are the values of the respective variables for a given observation. For each observation, a plot of this function over the interval $t = [-\pi,\pi]$ radians gives a curve that is unique to the observation but similar in shape and amplitude to the curves of observations having similar *x* values (*i.e.*, observations close to each other in multidimensional space). The order of the *x* variables in Eq. 1 usually is determined based on the contribution of the variable to the between-group differences, with

Table II—Classification by Andrews Function [a]

True Group	Predicted Group					Percent
	ESTR	PROG	ANDR	CORT	CARD	Correct[b]
ESTR	10	0	0	0	0	100.0
PROG	0	7	3	0	0	66.7
ANDR	0	2	5	0	2	55.6
CORT	0	4	0	6	0	60.0
CARD	0	0	1	0	6	85.7
Mean of group	−3.65	0.30	1.22	−0.07	3.33	
SD	0.98	0.67	1.29	0.24	0.90	

[a] For $t = 0.28$ rad, $F(x) = 0.707\,a + 0.267\,j + 0.964\,c + 0.514\,i + 0.858\,b$. Variables are ordered according to their contribution to the between-group variation. [b] Learning set results showed 74.5% correct. Compounds misclassified (predicted group) were: 11–13 (ANDR), 21 (PROG), 23 (CARD), 26 (CARD), 27 (PROG), 31 (PROG), 34 (PROG), 37 (PROG), 39 (PROG), and 45 (ANDR). Test set results showed 77.8% correct. Compounds misclassified (predicted group) were II (PROG) and III (ESTR).

the most significant variable being assigned to x_1 and the least significant variable assigned to the last term in the series.

Andrews function plots of prototypes of each therapeutic class are shown in Fig. 1. For this purpose, the group centroids (mean values) of each class were considered to be representative of the class. Other possibilities would be to use median or mode values of the variables for a given group. Each curve in Fig. 1 represents a typical member of its class. There are points along the t axis where between-group separations are large.

Figure 2 shows a plot of the total squared separation, summed over all possible pairs of groups (10 terms). The maximum in overall between-group separation occurs at 0.28 radian. By fixing t at this value, the Andrews function becomes a form of the discriminant function. A histogram of the values of this function for all of the observations is seen in Fig. 3.

Table II shows detailed classification results for the compounds. The results were obtained by determining the distance, in standard deviation units, of the Andrews function value for each observation from the mean value of each group. The observation then was classified into the group for which the absolute value of this distance was the smallest.

A comparison of the results in Table II with the corresponding results of a linear discriminant analysis (Table III) shows that the single Andrews discriminant function performed slightly worse in classifying the learning set (74% correct) compared to a discriminant analysis in which three canonical discriminant functions were used (83% correct). However, for the test set, the Andrews method correctly classified seven out of nine cases while the discriminant analysis was correct in only six cases.

Better classification results might be obtained from the Andrews function by selecting more than one value of t and generating several discriminant functions. Other values of t for which the between-group separation is large are indicated in Fig. 2. As an alternative, a separate value of t and its corresponding function could be selected for each pair of groups, and an observation could be classified based on the number of pairwise comparisons. Neither approach was tried during the present work.

***K* Nearest Neighbor Analysis**—This classification method conceptually is one of the simplest in the pattern recognition literature (22, 23). It differs from most minimum-distance classifiers in that the group centroid is not selected as the prototype of a given class. Instead, an observation is classified into the group to which a majority of its K nearest neighbors in multidimensional space belong. Values of K usually range from one to 10, but classification results suffer whenever the value of K approaches the average class size. In ties, the group showing the smallest aggregate distance is selected.

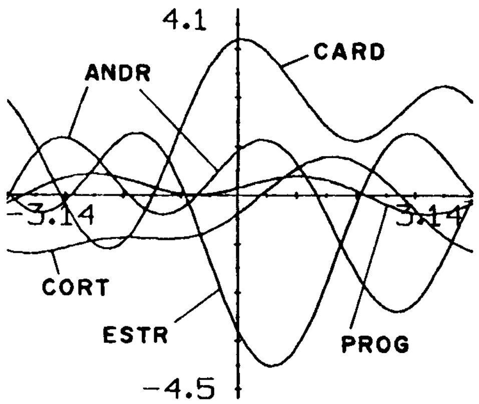


Figure 1—*Andrews function plots of the group centroids of each therapeutic category. The mean values of the position variables* a, j, c, i, *and* b *were used as coefficients in the Andrews function in the order given.*

Table III—Discriminant Analysis Results for Compounds in the Learning Set [a]

Variable	Discriminant Score Coefficients				
	ESTR	PROG	ANDR	CORT	CARD
a	−9.36	2.08	−0.36	−3.36	15.67
b	−1.02	−0.73	1.16	−0.38	1.56
c	−7.09	1.06	2.87	−0.99	6.35
i	−4.24	1.19	2.29	0.50	0.71
j	−1.14	−0.14	−3.16	2.10	2.89
Constant	−13.76	−2.08	−5.60	−3.50	−21.97

[a] Linear classification results were obtained using BMDP7M (27); average results were 82.6% correct. Quadratic classification results were calculated using the University of Wisconsin MULTDIS program (28); average results were 89.1% correct. Test set classification results (both linear and quadratic classification) were 66.7% correct.

Table IV shows the results of K nearest neighbor analyses for the learning and test sets of compounds. For the learning set, the results at all levels of K were comparable to the best results obtained by discriminant analysis. For the test set, four compounds were misclassified for all values of K except the first value. This total is one misclassification more than was found with the linear and quadratic discriminant analysis methods.

Multicategory Linear Learning Machine Analysis—In a manner similar to linear discriminant analysis, a linear learning machine seeks to define a function or boundary that will maximally separate a given group of observations from all other observations (13). However, instead of relying on a fixed statistical criterion (such as maximizing the ratio of between-group variation to within-group variation), a learning machine generates the classification function by iteration and feedback (24). A vector of weighting coefficients is derived for each group. Then, for each observation, a discriminant score is calculated for each group, and the observation is classified into the group for which the discriminant score is highest.

Table V summarizes the results of a multicategory linear learning machine analysis of the steroid data. A total of 201 iterations was performed, although the classification results ceased to improve after 186 cycles through the data. Thirty-nine of the learning set (84.8%) and six of the test set (66.7%) compounds were classified correctly by this method. In each case, the misclassified compounds and the groups into which they were placed were identical to those misclassified in a linear discriminant analysis of the data. This result occurred in spite of the fact that some notable differences existed both in the relative magnitude and in the sign between the weighting vectors derived by the learning machine (Table V) and those calculated in the discriminant analysis (Table III).

All of the misclassified compounds were placed by the learning machine

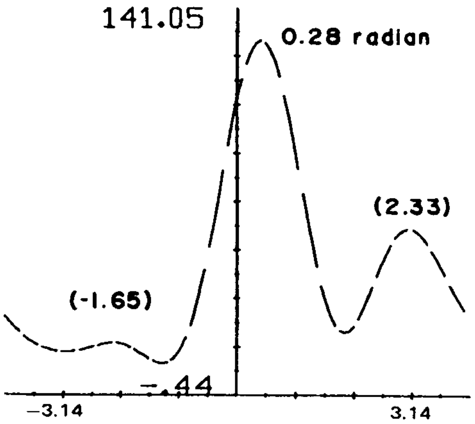


Figure 2—*Plot of the total sum of squares of differences between each possible pair of curves shown in Fig. 1. Values in parentheses are values of* t *(radians) for which local maxima exist; these values of* t *could be used to generate additional discriminant functions.*

Table IV—Results of *K* Nearest Neighbor Analysis [a]

K	Percent Correct Learning Set	 Test Set[b]
1	87.0	66.7
3	89.1	55.6
4	89.1	55.6
5	89.1	55.6
6	87.0[c]	55.6[d]

[a] Classification was performed in the space of the standardized position variables *a*, *b*, *c*, *i*, and *j*. [b] Each observation in the test set was classified on the basis of observations in the learning set, omitting other test set compounds. [c] Learning set compounds misclassified (predicted group) were 21 (PROG), 24 (PROG), 25 (PROG), 27 (PROG), 39 (PROG), and 45 (PROG). [d] Test set compounds misclassified (predicted group) were III (ESTR), IV (PROG), VI (CORT), and VII (PROG).

Table V—Multicategory Linear Learning Machine Results: Final Weighting Vectors [a]

Variable	ESTR	PROG	ANDR	CORT	CARD
a	0.357	0.646	0.469	−0.442	1.522
b	−1.250	0.747	1.392	−0.992	0.219
c	−0.828	0.373	0.719	−0.498	0.191
i	−0.643	0.194	0.113	0.010	0.209
j	0.527	0.283	−0.102	0.610	0.781
Constant	0.147	0.643	0.601	0.150	0.194

[a] Learning set results were 84.8% correct. Compounds misclassified (predicted group) were 13-15 (CORT), 16 (ANDR), and 19 (ANDR). Test set results were 66.7% correct. Compounds misclassified (predicted group) were IV (CORT), VI (CORT), and VIII (ANDR).

Table VI—Principal Component Models from a SIMCA Study

Group	Component[a]	*a*	*b*	*c*	*i*	*j*	Percent Correct in Group[b]
ESTR	1	−0.099	−0.150	−0.228	−0.189	−0.110	100.0
	2	−0.091	−0.967	0.052	−0.231	−0.004	
PROG	1	0.023	−0.018	0.025	0.004	−0.006	100.0
	2	0.208	0.177	−0.526	0.112	−0.797	
ANDR	1	−0.039	0.092	0.080	0.174	−0.139	55.6
	2	0.220	−0.235	−0.307	−0.895	0.035	
CORT	1	−0.097	−0.054	0.060	0.026	0.125	80.0
CARD	1	0.299	0.199	0.101	−0.052	0.167	85.7
	2	0.550	0.770	−0.271	−0.027	0.176	
	3	0.429	−0.032	0.841	−0.327	0.045	

[a] Cross-validation was performed to select the optimum number of components for each group, based on a partial *F* test that measures the significance of an added component, given the components already present. [b] Average learning set results were 84.8% correct. Compounds misclassified (predicted group) were 21 (PROG), 24 (PROG), 25 (PROG), 27 (PROG), 30 (PROG), 39 (PROG), and 45 (PROG). Average test set results were 66.7% correct. Compounds misclassified (predicted group) were IV (PROG), VI (CORT), and VII (PROG).

into the progestogen category. Examination of two- and three-dimensional scatter plots and Andrews curves of the progestogens revealed that this class of compounds showed wide within-group variation with respect to the position variables selected. To be classified successfully by a linear learning machine, the compounds must be linearly separable in the space of the descriptor variables, which perhaps is not the case here.

Statistical Isolinear Multicomponent Analysis—A more sophisticated method of pattern recognition than those mentioned previously was developed by Wold (14). The SIMCA method involves generating a separate principal components model for each class of observations. Components are generated iteratively, and the number of components retained for each class may vary. For each group, the model so obtained is solved by least-squares techniques to determine the relationships between the original variables and the derived principal components.

An observation can be classified, and goodness of fit can be determined, by calculating the principal component scores of the observation for each group and then determining the theoretical *x* values the observation should have if it were a member of the given group. The sum of squares of the differences between the calculated *x* values and the observed ones (*i.e.*, the residuals from the regression models) is a measure of how well the observation fits the model for a particular group. An observation can be placed into the group for which this deviation is smallest or, depending on some preset criteria, it may be placed into a new class instead of into any existing group.

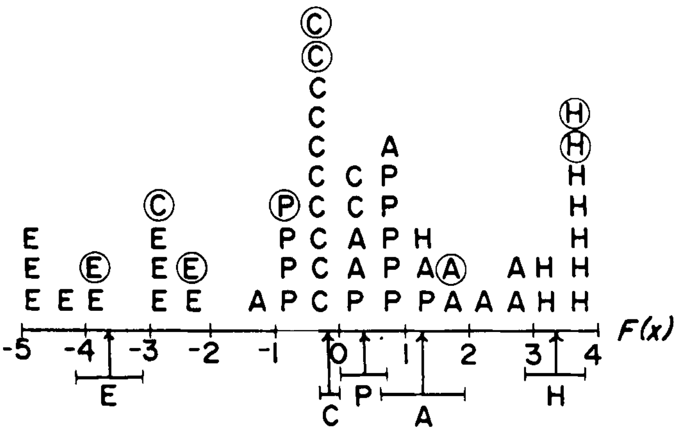


Figure 3—*Histogram of Andrews function values for the compounds in this study. Groups are estrogens* (*E*), *progestogens* (*P*), *androgens* (*A*), *corticosteroids* (*C*), *and cardiac steroids* (*H*). *Circled entries are test set compounds. Group means* (±SD) *are shown below the axis.*

Table VI summarizes the results of a SIMCA study of the data. The number of compounds misclassified by the SIMCA method for both the learning and test sets was the same as for the multiclass separator and the discriminant analysis techniques. However, the identities of the compounds misclassified were different. Those misclassified by the SIMCA method had more in common with the compounds misclassified by the *K* nearest neighbor analysis. This result may reflect a fundamental difference between the multiclass separator and discriminant analysis on the one hand and SIMCA and *K* nearest neighbor analysis on the other. The SIMCA and *K* nearest neighbor methods rely more on local properties of the data set (*i.e.*, the distribution of compounds in the immediate neighborhood of the compound being classified). By contrast, the discriminant function methods rely on a more global view of the data set, considering all groups at once or, in the case of the multiclass separator, comparing a given group with all other groups combined.

CONCLUSIONS

First-order molecular connectivity values contain information that can be used, with varying degrees of success, in the classification of molecular structures into their correct therapeutic categories using pattern recognition techniques. This has been demonstrated using a template-based method of variable assignment. Of course, any method of assigning position variables may be criticized on the basis of its subjectivity. An alternative method, involving whole-molecule molecular connectivity terms, is being studied, and encouraging results have been obtained[3].

For the compounds in this study, pattern recognition methods did not give better classification results than those obtained with the traditional discriminant analysis techniques. The latter methods, due to their statistical nature, have the advantage that stepwise procedures may be employed for variable selection. Such stepwise techniques are not used as commonly in most pattern recognition research.

The classification of observations *per se* is not usually considered a sufficient goal for research of the type reported here. The use of model-building methods of analysis, such as SIMCA, provides a basis for extending such investigations to include both qualitative and quantitative predictions (9). However, it is important that a variety of descriptor variables be studied to determine which variables are most suitable for particular purposes. Molecular connectivity indexes, because of their close relationship to molecular structure, have unique potential for classification problems.

REFERENCES

(1) D. R. Henry and J. H. Block, *J. Med. Chem.*, **22**, 465 (1979).

(2) P. A. Lachenbruch, "Discriminant Analysis," Hafner, New York, N.Y., 1975, p. 40.

[3] See D. R. Henry, Ph.D. thesis, Oregon State University, Corvallis, Ore. (to be published).

(3) L. Kanal, *IEEE Trans. Inform. Theory,* **IT-20,** 697 (1974).

(4) Y. C. Martin, "Quantitative Drug Design–A Critical Introduction," Dekker, New York, N.Y., 1977.

(5) A. J. Stuper and P. C. Jurs, *J. Pharm. Sci.,* **67,** 745 (1978).

(6) A. Cammarata and G. K. Menon, *J. Med. Chem.,* **19,** 739 (1976).

(7) G. K. Menon and A. Cammarata, *J. Pharm. Sci.,* **66,** 304 (1977).

(8) W. J. Dunn, III, and S. Wold, *J. Med. Chem.,* **21,** 1001 (1978).

(9) W. J. Dunn, III, and Y. C. Martin, *ibid.,* **21,** 922 (1978).

(10) D. F. Andrews, in "Discriminant Analysis and Applications," T. Cacoullos, Ed., Academic, New York, N.Y., 1973, pp. 37–47.

(11) E. Fix and J. L. Hodges, U. S. School of Aviation Medicine, Project 21-49-004, Report 4, Randolph Field, Tex., 1951; reprinted in "Machine Recognition of Patterns," A. K. Agrawala, Ed., IEEE Press, New York, N.Y., 1977, pp. 261–279.

(12) T. M. Cover and P. E. Hart, *IEEE Trans. Inform. Theory,* **IT-13,** 21 (1967).

(13) N. J. Nilsson, "Learning Machines," McGraw-Hill, New York, N.Y., 1965.

(14) S. Wold, *Pattern Recognition,* **8,** 127 (1976).

(15) S. Wold and M. Sjöstrom, in "Chemometrics–Theory and Practice," B. R. Kowalski, Ed., ACS Symposium Series No. 52, American Chemical Society, Washington, D.C., 1977, pp. 243–282.

(16) R. Gnanadesikan, "Methods for Statistical Analysis of Multivariate Observations," Wiley, New York, N.Y., 1977, pp. 203–225.

(17) D. L. Duewer, J. R. Koskinen, and B. R. Kowalski, "ARTHUR," available from Alice M. Harper, Department of Chemistry, University of Georgia, Athens, GA 30602.

(18) A. M. Harper, D. L. Duewer, B. R. Kowalski, and J. L. Fasching, in "Chemometrics–Theory and Practice," B. R. Kowalski, Ed., ACS Symposium Series No. 52, American Chemical Society, Washington, D.C., 1977, pp. 14–52.

(19) M. J. Green and B. N. Lutsky, in "Annual Reports in Medicinal Chemistry," vol. 11, F. H. Clarke, Ed., Academic, New York, N.Y., 1976, chap. 16, pp. 149–157.

(20) L. B. Kier and L. H. Hall, *Eur. J. Med. Chem.,* **12,** 307 (1977).

(21) D. F. Andrews, *Biometrics,* **28,** 125 (1972).

(22) P. E. Hart, *IEEE Trans. Inform. Theory,* **IT-14,** 515 (1968).

(23) J. H. Friedman, F. Baskett, and L. J. Shustek, *IEEE Trans. Comput.,* **C-24,** 1000 (1975).

(24) J. T. Tou and P. Gonzales, "Pattern Recognition Principles," Addison-Wesley, Reading, Mass., 1974, pp. 181–186.

(25) D. S. Fullerton, in "Textbook of Organic Medicinal and Pharmaceutical Chemistry," 7th ed., C. O. Wilson, O. Gisvold, and R. F. Doerge, Eds., Lippincott, Philadelphia, Pa., 1977, chap. 20, pp. 731–823.

(26) W. C. Cutting, "Cutting's Handbook of Pharmacology," 5th ed., Appleton-Century Crofts, New York, N.Y., 1972.

(27) "BMDP-Biomedical Computer Programs–1977," W. J. Dixon, Ed., University of California Press, Berkeley, Calif., 1977, pp. 711–733.

(28) R. A. Eisenbeis and R. B. Avery, "Discriminant Analysis and Classification Procedures—Theory and Applications," Heath, New York, N.Y., 1972.

ACKNOWLEDGMENTS

Funds for computer time were supplied by the Oregon State University Computer Center.

Structural Information from Molecular Connectivity $^4\chi_{PC}$ Index

LEMONT B. KIER

Received June 5, 1978, from the *Department of Pharmaceutical Chemistry, Medical College of Virginia, Virginia Commonwealth University, Richmond, VA 23298.* Accepted for publication March 26, 1979.

Abstract □ The molecular connectivity $^4\chi_{PC}$ index was examined for its ability to describe uniquely molecules containing substituted benzene rings. The subgraphs comprising this index were shown to encode information about the number, placement, and type of ring substituents. Several examples illustrate the ability of the index to describe structure-influencing properties.

Keyphrases □ Molecular connectivity—description of molecules containing substituted benzene rings by $^4\chi_{PC}$ index □ Structure–activity relationships—molecules containing substituted benzene rings, interpretation of structure–activity relationship from molecular connectivity $^4\chi_{PC}$ index

Since the development of a new method of molecular structure quantitation called molecular connectivity, it has been utilized in numerous structure–activity relationship studies (1–7). The numerical indexes computed for each molecule are rich in information content; hence, constellations of indexes are of considerable value in describing structural features contributing to the numerical value of a physical property or biological activity. This study explored the information content of one important index, $^4\chi_{PC}$, and revealed how it plays a prominent role in several structure–activity relationship analyses.

THEORY

The molecular connectivity description of molecular structure gives rise to several numerical indexes of the general form $^m\chi_t$, where m is the order of the molecular fragment and t is the type. Indexes may be of the simple connectivity (unweighted adjacency) or valence level. The indexes are weighted counts of fragments within a molecule, conveying information about topological features such as molecular size, branching, cyclization, unsaturation, and heteroatom location and type.

One distinct advantage of a molecular connectivity analysis of structure in a structure–activity relationship study is that the indexes correlating with activity in a regression analysis can be interpreted directly in terms of structural fragments meaningful to the medicinal chemist (8–10). Depending on the study, various indexes will emerge from searches with one or more variables, each conveying various amounts of structural information.

It has become apparent in the studies conducted in these laboratories that certain patterns of index appearance are found in analyses of molecular structure using molecular connectivity. One noteworthy appearance is the $^4\chi_{PC}$ (or the $^4\chi^v_{PC}$) index in studies on the structures of molecules containing substituted benzene rings. This index frequently is important as a second or third variable in regression analyses on molecules in which the benzene rings possess different numbers, positions, and types of substituents.

This recurrence led to the belief that the $^4\chi_{PC}$ index carries a high degree of information content in these structural classes that is common to many drug molecules. The purpose of this report is to analyze and

0022-3549/ 80/ 0900-1034$01.00/ 0