Articles

Classification of Drugs by Discriminant Analysis Using Fragment Molecular Connectivity Values

Douglas R. Henry and John H. Block*

School of Pharmacy, Oregon State University, Corvallis, Oregon 97331. Received June 19, 1978

An investigation was made into the use of linear and quadratic discriminant analysis, along with K nearest-neighbor analysis, in the classification of a set of 51 compounds which were divided into five therapeutic categories. By superimposing each compound on a pattern structure, as first proposed by Cammarata, eight positions were assigned on the molecule. Each position was coded with the numerical value of a descriptor index. Relative molar refraction, which was the index used by Cammarata, was compared with a number of molecular connectivity indices. For each of the indices studied, it was found that only four of the eight positions contributed significantly to between-class differences. It was also found that first-order molecular connectivity, calculated as the sum of the contributions of each of the bonds joining a given position, resulted in consistently fewer misclassifications as compared with the other indices. Using first-order molecular connectivity, validation procedures were performed on the original set of compounds, on random samples drawn from this set, and on a set of ten compounds not included in the analysis. The results obtained were highly data dependent, but they, nevertheless, suggest that molecular connectivity indices should prove useful in structural classification procedures.

Throughout the last decade, a good deal of research in medicinal chemistry has been directed toward the study of quantitative relationships between physicochemical properties and the levels of biological activity of drugs.[1,2] Such QSAR's lend themselves readily to the application of the usual univariate statistical methods of regression analysis and analysis of variance. Lately, however, interest has arisen in the study of a form of a qualitative structure–activity relationship in which one attempts to classify drugs according to the type of activity shown. This problem of classification, though long examined and used in other disciplines,[3,4] is relatively new to medicinal chemistry[5] and to chemistry in general.[6] It can be handled through various techniques, such as pattern recognition[7,8] and learning machine programs,[9,10] or by traditional statistical methods. The most common and appropriate of the latter is the method of discriminant analysis.[11-13]

Recently, Prakash,[14] Martin,[15] and Dunn[16] have used discriminant analysis for the classification of drugs according to the level of activity, using traditional physicochemical properties as independent variables. As Martin has demonstrated, the case for classification in the two-group example can be handled conveniently by a univariate regression analysis in which the dependent variable assumes discrete values of 0 (or −1) and 1. The problem of multigroup classification must, nevertheless, be handled by multivariate techniques.[17-19]

Cammarata and Menon have reported the use of one multivariate technique, principal component analysis, as a preprocessing step in a study of the clustering of similar compounds according to their structural features, as expressed by molar refraction.[20] They have applied this technique to the pattern recognition of several sets of compounds of diverse biological activities.[21] These authors suggested using the largest principal components of the overall correlation matrix as inputs to a pattern recognition machine. By contrast, Tou has suggested the use of the smallest principal components of the individual within-groups covariance matrices.[22] Other applications which are similar to principal component analysis include the Karhunen–Loeve transform[23] and the SIMCA method.[8,24]

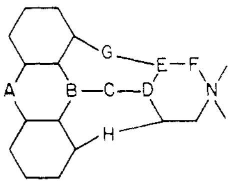
Discriminant analysis resembles principal component analysis in that linear combinations of the original variables are constructed. The goal, however, is not to maximize the overall variance, rather it is to maximize the ratio of the between-groups (hypothesis) variance to the within-groups (error) variance. In general, whenever there is prior information available about the class to which a particular observation belongs a discriminant analysis will be more useful than a principal component analysis for deciding which of the variables are of value in determining between-groups differences and for classification purposes. Thus, one purpose of this report is to extend some of the concepts which Cammarata has introduced, by the application of discriminant analysis to the multigroup classification of a number of the compounds which have already been studied by principal component analysis.

Relative molar refraction was the descriptor index which Cammarata selected for coding various positions on a given structure. In this context, each position becomes a separate variable, and the terms position and variable can be used interchangeably. This is in contrast to some QSAR studies where the descriptor index itself is considered the variable, when measured at one or more positions. For each of the compounds, the various positions were identified by superimposing the structure on a pattern which included all the positions of interest. This pattern is seen in Table I. As Cammarata has pointed out, it is this process of superimposing which is the least objective step in the method, since it requires some judgement on the part of the researcher.[20]

As alternatives to the use of molar refraction, a number of fragment molecular connectivity indices were selected as descriptors.[25] Molecular connectivity indices, whether applied to whole molecules or to structural fragments, possess a number of useful qualities which molar refraction values lack.[26] Foremost is the fact that molecular connectivity values can be calculated unambiguously, given the structure or the connectivity (adjacency) matrix of the molecule. A Fortran program is available for this purpose.[27] Also, molecular connectivity values can uniquely describe positions on a molecule which may show only subtle differences between one another, as in predominantly aliphatic systems. By comparison, the molar refraction value of an aliphatic carbon atom is usually considered constant, regardless of whether the atom is

0022-2623/79/1822-0465$01.00/0

Table I. Raw MC_1 Values for the Compounds Used in the Analyses[a]



		classification[b]			position							
no.	compound	true	lin.[c]	quad.[d]	A	B	C	D	E	F	G	H
1	pyrroliphene	A	A	A	0.000	1.096	0.000	1.274	0.724	0.000	0.707	0.408
2	levoprome	A	H	H	0.408	0.763	0.000	0.724	1.393	0.724	0.000	0.000
3	carbiphene	A	A	A	0.000	0.954	0.687	0.816	0.816	0.816	0.000	0.493
4	fentanyl	A	A	A	0.000	0.705	0.000	0.816	0.908	0.816	0.000	0.781
5	dextromoramide	A	A	A	0.000	1.039	0.000	1.274	0.724	0.000	0.000	0.678
6	methadone	A	P	A	0.000	1.104	0.000	0.854	0.816	0.000	0.000	0.808
7	propoxyphene	A	A	A	0.000	1.096	0.000	1.274	0.724	0.000	0.707	0.408
8	carbamazepine	A	D	A	0.911	0.671	0.716	0.289	0.000	0.000	0.000	0.000
9	profadol	A	A	A	0.000	1.311	0.000	0.854	0.816	0.000	0.000	0.854
10	tilidine	A	A	A	0.000	1.142	0.762	0.742	0.667	0.592	0.000	0.658
11	prodilidine	A	A	A	0.000	1.096	0.000	0.854	0.816	0.000	0.000	0.408
12	imipramine	D	D	D	1.207	0.763	0.000	0.816	1.000	0.816	0.000	0.000
13	protriptylene	D	D	D	0.911	0.986	0.000	0.908	1.000	0.854	0.000	0.000
14	amitriptylene	D	D	D	1.207	0.789	0.000	0.697	0.908	0.816	0.000	0.000
15	nortriptylene	D	D	D	1.207	0.789	0.000	0.697	0.908	0.854	0.000	0.000
16	doxepin	D	D	D	0.846	0.789	0.000	0.697	0.908	0.816	0.000	0.000
17	desipramine	D	D	D	1.207	0.763	0.000	0.816	1.000	0.854	0.000	0.000
18	dimethindene	H	A	A	0.000	0.789	0.000	0.957	0.854	0.816	0.707	1.155
19	methdilazine	H	H	H	0.408	0.763	0.000	0.724	1.225	0.724	0.000	0.000
20	promethazine	H	H	H	0.408	0.763	0.000	0.724	1.244	0.000	0.000	0.000
21	methapyrilene	H	H	H	0.000	0.856	0.000	0.816	0.816	0.000	0.670	0.000
22	ethopropazine	H	H	H	0.408	0.763	0.000	0.724	1.244	0.000	0.000	0.000
23	cyproheptadine	H	D	D	0.911	0.750	0.000	0.957	0.854	0.816	0.000	0.000
24	cyclizine	H	C	H	0.000	0.836	0.000	0.891	0.816	0.816	0.000	0.000
25	diphenhydramine	H	C	H	0.000	0.813	0.000	0.524	0.789	0.816	0.000	0.000
26	tripellenamine	H	H	H	0.000	0.856	0.000	0.816	0.816	0.000	0.577	0.000
27	doxylamine	H	P	P	0.000	1.204	0.000	0.493	0.789	0.816	0.500	0.000
28	trimeprazine	H	H	H	0.408	0.763	0.000	0.724	1.394	0.724	0.000	0.000
29	orphenadrine	C	H	C	0.000	0.818	0.000	0.524	0.789	0.816	0.500	0.000
30	diphenidol	C	P	P	0.000	1.077	0.000	0.844	1.000	0.816	0.224	0.000
31	benztropine	C	C	P	0.000	0.813	0.471	1.052	0.816	1.075	0.000	0.000
32	poldine	C	C	C	0.000	0.974	0.658	0.493	0.697	0.666	0.224	0.000
33	diphemanil	C	H	C	0.000	0.750	0.000	0.957	0.854	0.789	0.000	0.000
34	thiphemanil	C	C	H	0.000	0.866	0.697	0.493	0.789	0.816	0.000	0.000
35	methixene	C	H	C	0.408	0.986	0.000	0.816	0.724	0.000	0.000	0.000
36	piperidolate	C	C	H	0.000	0.866	0.697	0.440	0.724	0.000	0.000	0.000
37	adiphenine	C	C	C	0.000	0.866	0.697	0.493	0.789	0.816	0.000	0.000
38	pentapiperium	C	C	C	0.000	0.911	0.697	0.440	0.908	0.789	0.000	0.000
39	oxyphenonium	C	C	C	0.000	1.012	0.658	0.493	0.789	0.789	0.224	0.000
40	methantheline	C	C	C	0.408	0.866	0.697	0.493	0.789	0.789	0.000	0.000
41	glycopyrrolate	C	C	A	0.000	1.012	0.658	0.440	0.697	0.000	0.224	0.000
42	alverine	C	H	C	0.000	0.000	0.000	0.000	1.000	0.816	0.854	0.854
43	pipenzolate	C	C	C	0.000	0.974	0.658	0.440	0.697	0.000	0.224	0.000
44	mepenzolate	C	C	C	0.000	0.974	0.658	0.440	0.697	0.000	0.224	0.000
45	hexocyclium	C	C	C	0.000	1.116	0.670	0.949	0.816	0.789	0.224	0.000
46	tridihexethyl	C	P	C	0.000	1.116	0.000	0.854	0.789	0.000	0.224	0.000
47	isometheptine	P	C	P	0.000	1.289	0.697	0.908	0.908	1.274	0.000	0.000
48	cycrimine	P	P	P	0.000	1.116	0.000	0.854	0.816	0.000	0.224	0.000
49	trihexphenidyl	P	P	P	0.000	1.116	0.000	0.854	0.816	0.000	0.224	0.000
50	procyclidine	P	P	P	0.000	1.116	0.000	0.854	0.816	0.000	0.224	0.000
51	biperidine	P	P	P	0.000	1.077	0.000	0.854	0.816	0.000	0.224	0.000

[a] Positions were assigned by superimposing the structure of the compound on the structure shown.[20] [b] Group codes are: analgesics, A; antidepressants, D; antihistamines, H; anticholinergics, C; antiparkinsonians, P. [c] Linear classification performed using position variables *a*, *b*, *c*, and *h*. The BMDP7M program was used, and the results shown were obtained using the Lachenbruch or jackknife holdout procedure. [d] Quadratic classification results obtained using the first three canonical discriminant functions; the program used was MULTDIS.

primary, secondary, or tertiary. Finally, molecular connectivity values have been shown to correlate well with many of the more common physical and chemical properties which have been used in QSAR studies[28] and with biological activities as well.[29] Thus, in addition to determining the value of discriminant analysis in the multigroup classification of drugs, a second purpose of this report is to compare the performance of molar refraction with several fragment molecular connectivity indices when applied to this problem.

Experimental Section

Calculations were performed on the Oregon State University CDC Cyber 70 Model 73-16 computer. Software which was used included SPSS and SPSSONLINE discriminant analysis programs,[30] the University of California BMDP7M discriminant analysis

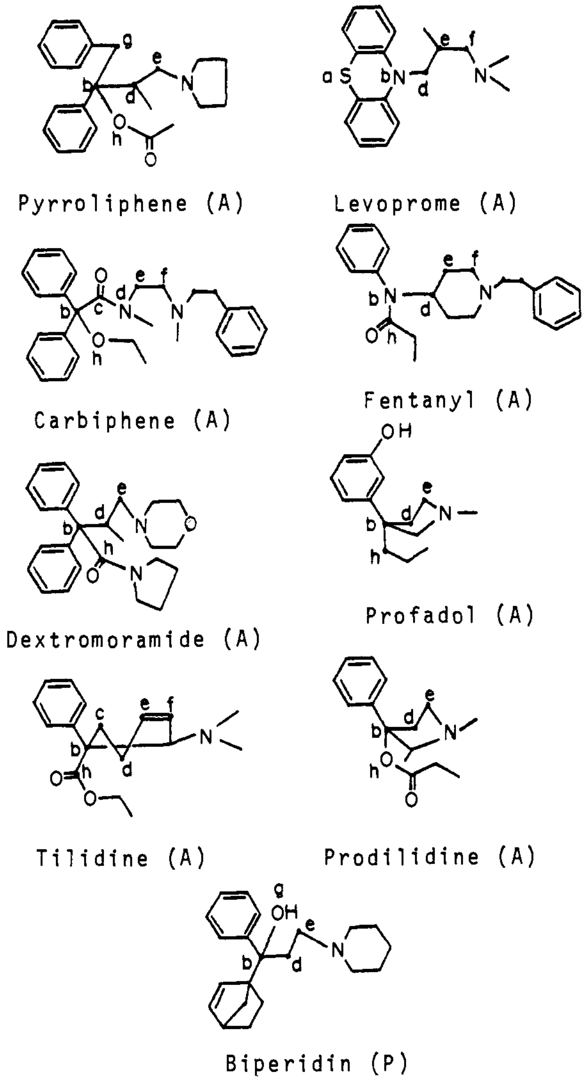


Figure 1. Structures of the compounds added to the design set of Cammarata.[20] The lower-case letters indicate the position assignments relative to the pattern structure of Table I. All the compounds are analgesics except biperidin, which is an antiparkinsonian.

program,[31,32] and the University of Wisconsin MULTDIS program.[11] Also, use was made of the Oregon State Statistical Interactive Programming System (SIPS)[33] and the Oregon State Conventional Aid to Research (OSCAR).[34]

Coding the Molecules. In one paper, Cammarata studied 43 compounds falling into six categories.[20] Of these, 42 were selected, omitting the single antipsychotic drug promazine. To give a larger selection of compounds, an additional eight analgesics and one antiparkinsonian drug were selected from a text.[35] In this way, a total of 51 compounds representing five relatively well-defined classes were obtained, for an average of ten compounds per class. The position assignments were made in accordance with Cammarata's designations. In the case of the added compounds, the positions were assigned by visually superimposing planar representations of the structures on the pattern shown in Table I. In most cases, the assignments were not ambiguous, and in those cases where some question existed, assignment was made by analogy to similar compounds in Cammarata's data set. Figure 1 shows the structures and the position assignments of the additional compounds. The relative molar refraction values were used as defined by Cammarata.[20] As additional descriptor indices, four fragment molecular connectivity indices were adopted for use in this study.

(1) **Vertex valence (δ^v or Δ) values** were included as the most basic numerical index of the connectedness of a given position. They were calculated from the formula given by Kier.[36] By convention for this study, those positions coded with zero for molar refraction, indicating the absence of a substituent at the position, were coded zero in Δ as well. Another convention was to sum the Δ values in those cases where a subgroup containing more than one atom occupied a given position. Thus, a disubstituted ethylene group –C=C– in which each carbon has $\Delta = 3$ would be assigned an overall Δ value of 6, which makes it electronically equivalent to an oxy group.

Table II. Correlation Matrix Showing Pairwise Comparisons of the Descriptor Indices[a]

	MR	Δ	MC_0	MC_1	MC_n
MR	1.0	0.533	0.731	0.403	0.165
Δ		1.0	0.585	0.681	0.192
MC_0			1.0	0.344	0.207
MC_1				1.0	0.140
MC_n					1.0

[a] The standardized values at each of the eight positions on each of the 51 compounds were compared to the corresponding values for the other indices. The number of points for each correlation is 51 × 8 = 408.

(2) **Zero-order molecular connectivity ($^0\chi^v$ or MC_0) values** were calculated as $(\Delta^{-1/2})$. The same Δ values as calculated above were used, and again the convention of zero was adopted for missing substituents. Although zero has physical significance and is a computational possibility for Δ values, a zero for any higher order molecular connectivity term is undefined. In addition, zero represents a break in the continuum of molecular connectivity values; as one progresses from lower to higher levels of connectedness, as represented by Δ, the higher order molecular connectivity terms approach 0. This value then implies infinite connectedness and not the absence of connectedness. This is of no consequence for classification problems, but it could have significance in other types of structure–activity relationship studies. An alternative to using zero for missing substituents would be to leave the entry in the data table blank, resulting in an incomplete data vector. The use of incomplete observations in discriminant analysis has been discussed by Kittler.[37] Most available discriminant analysis packages do not have the capability of dealing with missing data, so it was decided to use zero entries instead.

(3) **First-order molecular connectivity ($^1\chi^v$ or MC_1) values** can be defined for each bond in a structure joining atoms i and j as $(\Delta_i{\cdot}\Delta_j)^{-1/2}$. For this report, the MC_1 value at a given position was defined to be the sum of the MC_1 values of each of the bonds joining the position. In cases of multiple atoms being considered as a single position, the MC_1 term was simply the sum of the MC_1 terms for each bond included in the substructure, with no bond being counted more than once. The raw MC_1 values for the compounds used in this study are shown in Table I.

(4) **Higher-order molecular connectivity ($^n\chi^v$ or MC_n) terms** were calculated as the inverse root of the product of all the Δ values at, and immediately adjacent to, a given position, that is, all the Δ values used in the calculation of the MC_1 term: $MC_n = [(\Delta_0)(\Delta_1)...(\Delta_n)]^{-1/2}$. In this application, the order of a higher-order term clearly depends on the number of adjacent vertices. All higher order terms were grouped as a single variable, and no attempt was made to distinguish among orders or to separate the various types of higher order terms (e.g., path, cluster, etc.), although this can be done.[24]

An initial examination of the raw data showed that molecular connectivity indices in general appeared to differentiate better among the various positions on the molecules than did molar refraction. It was also noted that the values of MC_n were much smaller in magnitude than those of the other indices. Consequently, following the recommendation of VandeGeer,[38] each of the variables (i.e., positions) was standardized to a mean of 0 and a variance of 1. This process, which is sometimes called autoscaling, will not alter any of the classification results, since the ordinal postions of the points representing the observations do not change with respect to each other.

A correlation matrix showing the relationships among the various descriptor indices is seen in Table II. The coefficients in this table are not large, indicating that, although the indices may contain much the same information about the structures, the manner in which this information is expressed varies from

Table III. Positions Which Showed Constant Values (i.e., Zero Variance) Within a Given Group

	index			
group	MR	Δ and MC_0	MC_1	MC_n
analgesics				
antidepressants	*c*, *e*, *f*, *g*, *h*	*c*, *e*, *f*, *g*, *h*	*c*, *h*	*g*, *h*
antihistamines	*c*	*c*	*c*	*c*
anticholinergics	*e*	*e*		
antiparkinsonians	*a*, *d*, *e*, *h*	*a*, *b*, *d*, *e*, *h*	*b*, *h*	*a*, *h*

Table IV. *F* Values for the Decrease in Significance from the Full Eight-Variable Model[a]

index	best 2	best 3	best 4	best 5
MR	8.3 (*a*, *h*)	3.2 (*a*, *c*, *h*)	1.4 (*a*, *c*, *e*, *h*)	0.6 (*a*, *c*, *d*, *e*, *h*)
Δ	11.7 (*c*, *h*)	4.3 (*c*, *g*, *h*)	1.5 (*b*, *c*, *g*, *h*)	0.8 (*b*, *c*, *d*, *g*, *h*)
MC_0	13.2 (*a*, *h*)	5.0 (*a*, *d*, *h*)	2.1 (*a*, *d*, *e*, *h*)	0.9 (*a*, *c*, *d*, *e*, *h*)
MC_1	7.8 (*a*, *h*)	2.9 (*a*, *c*, *h*)	0.9 (*a*, *b*, *c*, *h*)	0.4 (*a*, *b*, *c*, *f*, *h*)
MC_n	10.1 (*a*, *d*)	4.0 (*a*, *d*, *h*)	1.8 (*a*, *d*, *e*, *h*)	1.0 (*a*, *d*, *e*, *g*, *h*)
df	8, 78	12, 104	16, 121	20, 132
$F_{0.01}$	2.75	2.36	2.15	2.02

[a] A large *F* means a large decrease in significance. Variables selected are in parentheses. The critical $F_{0.01}$ values are given for comparison.

one index to the next. As is likely to be the case when only a limited number of compounds are studied, it was found that some of the variables within a given group showed zero variance, which is a clear violation of multivariate normal assumptions (Table III). Rather than drop these variables at the beginning, which would have been the most conservative approach, it was decided to keep all the variables for the analyses and to rely on stepwise selection procedures to remove the variables which did not contribute to between-groups differences. In one sense, a constant variable within a given group can be viewed as a perfect descriptor, if it varies from group to group.

Results and Discussion

Selection of Significant Variables. Most discriminant analysis procedures have stepwise options for including the most significant variable, and/or dropping the least significant one, from a given set of variables. This is necessary, since the inclusion of a sufficient number of variables will guarantee good classification results, even when no true differences exist among the groups. An observation/variable ratio of 10 has been suggested by one author,[39] while others have recommended values ranging from 2 to 20.[12]

A complete stepwise selection procedure was performed for each of the indices, and every possible combination of variables at a given subset level was examined to find the combination of variables with the highest *F* ratio derived from Wilk's λ. This all-subsets option is a part of the MULTDIS program,[11] but it was found that, with very few exceptions, the best subsets selected by this procedure were the same as the subsets that selected one variable at a time by the SPSS or BMDP programs. At each subset level it was possible to test for the loss of information from the full eight-variable model. The results are summarized in Table IV. There is clearly no loss of information in dropping from eight variables to five. The loss at the four-variable level is marginal, but it becomes significant at three variables. Consequently, it was decided to use the best four-variable subset in the discriminant analysis classification procedures as a basis for comparing the descriptor indices.

Table V. Variable Selection Table Showing the Five "Best" Five-Variable Models with the Associated *F* Values[a]

	variable								
index	*a*	*b*	*c*	*d*	*e*	*f*	*g*	*h*	$F_{20,140}$
MR	X		X	X	X			X	6.01
	X		X		X	X		X	5.88
	X	X	X		X			X	5.87
	X		X		X		X	X	5.49
	X		X	X		X		X	5.49
Δ		X	X	X			X	X	9.26
	X	X	X				X	X	9.01
		X	X		X		X	X	8.98
		X	X			X	X	X	8.86
	X	X	X	X				X	8.07
MC_0	X		X	X	X			X	8.45
	X	X		X	X			X	7.89
	X			X	X		X	X	7.69
	X	X	X		X			X	7.54
	X			X	X	X		X	7.24
MC_1	X	X	X			X		X	9.18
	X	X	X				X	X	8.67
	X	X	X	X				X	8.57
	X	X	X		X			X	8.56
	X		X	X		X		X	8.49
MC_n	X			X	X		X	X	5.81
	X		X	X	X			X	5.68
	X	X		X	X			X	5.48
	X			X	X	X		X	5.45
	X	X	X				X	X	5.27

[a] Note that there are many nearly equivalent models for each index, but selection of the four most significant variables is easy.

Just as is the case in regression analysis, it is often found in discriminant analysis that a number of combinations of a given set of variables will show equal or nearly equal levels of significance. This is illustrated in Table V, which shows the five best five-variable subsets for each of the indices. In each case, it is seen that there is virtually no choice to be made among the various subsets. However, it is easy to identify the four most significant variables as being those which occur most frequently in the table. In each case, the best four-variable subset selected in this manner corresponds exactly to the subset selected by the stepwise procedure. All this points to the validity of the stepwise selection procedures in the SPSS and BMDP discriminant analysis programs.

Classification by Discriminant Analysis. The concept of a discriminant function as the linear combination of a set of variables which best differentiates between two groups was first developed by Fisher.[40] Classification into one or the other of these groups can be performed on the basis of the value the discriminant function assumes for a given observation. In the case of several groups, classification can be made on the basis of so-called discriminant scores. This is termed linear classification, and the assumption is made that the within-groups dispersion is the same for all the groups. This is the type of classification procedure that is used in most discriminant analysis programs.

If the within-groups covariance matrices are not assumed to be equal, a quadratic classification procedure will often produce better results. Some discriminant analysis programs, such as the MULTDIS and SAS programs,[41] can perform true quadratic classification. Simply including all squared and cross-product terms in a linear discriminant function, although it results in an expression which resembles the quadratic discriminant function, will not give true quadratic classification results. There are, however, a class of so-called "quadric" discriminant functions which can be obtained by the input of squared and cross-product

Table VI. Linear Test-Space Classification Results[a]

		% correct	
index	variables	ordinary	Lachenbruch
MR	*a, c, e, h*	52.9	49.0
Δ	*b, c, e, h*	64.7	58.8
MC_0	*a, d, e, h*	68.6	60.8
MC_1	*a, b, c, h*	78.4	70.6
MC_n	*a, d, e, h*	68.6	64.7

[a] Results were obtained in the space of the best four-variable subset for each of the indices. The position variables used are shown for each index. Results were obtained using standardized variables with BMDP7M.

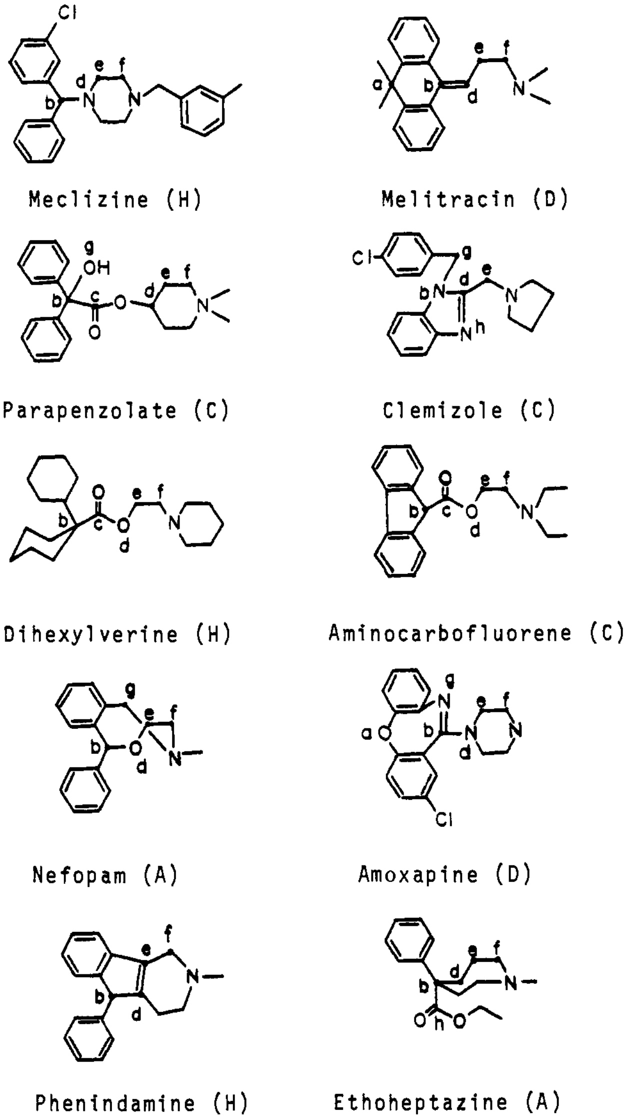
terms into a linear learning machine.[10] These functions are obtained by nonparametric pattern-recognition methods, and they do not have any direct relationship to the quadratic discriminant function.

Test Space Classification. When the classification process is performed in the space of the original variables, it is termed test space classification. The BMDP7M and BMDO7M programs perform this type of classification. Table VI shows the linear classification results obtained using the best four-variable subset for each of the indices. These results were obtained using the BMDP7M program, which has the option of classifying according to the Lachenbruch holdout or "jackknife" procedure.[42] This allows the classification of each observation on the basis of discriminant scores calculated using the other (N-1) observations. This reduces the bias caused by classifying an observation according to rules which have been derived from that observation, and the method seems to be the best compromise for small sample sizes.

The results in Table VI were obtained by assuming equal prior probabilities for each of the groups. Specification of prior probabilities is necessary, since the actual classification procedure follows a Bayes rule.[3] It was found in each case examined that the use of equal priors gave classification results which were as good as, or better than, those obtained using prior probabilities based on observed group membership, which is another commonly used technique. The use of equal priors reduces the linear classification problem to a simple minimum distance classifier. The results in Table VI indicate that the MC_1 index is capable of differentiating better among the various therapeutic classes than are the other indices.

Using the MC_1 index as an example, it was found in general that a large amount of overlap existed between the anticholinergics and the antihistamines, while the antidepressants and the antiparkinsonian drugs, because of the homogeneity in their structures, tended to be classified correctly. Some specific examples of misclassification could easily be rationalized, such as the classification of cyproheptadiene, a tricyclic antihistamine, as an antidepressant or the classification of the phenothiazine analgesic, levoprome, as an antihistamine. The specific linear classification results are shown in the third column of Table I.

Piecewise Linear Classification. One can take advantage of prior knowledge about group membership to construct what amounts to several short discriminant function segments in the immediate neighborhood of each observation. These have been termed piecewise linear discriminant functions.[10] A common pattern-recognition technique which simulates piecewise linear functions is the *K* nearest-neighbor (KNN) method.[43] This method, which can easily by implemented for sample sizes up to a few hundred, classifies an observation according to the class to which a majority of its *K* nearest neighbors in space belong. The value of *K* may range from 1 to any number, but values below 10 are commonly used.



Meclizine (H)

Melitracin (D)

Parapenzolate (C)

Clemizole (C)

Dihexylverine (H)

Aminocarbofluorene (C)

Nefopam (A)

Amoxapine (D)

Phenindamine (H)

Ethoheptazine (A)

Figure 2. Structures of the compounds in the test set.

In practice, the simplest approach to KNN classification is to prepare a distance matrix which contains the Euclidean distance between each point and every other point.[7] For each observation, this matrix is scanned to find the *K* nearest neighbors to that observation. A "vote" is taken among the neighbors, and the observation in question is classified according to the majority outcome. In the case of a tie, the observation can be classified into the group showing the smallest aggregate distance from the observation.

It was felt that it would be interesting to compare the five descriptor indices with each other, with respect to performance in a simple KNN classification. Accordingly, for each of the descriptor indices, the 51 compounds were subjected to two KNN classifications. First, results were obtained using all eight of the position variables. Then, similar results were generated using the best four-variable subset used previously for the discriminant analyses. For all eight variables, the classification results were virtually the same for each of the indices, for values of *K* up to 10. Between 60 and 70% were correctly classified in each case. When only the four most discriminating variables were used, the MC_1 index was clearly superior to the others, in terms of correct classifications. This is shown in Figure 3, which plots the percent correct classification vs. the number of nearest neighbors used in the analysis.

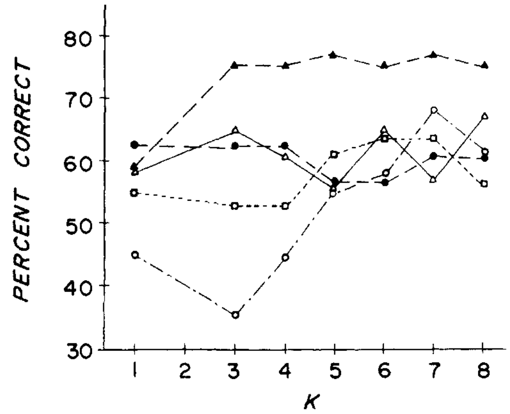


Figure 3. K nearest-neighbor classification results for each of the indices. The best four-variable subset was used in a simple KNN classification process using the values of K indicated. The results for $K = 2$ are, by definition, the same as those for $K = 1$. The indices used were molar refraction (●), Δ (□), MC_0 (○), MC_1 (▲), and MC_n (△). For this set of compounds, the MC_1 index uniformly gives better classification results for any values of K greater than 1.

Reduced Space Classification. An alternative interpretation of the term discriminant function can be derived from the eigenvalues and eigenvectors of the matrix formed by dividing the between-groups sum of squares matrix **B** by the within-groups sum of squares matrix **W**. The eigenvalues of this $\mathbf{W}^{-1}\mathbf{B}$ matrix, like those derived in a principal component analysis, account for the variance represented by the sum of the diagonal elements, or trace, of $\mathbf{W}^{-1}\mathbf{B}$. The elements of the associated eigenvectors, when appropriately normalized, are considered the coefficients of another form of discriminant function. In fact, they define the linear combinations of the variables which best differentiate among all the groups, rather than just a given pair of groups. Using standardized variables, it is found that the magnitude of the coefficient of a given variable in such a linear combination is proportional to the contribution of the variable to the between-groups variation.

Some authors prefer to term these linear combinations the canonical discriminant functions.[32] For k groups and p variables, a maximum of $k - 1$ or p functions can be derived, whichever is smaller. The canonical discriminant functions are most useful for reducing the dimensionality of the data and for graphical presentation. The space of the canonical discriminant functions is termed the reduced discriminant space, and classification can be performed in this space, just as in the space of the original variables. Using all the canonical discriminant functions gives linear classification results which are identical to the results obtained in the test space.[17] It is often found that not all the canonical functions are statistically significant. Reducing the number of functions will sometimes result in poorer classification results, but it reduces the computations necessary and it makes the points easier to visualize.

Unlike the BMDP7M program, which performs linear classification in the test space, the SPSS discriminant program performs classification in the reduced space of the canonical discriminant functions. Table VII summarizes the canonical discriminant functions that were derived for the MC_1 index, using the variables a, b, c, and h. Since the variables were initially standardized, the raw and the standardized discriminant function coefficients were the same.

For the MC_1 index, the first three canonical discriminant functions account for over 96% of the between-groups variation in the data. A plot of the compounds of Table I in the space of these canonical discriminant functions is seen in Figure 4. Approximate dispersions of the groups are also shown. The group centroids lie at the centers of these dispersions. For linear classification, the boundary between any two groups is simply the plane perpendicular to the line segment joining the centroids of the two groups. Points on one side of the plane are classified into one group and those on the other side into the second group. For each of the five descriptor indices, it was found that the first three canonical discriminant functions accounted for over 95% of the between-groups variance. Ordinary linear classification in reduced three-space gave the classification results shown in the first column of Table VIII.

Table VII. Derivation of the Canonical Discriminant Functions Which Maximize $\mathbf{W}^{-1}\mathbf{B}$ for the MC_1 Index

	function			
	1	2	3	4
variable				
a	2.098	−0.161	−0.511	0.237
b	0.425	−0.563	0.124	0.977
c	−0.393	−0.192	−1.114	0.060
h	0.609	−1.270	−0.205	−0.316
eigenvalue	3.534	0.843	0.276	0.130
% of trace of $\mathbf{W}^{-1}\mathbf{B}$	73.9	17.6	5.8	2.7
Wilk's λ	0.083	0.376	0.694	0.885
% signif	0.0	0.0	0.2	1.8
group centroids in reduced space				
analgesics	0.356	−1.650	−0.004	−0.046
antidepressants	4.439	0.664	−0.361	0.158
antihistamines	0.061	0.527	0.707	−0.384
anticholinergics	−1.447	0.403	−0.520	−0.069
antiparkinsonians	−1.103	0.252	0.548	0.826

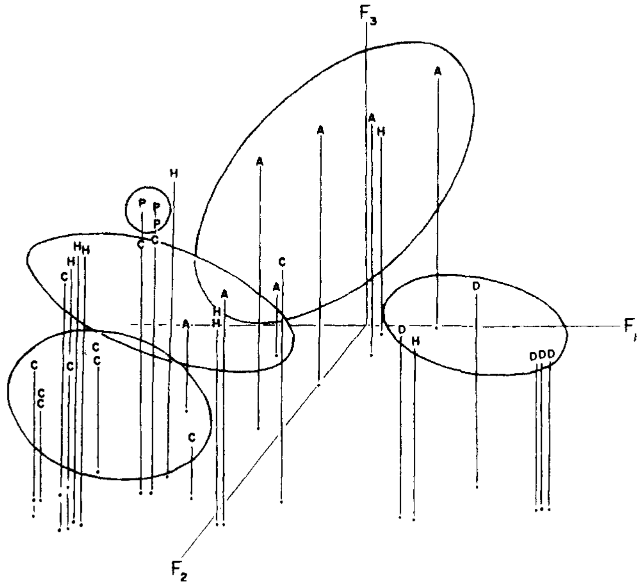


Figure 4. Plot of the design set compounds in the space of the first three canonical discriminant functions of the variables a, b, c, and h using the MC_1 index. Multiple observations at a single point are not shown, and the enclosed regions are only approximations of the within-groups dispersions. The groups are analgesics (A), antidepressants (D), antihistamines (H), anticholinergics (C), and antiparkinsonians (P).

Quadratic Classification. An examination of Figure 4 suggests that the various therapeutic classes, at least for the MC_1 index, are not dispersed to the same degree in space; that is, their within-groups covariance matrices are not equal. In such a case, the optimal classification procedure is by a quadratic discriminant function.[12] This requires the inversion of the within-groups matrices. For

Table VIII. Reduced Space Classification Results Using the Three Most Significant Discriminant Functions[a]

index	% correct	
	linear	quadratic
MR	52.9	60.8
Δ	66.7	76.4
MC_0	64.7	74.5
MC_1	70.6	82.4
MC_n	66.7	70.5

[a] Linear results were obtained using the SPSS discriminant program; quadratic results were obtained using MULTDIS.

each of the indices in this study, using the original four-variable subsets, there was at least one therapeutic category for which the inverse of the covariance matrix was not defined, due to zero variance of one or more of the variables (Table III). Consequently, quadratic classification could not be performed in the space of the original variables.

It was found, however, that nonzero covariance matrices could be obtained in the space of the first three canonical discriminant functions. When quadratic classification was performed in this reduced space, using the MULTDIS program, the results shown in the second column of Table VIII were obtained. As expected, there is improvement in the classification results for each of the indices. The specific classification results for the MC_1 index are shown in the fourth column of Table I. The best results were obtained using the MC_1 index, which is consistent with all the other classification results.

Validation of the Results. It is common practice in classification problems to validate the classification equations or the discriminant functions by classifying a test set of observations. This test set may be a subset of the original design or learning set or it may include completely new compounds. It was decided to use both of these approaches in validating the quadratic reduced space classification procedure for the MC_1 index, since this was the classification method which gave the best results using the design set of compounds.

Random sampling of all the compounds in the design set was performed using a random-number generator. A sample set of ten observations was selected, without replacement, from the full set of 51 compounds. The sample so obtained was subjected to the same quadratic classifier which generated the results in Table VIII. This procedure was repeated five times. This gave classification results of 7/10, 8/10, 8/10, 6/10, and 9/10 correct, which average to 76%. This is comparable to the quadratic result noted in Table VIII for the MC_1 index.

A test set of ten compounds not included in the original analysis was selected. The structures and position assignments of these compounds are shown in Figure 2. A quadratic classification using the first three discriminant functions of the MC_1 values at positions *a*, *b*, *c*, and *h* gave the results shown in Table IX. It is seen that only six of the ten were correctly classified, for the misclassified compounds, the second most likely group was the correct group in only one of the four instances. These results indicate that the quadratic classifier which was used, although it was the best for the design set of compounds, was not necessarily optimal for more general use.

Conclusions. It is clear that discriminant analysis can be used for the multigroup classification of drugs into therapeutic categories, based on structural features. Furthermore, fragment molecular connectivity values, and in particular first-order molecular connectivity terms, have been shown to perform better as descriptors of molecular features when applied to a particular classification problem than relative molar refraction values. Other results, obtained with steroids and phenylethylamines, have confirmed this conclusion.[44] Indeed, it is likely that classification problems may well be one of the more appropriate uses of molecular connectivity indices, since the values are derived directly from the structures of the molecules they represent.

Table IX. Quadratic Reduced Space Classification Results for Test Compounds[a]

compound[b]	true class	probable class[c]				
		highest	⟶			lowest
meclizine	H	H	C	P	A	D
melitracin	D	D	H	A	P	C
parapenzolate	C	C	P	A	H	D
clemizole	C	H*	C	P	A	D
dihexylverine	H	C*	P	A	H	D
aminocarbofluorene	C	C	P	A	D	H
nefopam	A	H*	C	P	A	D
amoxapine	D	D	H	A	P	C
ethoheptazine	A	C*	P	A	H	D
phenindamine	H	H	C	P	A	D

[a] The first three canonical discriminant functions of the positions *a*, *b*, *c*, and *h*, using the MC_1 index, were used. Equal prior probabilities of group membership were assumed. [b] Structures are shown in Figure 2; group codes are as before. [c] Asterisks indicate misclassifications.

Some shortcomings of the approach used in this report should be addressed. Foremost, of course, is the method of assigning the variable positions on the molecules. Although some data handling programs, like the PROPHET system,[45] are capable of accepting and manipulating structural, rather than numerical, input, such systems are not yet widely available. The classical approaches to quantitative medicinal chemistry have evolved around decision-theoretic processes in which the variables have always been assigned, either directly or indirectly, by the researcher. The problem of superimposing a chemical structure on a given pattern is part of the essence of medicinal chemistry. The eventual solution of this problem by computer methods is far from trivial; it may in fact require the use of so-called syntactic methods of pattern recongnition.[46] Until this problem can be efficiently resolved it is likely that the experience and intuition of the researcher will have to guide the assignment of positions on the molecules. No attempt was made in the course of this work to change or to optimize the initial assignments once they were made.

Of course, an alternative to assigning specific position variables would be to use whole-molecule molecular connectivity values as variables. This approach was not used here, since it was desired to compare the discriminating power of molecular connectivity indices directly with that of molar refraction values. The use of whole-molecule values would allow consideration of higher order path, cluster, and chain terms as separate variables. This would give a wide variety of values with which to work. Since the cost of calculating higher order connectivity terms increases with order, these terms would be logical candidates for sequential pattern-recognition techniques in which the ease or cost of measurement of the variable is considered, as well as the information it contains.[47]

One characteristic which parametric and nonparametric classification procedures share alike, especially if the design set is small, is a strong dependence on the data. This is evidenced by the relatively poor performance of the quadratic classifier on a completely new test set of compounds. This does not reduce the validity of the method, especially with regards to comparing the indices as to discriminatory power. As the design set grows in size, provided the assumption holds that there exists a systematic relationship between the therapeutic category and the variables being used, classification results for test compounds would be expected to improve.

The compounds in this study, though they spanned a wide variety of therapeutic classes, were fairly similar in structure; most were diphenylmethyl-based compounds. Also, the therapeutic categories were highly simplified, and much overlap existed. Although it was clear that some of this overlap was reflected in the classification results, no systematic study of this phenomenon was performed. Such a study could serve as further validation of any classification technique, especially if the calculated order of classification (e.g., Table IX) matched to some extent the observed order of therapeutic usefulness. Similarly, a potential exists for identifying possible side effects, provided such side effects are related to the structural variables selected.

Finally, the use of a statistical classification procedure like discriminant analysis invites a comparison with nonstatistical method of pattern recognition. The single nonstatistical method used in this report, the KNN analysis, produced classification results which were not very different from those obtained using discriminant analysis. This may not be true for other data sets, or even for other pattern-recognition techniques when applied to the compounds studied here. It does seem that the use of molecular connectivity indices in pattern-recognition research is worthy of greater attention.

Acknowledgment. Computer time was furnished from unsponsored research funds supplied by the Office of the Dean of Research.

References and Notes

(1) Y. C. Martin, "Quantitative Drug Design, A Critical Introduction", Marcel Dekker, New York, 1978.
(2) W. P. Purcell, G. E. Bass, and J. M. Clayton, "Strategy of Drug Design—A Guide to Biological Activity", Wiley-Interscience, New York, 1973.
(3) C. R. Rao, "Advanced Statistical Methods in Biometric Research", Hafner Press, Darien, Conn., 1952.
(4) W. R. Atchley and E. H. Bryant, "Multivariate Statistical Methods Among Groups Covariation", Halsted Press, New York, 1975, p 108 ff.
(5) A. J. Stuper and P. C. Jurs, *J. Pharm. Sci.*, **67**, 745 (1978).
(6) B. R. Kowalski, *Anal. Chem.*, **47**, 1152A (1975).
(7) J. T. Tou and R. C. Gonzales, "Pattern Recognition Principles", Addison-Wesley, Reading, Mass., 1974.
(8) W. J. Dunn, S. Wold, and Y. C. Martin, *J. Med. Chem.*, **21** 922 (1978).
(9) P. C. Jurs and T. Isenhour, "Chemical Applications of Pattern Recognition", Wiley-Interscience, New York, 1975.
(10) N. A. Nilsson, "Learning Machines", McGraw-Hill, New York, 1965.
(11) R. A. Eisenbeis and R. B. Avery, "Discriminant Analysis and Classification Procedures—Theory and Applications", Heath, New York, 1972.
(12) P. A. Lachenbruch, "Discriminant Analysis", Hafner Press, New York, 1975, and references therein.
(13) T. Cacoullous, Ed., "Discriminant Analysis and Applications", Academic Press, New York, 1973.
(14) G. Prakash and E. Hodnett, *J. Med. Chem.*, 21, 369 (1978).
(15) Y. C. Martin, J. B. Holland, C. H. Jarboe, and N. Plotnikoff, *J. Med. Chem.*, 17, 409 (1974).
(16) W. J. Dunn and M. J. Greenberg, *J. Pharm. Sci.*, **66**, 1416 (1977).
(17) M. Tatsuoka, "Multivariate Analysis: Techniques for Educational and Psychologial Research", Wiley, New York, 1971, pp 157–184.
(18) D. F. Morrison, "Multivariate Statistical Methods", 2nd ed, McGraw-Hill, New York, 1976, pp 230–246.
(19) R. Gnanadesikan, "Methods for Statistical Data Analysis of Multivariate Observations", Wiley, New York, 1977, pp 82–103.
(20) A. Cammarata and G. K. Menon, *J. Med. Chem.*, **19**, 739 (1976).
(21) G. K. Menon and A. Cammarata, *J. Pharm. Sci.*, **66**, 304 (1977).
(22) J. T. Tou, *Pattern Recognition*, *1*, 3 (1968).
(23) M. Loeve, "Probability Theory", Van-Nostrand, Princeton, N.J., 1955.
(24) S. Wold, *Pattern Recognition*, **8**, 127 (1976).
(25) L. B. Kier, L. H. Hall, W. J. Murray, and M. Randic, *J. Pharm. Sci.*, **64**, 1971 (1975).
(26) L. B. Kier and L. H. Hall, "Molecular Connectivity in Chemistry and Drug Research", Academic Press, New York, 1976, pp 16–20.
(27) Program XFUNC, available from L. H. Hall, Eastern Nazarene College, Quincy, Mass. 02170.
(28) W. J. Murray, *J. Pharm. Sci.*, **66**, 1352 (1977).
(29) L. B. Kier and L. H. Hall, *J. Med. Chem.*, **20**, 1631 (1977).
(30) N. H. Nie, C. H. Hull, J. G. Jenkins, K. Steinbrenner, and D. H. Bent, "SPSS—Statistical Package for the Social Sciences", 2nd ed, McGraw-Hill, New York, 1975, pp 434–467.
(31) W. J. Dixon, Ed., BMDP—Biomedical Computer Programs—1977", University of California Press, Berkeley, Calif., 1977, pp 711–733.
(32) R. I. Jennrich, in "Statistical Methods for Digital Computers", Volume III, K. Englein, Ed., Wiley-Interscience, 1977, Chapter 5, pp 76–95.
(33) K. Rowe and JoAnn Barnes, "Statistical Interactive Programming System (SIPS)", Oregon State University Press, Corvallis, OR, 1976.
(34) OSU Computer Center, "Oregon State Conversational Aid to Research (OSCAR)—a Users Manual", Oregon State University Press, Corvallis, OR, 1969.
(35) C. O. Wilson, O. Gisvold, and R. F. Doerge, "Textbook of Organic Medicinal and Pharmaceutical Chemistry", 7th ed, Lippincott, Philadelphia, 1977.
(36) L. B. Kier and L. H. Hall, *Eur. J. Med. Chem.*, **12**, 307 (1977).
(37) J. Kittler, *IEEE Trans. Comput.*, **c-27**, 367 (1978).
(38) S. VandeGeer, "Introduction to Multivariate Analysis for the Social Sciences", W. H. Freeman, San Francisco, Calif., 1971, p 244.
(39) B. Kowalski, *Comput. Chem. Biochem. Res.*, **2**, 1–76 (1974).
(40) R. A. Fisher, *Ann. Eugenics*, **7**, 179 (1936), reprinted in "Machine Recognition of Patterns", A. Agrawala, Ed., IEEE Press, New York, 1977.
(41) Information can be obtained from the SAS Institute, P.O. Box 10522, Raleigh, N.C. 27605.
(42) P. Lachenbruch, *Biometrics*, **23**, 639 (1967).
(43) H. C. Andrews, "Introduction to Mathematical Techniques in Pattern Recognition", Wiley-Interscience, New York, 1972, p 87.
(44) This laboratory, unpublished results.
(45) W. F. Raub, *Fed. Proc., Fed. Am. Soc. Exp. Biol.*, **33**, 2390 (1974).
(46) R. C. Gonzales and M. G. Thomason, "Syntactic Pattern Recognition—An Introduction", Addison-Wesley, Reading, Mass., 1978.
(47) K. S. Fu, "Sequential Methods in Pattern Recognition and Machine Learning", Academic Press, New York, 1968.