

This was partially dissolved in 200 mL of absolute EtOH, treated with 0.5 g of PtO₂, and hydrogenated on the Parr shaker at up to 50 psi for 24 h. During this time the bottle was vented and refilled with hydrogen several times. The catalyst was removed by filtration and the filtrate was taken to dryness in vacuo leaving a foam which failed to crystallize. It was dissolved in warm *i*-PrOH, treated with charcoal, and filtered. The filtrate was acidified with a solution of HCl in *i*-PrOH. Ether was added, precipitating a mixture of gum and solid. This was twice dissolved in warm *i*-PrOH and precipitated by adding ether. The second time, the first material that precipitated was somewhat gummy. This was removed and addition of more ether to the filtrate gave a near white solid which was dried several days in vacuo over P₂O₅ at 56 °C to give **29**, 4.3 g (28%), softening and decomposition at 125–160 °C. Anal. (C₁₇H₂₈N₂O₃·2HCl·H₂O) C, H, N, Cl.

2,3-cis-5-[3-[(1,1-Dimethylethyl)amino]-2-hydroxypropoxy]-1,2,3,4-tetrahydro-8-iodo-2,3-naphthalenediol (25). To a solution of 25.0 g (0.081 mol) of **15** in a mixture of 250 mL of H₂O and 8.5 mL of concentrated hydrochloric acid at 15 °C was added dropwise over 30 min a solution of 13.2 g of ICl in 17.5 mL of concentrated hydrochloric acid. When approximately 80% of the ICl had been added, the solution became cloudy and an oil precipitated from the reaction mixture. The solution was allowed to warm to room temperature and then stirred for an additional 30 min, during which time the oily precipitate dissolved. The solution was made basic by the addition of 25 mL of 50% aqueous NaOH and then extracted with CH₂Cl₂ (5 × 250 mL). Concentration of the CH₂Cl₂ extracts in vacuo gave an oil which crystallized on standing. Recrystallization from CHCl₃ gave 15.8 g (45%) of **25**, mp 140–141 °C. Anal. (C₁₇H₂₆NO₄I) C, H, N, I.

4-(2,3-Epoxypropoxy)-5,6,7,8-tetrahydro-1,2-naphthalenediol (57g). To a well-stirred solution of 26.8 g (0.10 mol) of freshly prepared potassium nitrosodisulfonate (Fremy's salt) in 1.80 L of H₂O and 0.18 L of 1/6 M KH₂PO₄ at 0–5 °C was added a solution of 9.24 g (0.042 mol) of epoxide **57p** in 250 mL of ether. The mixture was stirred at 0–5 °C for 30 min, followed by addition (in one portion) of a solution of 26.8 g (0.10 mol) of Fremy's salt in 1.8 L of H₂O and 0.18 L of 1/6 M KH₂PO₄ precooled to 0–5 °C. The mixture was stirred for an additional 30 min, CHCl₃ was added, and the layers were separated. The aqueous layer was thoroughly extracted with CHCl₃; the combined organic extracts were washed with saturated aqueous NaCl, dried, and concentrated in vacuo to give 9.0 g (85%) of solid red-orange quinone **58**.

A suspension of the above quinone **58** (9.0 g) in 250 mL of EtOAc was hydrogenated in the presence of 1.0 g of 5% Pd/C (Parr shaker). After uptake of 1 equiv of H₂ (10 min), the solution was warmed briefly to dissolve precipitated product, the catalyst

filtered off (Celite), and the filtrate concentrated in vacuo to an off-white solid. Trituration with ether gave 8.0 g (89%) of crystalline epoxide **57g**.

Acknowledgment. We thank our Analytical Department staff for microanalyses and spectra, Dr. Clayton Bennett for some intermediates, and our Pharmacology Department staff for biological data.

References and Notes

- (1) B. N. C. Prichard, *Drugs*, **7**, 55 (1974).
- (2) L. Hansson, *Acta Med. Scand., Suppl.*, **550**, 1 (1973).
- (3) F. D. Simpson, *Drugs*, **7**, 85 (1974).
- (4) J. D. Fitzgerald, *Clin. Pharmacol. Ther.*, **10**, 292 (1969).
- (5) B. N. Singh and D. E. Jewitt, *Drugs*, **7**, 426 (1974).
- (6) J. R. Blinks, *Ann. N.Y. Acad. Sci.*, **139**, 673 (1969).
- (7) C. S. Liang and W. B. Hood, Jr., *Circ. Res.*, **35**, 272 (1974).
- (8) S. A. Stephen, *Am. J. Cardiol.*, **18**, 463 (1966).
- (9) A. F. Crowther, R. Howe, and L. H. Smith, *J. Med. Chem.*, **14**, 511 (1971).
- (10) G. Shtacher, M. Erez, and S. Cohen, *J. Med. Chem.*, **16**, 516 (1973).
- (11) R. J. Lee, D. B. Evans, S. H. Baky, and R. J. Laffan, *Eur. J. Pharmacol.*, **33**, 371 (1975).
- (12) D. B. Evans, M. T. Pescha, R. J. Lee, and R. J. Laffan, *Eur. J. Pharmacol.*, **35**, 17 (1976).
- (13) K. K. Wong, J. Dreyfuss, J. M. Shaw, J. J. Ross, and E. C. Schreiber, *Pharmacologist*, **15**, 245 (1973).
- (14) J. M. Shaw and J. Dreyfuss, *Fed. Proc., Fed. Am. Soc. Exp. Biol.*, **35**, 365 (1976).
- (15) A. J. Shand and R. H. Thompson, *Tetrahedron*, **19**, 1919 (1963).
- (16) J. F. Eastham and D. R. Larkin, *J. Am. Chem. Soc.*, **80**, 2887 (1958).
- (17) "Organic Syntheses", Collect. Vol. IV, Wiley, New York, N.Y., 1963, p 887.
- (18) D. J. Marshall and R. Deghenghi, *Can. J. Chem.*, **47**, 3127 (1969).
- (19) L. F. Fieser, *J. Am. Chem. Soc.*, **70**, 3165 (1948).
- (20) R. B. Woodward and F. V. Brutcher, *J. Am. Chem. Soc.*, **80**, 209 (1958).
- (21) The separation and resolution of the two diastereoisomers composing compound **15** will be the subject of a forthcoming publication.
- (22) A. F. Crowther and L. H. Smith, *J. Med. Chem.*, **11**, 1009 (1968).
- (23) C. F. Schwender, R. E. Pike, and J. Shavel, Jr., *J. Med. Chem.*, **18**, 211 (1975).

Structure-Activity Study of β -Adrenergic Agents Using the SIMCA Method of Pattern Recognition

W. J. Dunn, III,*¹ Svante Wold,

Research Group for Chemometrics, Institute of Chemistry, Umeå University, S-901 87 Umeå, Sweden

and Y. C. Martin

Abbott Laboratories, North Chicago, Illinois 60064. Received December 14, 1977

The SIMCA method of pattern recognition (PaRC) was used to analyze structure-activity data for a series of phenethylamine agonists and antagonists of the β -adrenergic receptor. On the basis of physicochemical substituent parameters the SIMCA method classified correctly 100% of the agonists and 88% of the antagonists. In addition, parameters derived from the class models were correlated with the biological activities of the agonists and antagonists, respectively. Test compounds not included in the initial data analysis were classified and their activities estimated. The applicability of pattern recognition in structure-activity studies in general is discussed.

Since the early reports of Hansch^{2a} and Free and Wilson^{2b} using multivariable regression to systematically analyze biological structure-activity data, a field of research interest has developed around the philosophy of relating structural changes within a class of pharmaco-

logically similar agents to changes in biological activity. The work of Hansch and his co-workers is especially significant in that it has shown that structurally and physicochemically similar substances can behave pharmacologically in regular and predictable manners.

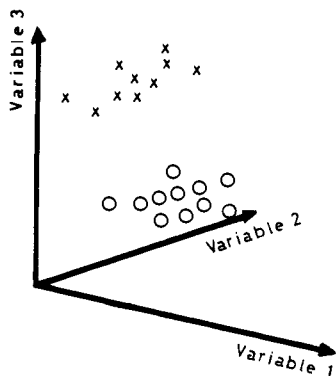
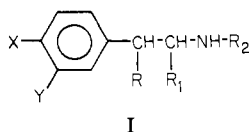


Figure 1. Graphical representation of the classification problem.

In order to treat biological structure–activity data of structurally and physicochemically similar substances which can fall into two or more pharmacological classes, other methods of data analysis are necessary.

Efforts to treat such more complex problems have relied on various methods of classification such as discriminant analysis,³ principal component analysis,^{4,5} and other pattern recognition methods.^{6–10}

This article discusses first the general applicability of pattern recognition methods to problems of structure activity. Second, it reports a specific application of the SIMCA method^{11,12} to the problem of classifying two series of phenethylamine agonists and antagonists of the β -adrenergic receptor using data reported by Lefkowitz et al.¹³ The general structure of these substances is I.



Scope of Classification Studies. In the terminology of classification and pattern recognition, the general scope of classification studies is to find rules for the classification of *objects* (in our case, drugs = chemical compounds) on the basis of the values of a number of variables characterizing these objects. Objects of the training set in a given study have a “known” class assignment and are used for the methodology to “learn” the characteristics of each class. Objects in the *test set* are initially of unknown class assignment. One of the goals of the analysis is to assign these objects to a class on the basis of the patterns found on the training set. In structure–activity studies, the variables are derived from the structure of the compounds, typically the presence of structural units, the length of side chains, etc.

The classification problem has a nice graphical interpretation as shown in Figure 1. The data characterizing each object (k) consist of the values of M variables, i.e., an M -dimensional vector ($y_{i,k}$). Hence, this vector can be represented as a point in an M -dimensional space obtained by giving each variable an orthogonal coordinate axis. The object vectors “known” to belong to one class are hopefully situated in a region of this space that is different from the site of object vectors of other classes. In this way, all classification methods can be seen as ways to mathematically describe the separation of these regions and of finding out in which region objects of the test set are situated.

In most recently published examples of the use of pattern recognition to treat biological structure–activity data, large training sets with large numbers of variables were used. In this example such is not the case. In order to form a basis for pattern recognition as applied in the

present paper, each class reference set should contain at least five to ten objects. Increasing the size of the data set to more than 20 objects in each class reference set usually does not lead to a marked improvement in the resulting information if the additional compounds are sufficiently similar to those in the reference set. The same argument holds for the number of variables characterizing the objects. At least five to ten variables are usually needed to make pattern recognition worthwhile and numbers far above 20 often lead to redundancies. Thus, it is not necessary to use large training sets to get good results. The methodology indicates how much information is contained in a given data set. The important thing is if this information is sufficient to answer the questions posed in the given problem.

Four levels of classification can be recognized.

(1) Classification into either of a number of defined classes, e.g., agonist or antagonist. Such methods as discriminant analysis and the linear learning machine are most often applied at this level. Since the result of such applications is mere classification, which often is trivial for the experienced pharmacologist, criticism of the use of the methods on this level in structure–activity studies has resulted.¹⁴

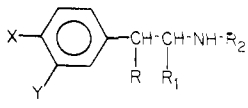
(2) Classification as in (1) above with the possibility that an object is an “outlier”, i.e., that it does not belong to any of the defined classes. In the present application, this would correspond to allowing for a compound to be neither an antagonist nor an agonist. In structure activity, this is usually the lowest realistic ambition level. Thus, the crucial point in the present study is not whether a compound is an agonist or antagonist but whether a new compound structurally similar to the previously investigated compounds is an agonist, antagonist, or *neither*. In M -dimensional space, this corresponds to a containment of each class in a closed structure. If a new object is inside a structure it belongs to the corresponding class; if it is outside all structures it is an “outlier”.

(3) At this level the scope is that of level 2 *plus* the ambition to relate the position of an object within a given class to a measured activity of the object (compound). This is the ambition level of the present investigation. On each of his investigated compounds, Lefkowitz and his co-workers measured agonist potency or antagonist potency and, also for the agonists, intrinsic activity. Thus, we wish to relate the structure of the compounds to their pharmacological class, as agonist or antagonist, and to their level of biological activity within their respective class. Mathematically this corresponds to containment of the objects in each class in a closed mathematical structure and finding a relation between the position within this structure and the measured activity.

(4) On the highest level, finally, several “effect” variables have been measured on each object and one wishes to relate all of these to the position of the objects within the classes. This problem is also common in structure–activity studies. Often several different measurements of the biological activity of each compound have been observed, say, the activity in different test systems such as rat, rabbit, dog, monkey, and man. In addition, the levels of side effects might be available on the same test systems.

Somewhat paradoxically, this problem is often simpler mathematically than the level 3 problem. The reason is that several “effect” variables can be used to define a systematic structure in an “effect space” in the same way as the characterizing variables are used to define a systematic structure in the measurement space discussed above. These two “systematic structures” can then be

Table I. Biological and Physicochemical Data for Agonists and Antagonists



object	X	Y	R	R ₁	R ₂	class	IA ^a	act. ^a	pK _b ^a	pK _a ^b	fPh ^c
1	18	18	OH	2	1	2	0.87	4.39	4.55	8.93	1.14
2	18	18	OH	3	1	2	0.71	4.42	4.74	8.93	1.14
3	18	18	OH	1	2	2	0.75	5.00	5.07	9.29	1.14
4	18	18	OH	1	4	2	1.00	5.85	5.77	9.90	1.14
5	18	18	OH	3	4	2	0.72	4.35	4.62	9.90	1.14
6	18	18	OH	3	5	2	0.64	4.51	4.41	9.93	1.14
7	18	18	OH	1	6	2	1.10	6.33	6.17	9.19	1.14
8	18	18	OH	1	7	2	1.10	6.37	6.17	9.19	1.14
9	18	18	H	1	8	2	0.25	4.68	4.33	10.03	1.14
10	18	18	H	1	9	2	0.25	5.04	4.62	10.29	1.14
11	18	18	OH	1	10	2	1.20	7.10	7.22	9.29	1.14
12	18	18	H	1	11	2	0.17	5.04	4.64	10.22	1.14
13	18	19	OH	1	4	2	0.28	6.00	5.62	9.94	-0.07
14	18	19	OH	1	4	2	0.24	5.48	6.19	9.77	-0.07
15	18	19	OH	1	12	2	0.27	7.10	7.85	9.29	-0.07
16	17	20	OH	1	1	1		3.51	4.08	8.93	2.66
17	17	19	OH	1	2	1		3.66	4.19	9.29	0.55
18	17	18	OH	1	2	1		3.87	4.28	9.29	1.36
19	17	18	OH	1	3	1		4.29	4.66	9.61	1.36
20	20	18	OH	1	4	1		5.89	5.38	9.90	2.04
21	18	17	OH	1	4	1		4.96	4.82	9.90	1.36
22	17	18	OH	2	1	1		4.52	4.46	8.93	1.36
23	20	20	OH	1	4	1		6.40	6.24	9.90	3.34
24	19	17	OH	1	4	1		5.80	5.89	9.90	0.55
25	17	17	OH	16	1	1		3.85	4.29	8.46	1.90
26	17	17	OH	2	2	1		4.07	5.04	9.29	1.90
27	21	17	OH	3	4	1		5.35	4.85	9.90	-0.94
28	18	17	OH	2	8	1		5.74	5.06	9.03	1.36
29	18	17	OH	2	13	1		6.62	5.85	8.16	1.36
30	18	17	OH	2	14	1		6.89	6.74	9.29	1.36
31	18	22	OH	2	13	1		7.22	7.12	8.16	1.04
32	17	23	OH	2	15	1		5.64	5.11	10.26	1.96
33	18	18	OH	1	1	0		4.04	<3.70	8.93	1.14
34	18	17	OH	1	1	0		<3.00	<3.70	8.93	1.36
35	18	17	H	1	1	0		<3.00	<3.70	8.93	1.36
36	18	18	H	1	1	0		<3.00	<3.70	8.93	1.14
37	17	17	H	1	1	0			<3.70	9.80 ^d	1.90

^a Reference 13. ^b Estimated using the procedure of Clark and Perrin.²⁵ ^c Corrected for proximity when required.
^d Pomona College Medicinal Chemistry Data Bank.

related to each other using rather simple mathematical-statistical methods.

Present Study. The data of Lefkowitz and his co-workers contained 37 compounds (objects of general structure I) of which 15 were agonists and 17 were antagonists. The compounds, norepinephrine, tyramine, octopamine, dopamine, and phenethylamine, were also included in the study as a test set. Norepinephrine is a weak agonist while the other four substances are neither agonists nor antagonists.

The biological evaluations were made on racemic mixtures. The affinity of all compounds for the β receptor was measured by the ability of the compound to displace (-)-[³H]alprenolol from its binding site in receptors partially purified from frog erythrocytes. In addition, the ability of the compounds to stimulate (agonists) or to inhibit agonist stimulation (antagonists) of the β -adrenergic coupled adenylate cyclase was also determined. These activities were reported as *K* values in μ M and were converted for this study to the p*K* scale in M units. For agonists both the affinity and the intrinsic activity were reported; the intrinsic activity was on a scale of isoproterenol equal to 1.0. For antagonists, the relative affinity for this reaction was reported.

Describing the Structure by a Set of Variables. A critical factor in classification studies in structure activity,

no matter which method is used, is the appropriate description of the objects being classified. In previously reported pattern recognition (PaRC) studies, various structural descriptors and measured variables were used, e.g., mass spectral data,⁶ molecular transforms,¹⁰ and descriptors which could be generated from a two-dimensional representation of the structures.^{5,7-9} The generation and utility of such descriptors have been discussed.^{15,16} In this report we take a different approach which may be considered an extension of the extrathermodynamic assumptions as used by Hansch and his co-workers. In describing the agonists and antagonists, physicochemical parameters have been used. It is assumed that differences in specific drug-receptor interactions determine the classification outcome. Therefore, physicochemical parameters which best model these interactions were used as the basis of classification. Descriptors relevant to the problem such as Hammett σ constants,¹⁷ hydrophobic constants,^{18,19} and steric constants^{20,21} were utilized. We have also included as a variable the experimentally determined receptor binding constant as reported. The binding constants for the objects in the test set could not be quantified, but upper limits were assigned. The use of these values will therefore introduce uncertainty into the classification of these objects. The objects and their descriptor variables are given in Table I.

The first objective of the present study is to find rules to classify compounds of general structure I as agonists or antagonists or neither. Since this is a classification problem a method of pattern recognition must be used, and a method operating at least on level 2 (see above) must be chosen. Also, since SIMCA is not designed to maximize the separation between the classes, but rather designed to describe each class in an operational sense, the method is not sensitive to a large number of variables as compared to the number of objects.

The second objective of the study is to relate the structural descriptors of the compounds to the measured level of activity. Here a multiple regression would seem appropriate. However, the number of variables (13) is too large to make a multiple regression numerically stable. We instead use the SIMCA method which extracts the systematic data structure in each class and expresses these as variable specific parameters (m_i and b_{ia}) and object specific (u_{ak}) terms. The latter parameters can then be related to the activity by a multiple regression since the u_{ak} parameters are few compared to the initial number of variables and the resulting multiple regression is stable. In conclusion, the SIMCA method is the only method presently developed which can fulfill the scope of the present problem, classification and prediction of the level of activity of the objects.

SIMCA Method. The basis of the pattern recognition method used in this report is that the data of objects in a class can be described by a principal components (PC) model.

$$y_{i,k} = m_i + \sum_{a=1}^A b_{ia} u_{ak} + e_{ik}$$

Here $y_{i,k}$ denotes the value of variable i observed on object k . The parameters of the model, m_i , b_{ia} , and u_{ak} , are estimated as to make the residuals e_{ik} minimal in the least-squares sense over all objects k and all variables i in the class.

Geometrically, this corresponds to an A -dimensional hyperplane in the M -dimensional measurement space which has the closest fit to the data points of the class. The values of the parameters m_i and b_{ia} define the central point and the direction coefficients of the plane, respectively. The values of the parameters u_{ak} define the position on this plane of the k th object in the training set. Thus, the SIMCA method describes the region of each separate class by means of a separate hyperplane and a closed area on this plane upon which all objects in the training set of that class are situated.

In addition, the residual standard deviation (RSD) of the class, s_{aq} , defines a confidence "slab" around the class plane within which an object should be situated to really belong to the class. This is shown in Figure 2.

Thence, SIMCA contains the objects in closed mathematical structures, "hyper-boxes", and thus always operates on at least ambition level 2 in the scheme discussed above. We see that the position within a class is defined by the values of the components u_{ak} . Hence, a problem on ambition level 3 or 4 is approached by SIMCA as finding relations between these parameters and the measured "effect" variables on the same objects, usually by some simple linear model, i.e., multiple regression. This corresponds to the "target rotation" of Weiner and Weiner²² in the closely analogous factor analysis and to a simplified path model of PLS type.²³

SIMCA has a theoretical foundation similar to that of polynomial models, i.e., on the basis of Taylor expansions. Thus, it can be shown that (a) provided that the variables

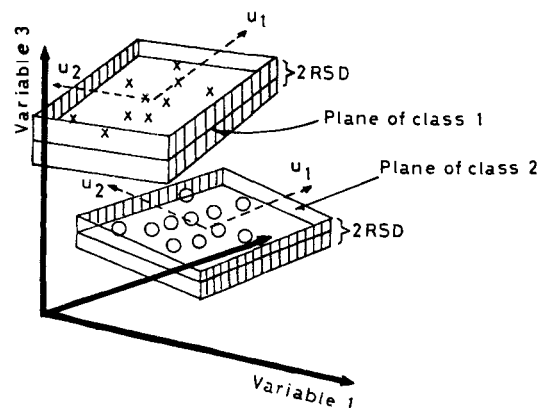


Figure 2. The SIMCA description of each class by means of a plane with RSD and u_i .

characterizing the objects of a class have certain continuity properties and that (b) the objects are realizations of a process with limited variability, that the data of the objects can be arbitrarily closely approximated by the above PC model with a limited number of product terms, A .

Since SIMCA fits the class models to the data of the objects in the training set by least squares, the method does not assume any particular distribution of the data. SIMCA can be considered to be a nonparametric method in the pattern recognition nomenclature.

The data analysis in the SIMCA framework is done as follows.

(1) For each class, q , the number of components, A_q , in the PC model required to describe the data of the objects in each class is determined. This is called determining the rank of the data matrix and is done by cross-validation.²⁴

(2) The parameters m_i^q , b_{ia}^q , and u_{ak}^q for $a = 1, 2, \dots, A_q$ are determined in each class model for the data of the training set. This corresponds to finding the A -dimensional plane of closest fit to the points of the class in M space.

(3) The resulting residuals of each class, e_{ik}^q , are used to assess the importance of each variable i .

Variables can be significant in classification in two ways: (1) in their ability to describe the class structure and (2) their ability to distinguish between classes. Their ability to describe class structure we call modeling power (ψ_i), while their ability to distinguish between classes we call discrimination power (ϕ_i). The modeling power of a variable is obtained by comparing its residual standard deviation, S_i , over all data in the training set with the corresponding standard deviation, $S_{i,y}$, of the training set data y .

$$\psi_i = 1 - S_i/S_{i,y}$$

A value for ψ_i approaching 1 indicates high modeling power while a value near 0 indicates low modeling power. This parameter can also be calculated for a variable over one class by comparing the corresponding within class standard deviations. This measures the modeling power of a variable in defining one single class. The discrimination power of a variable is a measure of the distance between two classes over all variables. This is obtained by comparing the fit of objects of the two classes (q and r) to their own class with that of their fit to the other class. This is given as

$$\phi_i^{q,r} = \left[\frac{(s_{i,r}^q)^2 + (s_{i,q}^r)^2}{(s_{i,r}^r)^2 + (s_{i,q}^q)^2} \right]$$

Here $s_{i,r}^q$ denotes the standard deviation of the residuals

Table II. u_i and RSD for Classes 1 and 2 and Test Set^a

object	u_1	u_2	u_3	RSD	
				class 1	class 2
class 2					
1	3.54	1.22	0.45	0.78	0.18
2	4.04	0.90	-0.79	1.00	0.28
3	0.84	0.79	0.71	0.70	0.42
4	-0.58	0.07	0.40	0.64	0.40
5	1.03	-1.49	-1.53	1.10	0.24
6	0.78	-1.89	-1.48	1.10	0.09
7	-1.29	0.62	-0.10	0.46	0.31
8	-1.18	0.71	-0.10	0.46	0.25
9	0.23	-0.90	1.24	0.90	0.20
10	-1.05	-1.60	0.89	0.90	0.19
11	-2.32	1.01	-0.76	0.63	0.20
12	-0.15	-1.08	1.22	0.98	0.33
13	-0.51	-0.02	0.45	0.63	0.38
14	-0.25	0.67	0.37	0.78	0.51
15	-3.14	0.98	-0.98	0.72	0.39
class 1					
16	3.65	0.66	-1.31	0.21	1.20
17	1.89	0.48	0.98	0.25	1.20
18	1.87	0.45	0.98	0.23	1.10
19	1.33	0.28	1.51	0.15	1.10
20	-1.31	1.43	0.82	0.50	1.30
21	-0.59	-0.30	1.04	0.80	0.34
22	3.11	0.70	-1.19	0.58	1.00
23	-1.52	1.14	0.79	0.74	1.30
24	-3.59	1.46	-0.88	0.89	1.70
25	3.66	0.73	-1.12	0.39	1.10
26	1.45	0.00	0.56	0.62	1.10
27	-4.28	2.46	-1.85	0.70	2.20
28	-0.28	-1.35	-0.64	0.50	0.46
29	-0.73	-2.59	-0.86	0.41	0.79
30	-1.73	-2.57	0.24	0.50	0.58
31	-1.05	-3.02	-0.95	0.50	0.56
32	-1.87	0.02	1.84	0.58	0.79
test set					
33	3.35	1.12	1.86	0.92	0.33
34	3.36	1.15	1.85	0.92	0.35
35	3.36	1.15	1.85	0.92	0.35
36	3.36	1.15	1.85	0.92	0.35
37	3.15	1.75	0.04	0.61	1.00

^a Class 1, RSD = 0.53; class 2, RSD = 0.31.

of the i th variable obtained when the objects in the r th reference set are fit to the q th class model.

The range of ϕ is $\phi \geq 1$ with values larger than 2 indicating good discrimination for that variable. In an analysis, variables with low ψ and ϕ are deleted and the

parameters m , b , and u are reestimated for the reduced data matrix.

(4) Classification of the objects in the test set is done by fitting the data for each object to each class model, q , i.e., the equation below with m_i^q and b_{ia}^q fixed to the values determined in step 2. The index p refers to the p th object.

$$y_{ip} - m_i = \sum_{a=1}^{A_q} t_{ap}^q b_{ia}^q + e_{ip}^q$$

Multiple regression is used with the t_{ap}^q 's being regression coefficients for the fit of the object p to the q class model. The residual standard deviation from each fit is a measure of the orthogonal distance of the object from the classes. Classification as being a member of a specific class or being a member of none of the classes can be made on the basis of this parameter.

Results

In applying SIMCA in this problem, the variables were first regularized to give them equivalent variance and means of zero. This procedure gives equivalent weight to the variables with small variation and to those with large variation and, thence, prevents masking of the variables with little variation by those with large variation.

Classification as Agonists or Antagonists. Our first objective was to find PC models which would separate the β agonists from the antagonists. Analysis with all variables showed that the agonists (class 2) could be well described by a PC model with two components ($A = 2$). The antagonists (class 1) could be described by a one-component model. On the basis of low modeling power and low discriminatory power the variables σ_m , f_{Ph} , B_m , and L_m were deleted from both classes. On the basis of the remaining variables (σ_p , L_p , B_p , f_{R1} , f_{R2} , E_{s-R2} , σ_{R2} , pK_b , and pK_a) PC models with $A = 3$ components sufficiently described both classes. The results are given in Tables II and III. The coefficients b_i of the variables are normalized such that

$$\sum_{i=1}^M b_i^2 = 1$$

On the basis of these models 100% (15/15) classification of the agonists and 88% (15/17) of the antagonists resulted.

Validation of these results was made by omitting every fourth object in each class making these actually a "mini" test set. The parameters in the PC models were estimated

Table III. b_{ia} Values

	pK_b	f_{R1}	f_{R2}	σ_{R2}^*	E_{s-R2}	pK_a	σ_m	σ_p	f_{Ph}	B_p	L_p	B_m	L_m
class 1													
m_i	-0.14	0.02	-0.20	0.07	0.12	-0.26		0.64		0.00	-0.14		
b_{i1}	-0.24	-0.16	-0.28	0.36	0.33	-0.25		-0.02		-0.52	-0.51		
b_{i2}	-0.33	-0.14	-0.42	0.08	0.25	0.40		0.58		0.30	0.14		
b_{i3}	-0.03	-0.28	0.06	-0.44	-0.35	0.56		0.05		-0.34	-0.40		
class 2													
m_i	0.15	-0.03	0.23	-0.08	-0.14	0.29		-0.73		0.00	0.16		
b_{i1}	-0.42	0.41	-0.44	0.45	0.51	-0.10		0.00		0.00	0.00		
b_{i2}	0.61	-0.31	-0.33	0.21	0.16	-0.60		0.00		0.00	0.00		
b_{i3}	-0.33	-0.84	-0.01	0.28	0.20	0.27		0.00		0.00	0.00		
class 3													
m_i	0.10	0.40	-0.13	0.05	0.04	-0.06	0.05		-0.06			-0.18	-0.09
b_{i1}	-0.37	0.42	-0.39	0.49	0.53	-0.10	0.00		0.00			0.00	0.00
b_{i2}	0.32	-0.71	-0.22	0.38	0.19	-0.41	0.00		0.00			0.00	0.00
class 4													
m_i	0.23	-0.67	0.77	-0.29	-0.40	0.81	0.46		-0.82			0.67	0.99
b_{i1}	-0.55	0.00	0.06	0.23	0.10	0.23	-0.19		0.36			-0.39	-0.50
b_{i2}	0.33	0.00	0.70	-0.12	-0.50	-0.12	-0.08		0.16			-0.17	-0.22

Table IV. u_i and RSD for Subclassification of Agonists^a

object	u_1	u_2	RSD	
			class 3	class 4
class 3				
1	3.04	1.25	0.11	1.30
2	3.56	0.31	0.39	1.60
3	0.25	1.00	0.52	0.76
4	-1.15	0.22	0.54	0.58
5	0.43	-2.12	0.15	1.10
6	0.20	-2.37	0.17	1.00
7	-1.79	0.58	0.34	0.59
8	-1.69	0.63	0.24	0.63
11	-2.85	0.50	0.44	0.81
class 4				
16	2.13	-0.24	0.71	0.23
10	-1.87	0.86	0.74	0.30
12	2.06	0.18	0.78	0.19
13	-1.39	-1.19	1.20	0.25
14	-1.67	-1.47	1.20	0.16
15	-3.01	1.86	1.30	0.09

^a Class 3 (strong agonists), RSD = 0.31; class 4 (weak agonists), RSD = 0.21.

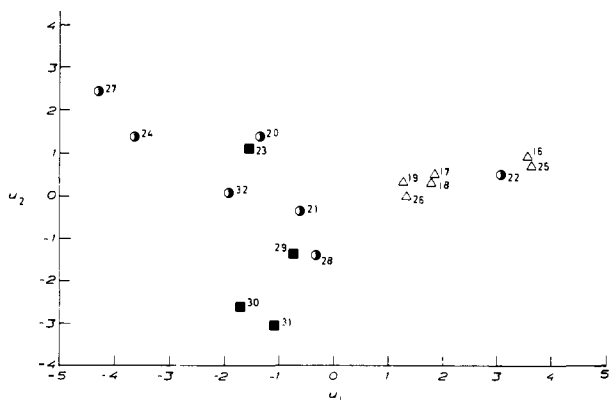


Figure 3. u_i plot for class 1: Δ = low activity, \circ = medium activity, \blacksquare = high activity.

from the reduced training set and this mini test set was then classified. The training set was then restored and another quarter was deleted to form another mini test set and so on until all compounds had been in such a test set one time and one time only. All objects in class 2 were correctly classified when placed in these test sets while objects 28 and 29 were misclassified in the validation of class 1.

Objects 33-37 in the test set were classified as discussed earlier with their RSD given in Table IV. Object 33 is norepinephrine and is correctly classified as an agonist. Objects 34-36, tyramine, octopamine, and dopamine, are also classified as agonists.

Lefkowitz classifies dopamine as an agonist even though it had no detectable intrinsic activity and its affinity was determined as an antagonist. Tyramine and octopamine were both classified as antagonists by Lefkowitz although neither [³H]alprenolol binding nor inhibition of isoproterenol stimulation of adenylate cyclase synthesis could be quantified. Compound 37 is phenethylamine, which had no detectable activity as agonist or antagonist. SIMCA classified it as an antagonist. Thus classification of objects 34-37 in itself is not sufficient to clarify their identity.

Relating the Position in the Class to Level of Activity. One of the basic assumptions in applying SIMCA and other PaRC methods in structure-activity studies is that structurally similar substances will cluster in the measurement space chosen to represent the objects. It, therefore, follows that within the measurement space, those

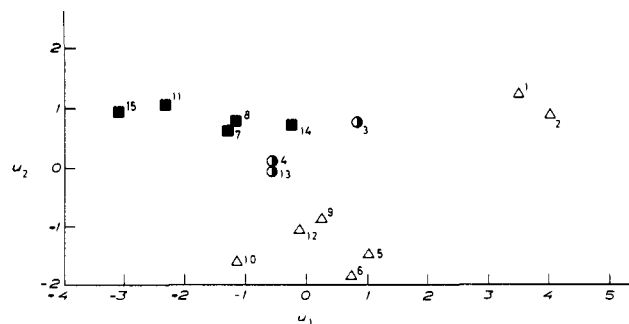


Figure 4. u_i plot for class 2: Δ = low activity, \circ = medium activity, \blacksquare = high activity.

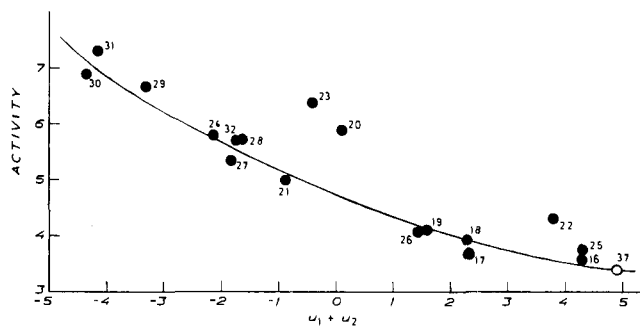


Figure 5. Activity estimation for class 1: \circ = test set.

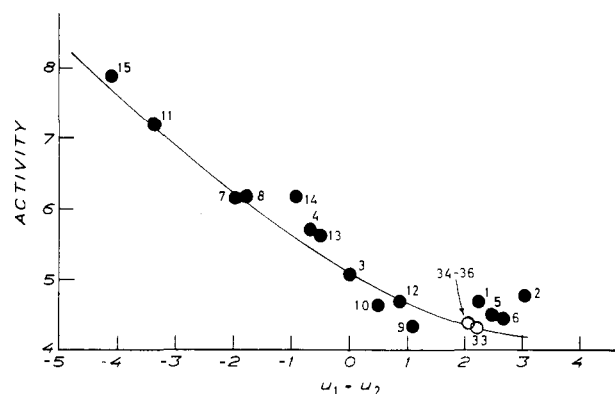


Figure 6. Activity estimation for class 2: \circ = test set.

objects with similar activities will cluster. As discussed above, the u_i vectors are related to the position of the objects within each class model in the measurement space. By plotting the u_1 vs. u_2 vectors for each class, as is shown in Figures 3 and 4, respectively, it can be seen that there is indeed an activity clustering. For class 1 the less active substances are in the region with positive u_1 and u_2 coordinates and the more active analogues have negative u_1 and positive u_2 coordinates. For class 2 the less active objects cluster in the region of negative u_1 and negative u_2 while the more active objects cluster with negative u_1 and positive u_2 coordinates. A plot of $(u_1 + u_2)$ against activity for class 1 (Figure 5) shows a significant relationship. Figure 6 shows a similar plot for class 2. This graphical analysis corresponds to a multiple regression relating for each class the activity to u_1 and u_2 . The graphical analysis is better for illustrative purposes but a multiple regression of activity as a function of the independent variables u_1 , u_2 , and u_3 will give the same result.

A $(u_1 - u_2)$ plot for the agonists in terms of intrinsic activity is given in Figure 7 and a graph of $(u_1 - u_2)/3$ with this activity (Figure 8) reveals a relationship between the u_i 's for the more active agonists while those with an intrinsic activity of 0.5 show no dependence on the u_i 's.

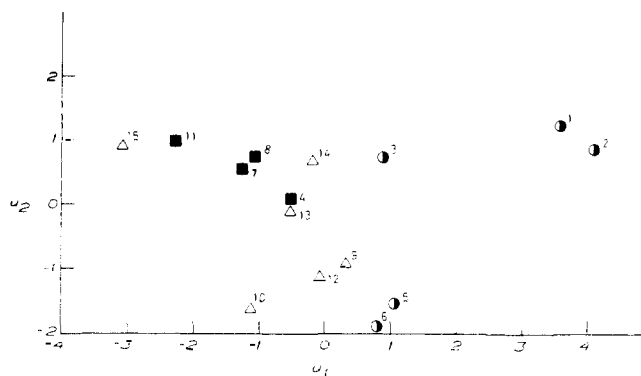


Figure 7. u_i plot for subclassification of agonists: Δ = low activity, \circ = medium activity, \blacksquare = high activity.

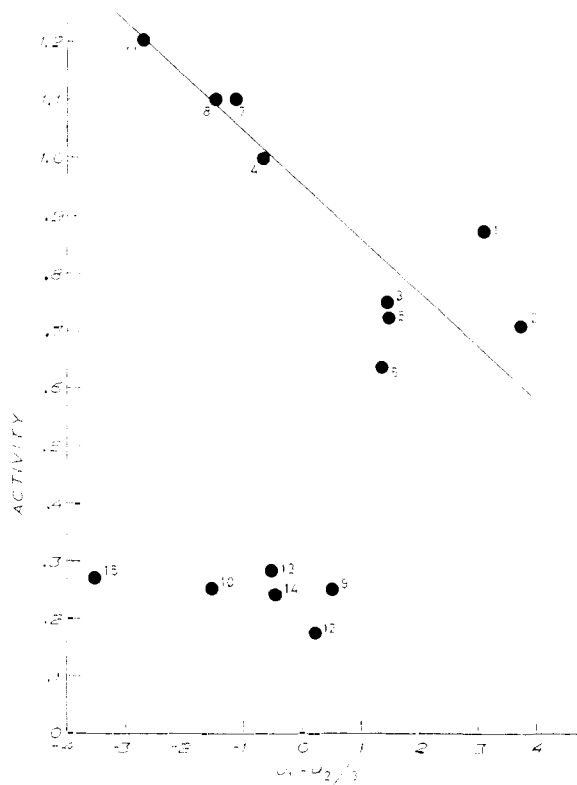


Figure 8. Activity estimation from subclassification of agonists.

The power of SIMCA is shown when this graphical analysis is extended to the objects in the test set. While classification is not conclusive in this case, the predicted low activity of each of the test objects as agonists or antagonists removes any ambiguity. It is necessary at this point to emphasize that the predicted activities are probably the upper limits for these substances due to the overestimation of the binding constants that is made.

Subclassification of the Agonists. At this point the agonists were divided into two groups on the basis of strongly active (objects 7, 9–15, and 18) and weakly active (objects 16, 17, and 19–22) as the graphical analysis suggests. This created classes 3 and 4, respectively. Application of SIMCA with the variables pK_b , f_{R_1} , f_{R_2} , σ_{R_2} , E_{s-R_2} , σ_p , f_{Ph} , L_m , B_m , L_p , and B_p and using a two-component model for each class gave correct classification of all members of both groups. A validation analogous to the one previously discussed where mini test sets were formed by objects from the training sets gave 100% correct classification. Due to the small number of objects in each class these results must be viewed as tentative but, nonetheless, the ability of a PaRC method to detect and

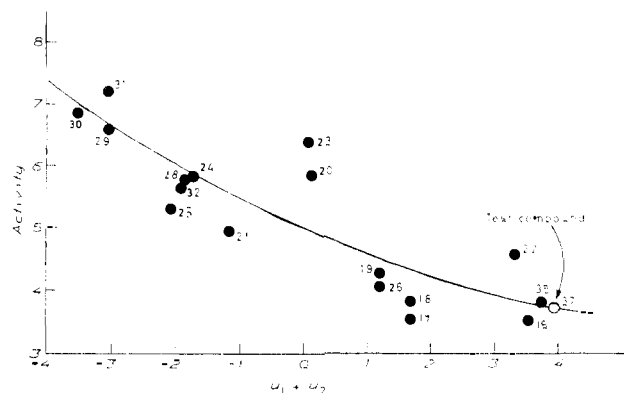


Figure 9. Activity estimation of antagonists using nonmeasured variables.

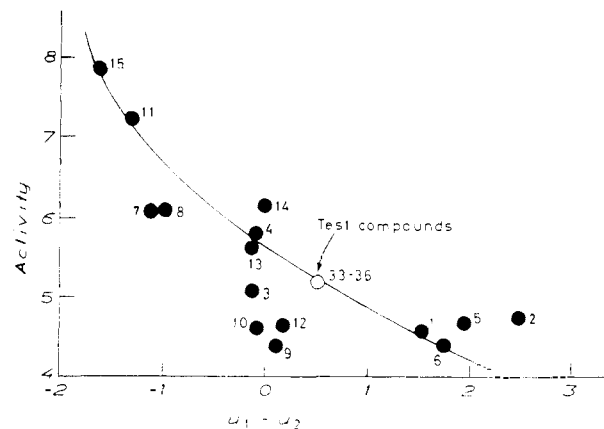


Figure 10. Activity estimation of agonists using nonmeasured variables.

separate classes on the basis of such small differences in structure is encouraging. The statistical data for this subclassification are given in Tables III and IV.

Predictions Based on Nonmeasured Variables Only. As a predictive tool, ideally, PaRC methods should be trained using nonmeasured variables that can be deduced from the molecular formula or can be looked up in tables. This would allow predictions to be made for objects without actually synthesizing them. In this study we have included the experimental receptor binding constant as a descriptor for each object in the training sets. The relevant predictive case would be that in which classification and quantification of activity could be done on an unknown object using only nonmeasured variables. In order to determine if our training sets could be used in such a manner, the PC models, which were derived for each training set with the binding constant included, were then applied to the training set with the pK_b set to zero. The classification results were 88% (15/17) for the antagonists with objects 21 and 30 missed and 87% (13/15) for the agonists with objects 11 and 15 missed. Activity predictions are given in Figures 9 and 10 and it can be seen that the predictability is good with the antagonists but only fair in the case of the agonists.

Summary of the SIMCA Analysis. The classification results can be summarized as the following. For class 1, using a PC model with $A = 3$ components and the variables σ^* , f_{Ph} , B_m , and L_m deleted, all were correctly classified. For class 2, using the same model and variables, 88% of the class was classified correctly. The model parameters are given in Tables V and VI. If a compound closely fits the model for class 1 with a RSD less than 0.53, it is classified as an antagonist. If a compound fits the model for class 2 with a RSD less than 0.31 it is an agonist, and

Table V. Parameters for Substituents R_1 and R_2

substituent no.	formula	f^a	$\sigma^{b,c}$	E_s^b
1	H	0.19	0.49	1.24
2	CH ₃	0.70	0.00	0.00
3	C ₂ H ₅	1.23	-0.10	-0.07
4	CH(CH ₃)O ₂	1.64	-0.19	-0.47
5	CH(CH ₃)CH ₂	2.35	-0.20	-0.51
6	CH(CH ₃)CH ₂ C ₆ H ₄ -4-OH	2.83	-0.13	-0.93
7	CH(CH ₃)CH ₂ C ₆ H ₃ -3,4-OCH ₂ O-	2.56	-0.13	-0.93
8	CH ₂ CH ₂ C ₆ H ₄ -4-OH	2.42	-0.08	-0.38
9	CH(CH ₃)(CH ₂) ₂ C ₆ H ₄ -4-OH	3.36	-0.13	-0.93
10	C(CH ₃) ₂ CH ₂ C ₆ H ₄ -4-OH	2.43	-0.30	-1.60
11	(CH ₂) ₃ C ₆ H ₄ -4-OH	2.95	-0.08	-0.38
12	C(CH ₃) ₂ CH ₂ C ₆ H ₅	3.80	-0.30	-1.60
13	CH(CH ₃)CH ₂ OC ₆ H ₅	2.77	-0.13	-0.93
14	CH(CH ₃)(CH ₂) ₂ C ₆ H ₅	3.90	-0.13	-0.93
15	C(CH ₃) ₃	2.24	-0.30	-1.60
16	CH ₂ CH ₂ OH	0.02 ^c		

^a Reference 19. ^b Pomona College Medicinal Chemistry Data Bank. ^c Corrected for proximity when required.

Table VI. Parameters for Substituents X and Y

substituent no.	formula	σ_p^a	σ_m^a	$B_p = B_m^b$	$L_p = L_m^b$
17	H	0.00	0.00	2.00	1.00
18	OH	-0.37	0.12	2.74	1.93
19	NHSO ₂ CH ₃ ^c	0.03	0.20	3.08	4.06
20	Cl	0.23	0.37	1.80	3.52
21	CH ₂ SO ₂ NH ₂	0.28 ^d		3.48	5.50
22	CH ₂ SO ₂ N(CH ₃) ₂		0.23 ^d	3.48 ^e	5.50 ^e
23	OCH ₃		0.12	2.87	3.98

^a Pomona College Medicinal Chemistry Data Bank. ^b Reference 21; $B = B_m$ in this reference. ^c L and B are assumed to be equal to NHSO₂CH₃. ^d Estimated value. ^e Assumed equal to substituent 21.

if it fits neither model within these RSD's it is an outlier (Table II).

Once an unknown has been classified, its activity can be estimated from the t_{ap} values obtained from the appropriate PC model for its class. Entering these values in Figure 5 or 6 will give an estimated activity for the unknown.

Agonists with intrinsic activity greater than 0.50 are well fit by a PC model with $A = 2$ components with the variables σ_p , B_p , and L_p deleted and the parameters given in Table IV. The activity of such an agonist can be predicted from the t_{ap} 's found on fitting the compound to this two-component model and the plot in Figure 8.

Discussion

The results reported here show that PaRC methods can give significant information from biological structure-activity data. In using PaRC methods for a specific problem, the problem must be formulated in such a way that the required level of classification as presented in this report is recognized. There are various methods of PaRC available and in the analysis of structure-activity data the choice of method should be based on the level of classification required.

The PaRC method is "trained" on reference sets of compounds of known class assignment. On these compounds, both "theoretical" variables and measured variables are included to stabilize the parameters in the class models as much as possible. When using these models for predictions of the behavior of new compounds, one needs fewer variables defined for these compounds; the predictions can be based on only "theoretical" variables deducible from the structure of the compounds.

Some comparison of SIMCA with multiple regression methods is in order at this point. A major philosophical difference with SIMCA compared to multiple regression

methods is that no *specific* model or relationship between structure and activity is assumed with SIMCA. One only assumes that such a relationship exists.

With multiple regression one is limited to the analysis of data on closely related compounds. A result of this is that classification is not possible since one considers all objects in the data set to be members of the same class. The concept of "outliers" then becomes cumbersome to deal with, while with PaRC their existence is a natural result of the analysis.

With multiple regression one is limited by the number of variables which can be used. This limitation applies not only to the number that can be initially selected but also to the number that can ultimately be used in the derived structure-activity relationship. This latter number should be as small as possible with a ratio of data points to variables of no less than five desirable.

This limitation does not apply to SIMCA. Its limitation is that of the number of components in the PC model used to describe a class. The number of components, A , should be less than approximately $M/3$ where M is the number of variables used to characterize the object. This advantage of SIMCA also holds when it is compared to other classification methods such as the linear learning machine and linear discriminant analysis.

Like multiple regression methods SIMCA can also be applied in the analysis of a single class of pharmacologically significant substances and activities of objects within the class can be estimated from the parameters of the derived PC model. In this case the problem is not one of classification.

To conclude, we are pleased with the performance of SIMCA in this first application of the method to structure-activity data. The information obtained from the data analysis regarding the significance of variables and, in particular, regarding the predictability of biological

activity is, in our view, most interesting.

Acknowledgment. We are grateful for support from the Swedish Natural Science Research Council and the Institute of Applied Mathematics, Stockholm.

References and Notes

- (1) On leave from the Department of Medicinal Chemistry, College of Pharmacy, University of Illinois/Medical Center, Chicago, Ill.
- (2) (a) C. Hansch, *Acc. Chem. Res.*, **2**, 232 (1969); (b) S. M. Free and J. W. Wilson, *J. Med. Chem.*, **7**, 398 (1964).
- (3) Y. C. Martin, J. B. Holland, C. H. Jarboe, and N. Plotnikov, *J. Med. Chem.*, **17**, 409 (1974).
- (4) P. H. A. Sneath, *J. Theor. Biol.*, **12**, 157 (1966).
- (5) A. Cammarata and G. K. Menon, *J. Med. Chem.*, **19**, 739 (1976); *J. Pharm. Sci.*, **66**, 304 (1977).
- (6) K. H. Ting, R. C. T. Lee, G. W. A. Milne, H. Shapiro, and A. M. Gaurino, *Science*, **180**, 417 (1973).
- (7) B. R. Kowalski and C. F. Bender, *J. Am. Chem. Soc.*, **96**, 916 (1974).
- (8) A. J. Stuper and P. C. Jurs, *J. Am. Chem. Soc.*, **97**, 182 (1975).
- (9) K. C. Chu, R. J. Feldman, M. B. Shapiro, G. F. Hazard, and R. I. Geran, *J. Med. Chem.*, **18**, 539 (1975).
- (10) L. J. Soltzberg and C. L. Wilkins, *J. Am. Chem. Soc.*, **99**, 439 (1977).
- (11) S. Wold, *Pattern Recognition*, **8**, 127 (1976).
- (12) S. Wold and M. Sjöström in "Chemometrics: Theory and Practice", ACS Symposium Series No. 52, B. R. Kowalski, Ed., American Chemical Society, Washington, D.C., 1977, p 243.
- (13) C. Mukherjee, M. C. Caron, D. Mulliken, and R. J. Lefkowitz, *Mol. Pharmacol.*, **12**, 16 (1976).
- (14) A. L. Perrin, *Science*, **183**, 551 (1974); R. J. Mathews, *J. Am. Chem. Soc.*, **97**, 935 (1975).
- (15) W. E. Brugger, A. J. Stuper, and P. C. Jurs, *J. Chem. Inf. Comput. Sci.*, **16**, 105 (1976).
- (16) A. J. Stuper, W. E. Brugger, and P. C. Jurs in "Chemometrics, Theory and Application", ACS Symposium Series No. 52, B. R. Kowalski, Ed., American Chemical Society, Washington, D.C., 1977, p 165.
- (17) L. P. Hammett, "Physical Organic Chemistry", 2nd ed, McGraw-Hill, New York, N.Y., 1971.
- (18) C. Hansch, A. Leo, S. H. Unger, K. H. Kim, D. Nikiatani, and E. J. Lien, *J. Med. Chem.*, **16**, 1207 (1973).
- (19) G. G. Nys and R. F. Rekker, *Eur. J. Med. Chem.*, **9**, 361 (1974).
- (20) R. W. Taft in "Steric Effects in Organic Chemistry", M. S. Newman, Ed., Wiley, New York, N.Y., 1956, p 556.
- (21) A. Verloop, W. Hoogenstraaten, and J. Tipker in "Drug Design", Vol. V, E. J. Ariens, Ed., Academic Press, New York, N.Y., 1971.
- (22) M. L. Weiner and P. H. Weiner, *J. Med. Chem.*, **16**, 655 (1973).
- (23) H. Wold, "On the Transition from Pattern Recognition to Model Building in Mathematical Economics and Game Theory. Essays in Honor of Oskar Morgenstern", R. Henn and O. Moeschlin, Ed., Springer-Verlag, Berlin, 1977.
- (24) S. Wold, *Technometrics*, in press.
- (25) J. Clark and D. D. Perrin, *Q. Rev., Chem. Soc.*, **18**, 295 (1966).

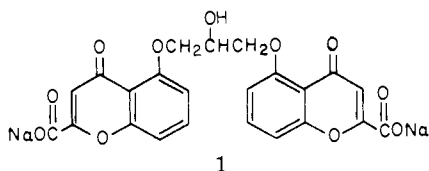
N,N'-(Phenylene)dioxamic Acids and Their Esters as Antiallergy Agents

John B. Wright,* Charles M. Hall,* and Herbert G. Johnson

Hypersensitivity Diseases Research, The Upjohn Company, Kalamazoo, Michigan 49001. Received February 17, 1978

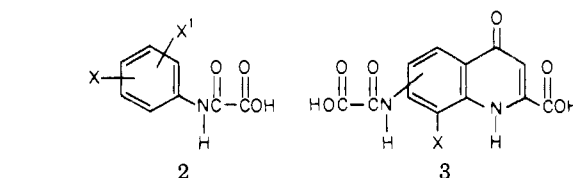
A series of dialkyl *N,N'*-(*m*-phenylene)dioxamates was synthesized by treatment of the requisite *m*-phenylenediamines with an alkyloxalyl chloride in the presence of triethylamine. Hydrolysis with sodium hydroxide solution gave the corresponding *N,N'*-(*m*-phenylene)dioxamic acids. Several *N,N'*-(*p*-phenylene)dioxamic acids were synthesized also in the same manner starting with the requisite *p*-phenylenediamines. These compounds were tested in the rat passive cutaneous anaphylaxis (PCA) assay. When tested *iv*, activity was found in the *N,N'*-(*m*-phenylene)dioxamic acids up to 2500 times that shown by disodium cromoglycate [50% inhibition at 0.001 mg/kg for *N,N'*-(2-chloro-5-cyano-*m*-phenylene)dioxamic acid (compound 61)]. Oral activity was seen in this series of compounds with duration of activity up to 120 min. Oral activity was detected in diethyl *N,N'*-(2-chloro-5-cyano-*m*-phenylene)dioxamate (compound 38) at levels of drug as low as 0.1 mg/kg.

Disodium cromoglycate (1) is an antiasthma agent that



is thought to act by inhibition of the liberation of the mediators of allergic reactions initiated by antigen-antibody reactions.¹ This activity may be measured conveniently in rats by means of the passive cutaneous anaphylaxis (PCA) reaction.²

Previously, we³ and others⁴ have reported that mono-dioxamic acids of the type 2 and 3 possess this same activity to an appreciable extent.



Disodium cromoglycate (1), as may be seen from its structure, possesses a "bis-functionality". A high order of activity was noted⁵ also in the fused-ring quinaldic acids which also possess a bis-functionality. In order to explore the importance of this bis-functionality on the biological activity we synthesized and studied biologically a series of *N,N'*-(phenylene)dioxamic acids and their esters. The results of this study are described below.

Chemistry. Synthesis of diethyl *N,N'*-(*m*-phenyl-