Journal of Medicinal Chemistry

© *Copyright 1986 by the American Chemical Society*

Volume 29, Number 11 November 1986

Computer-Aided Studies of the Structure–Activity Relationships between the Structure of Some Steroids and Their Antiinflammatory Activity

Terry R. Stouch and Peter C. Jurs*

The Pennsylvania State University, University Park, Pennsylvania 16802. *Received December 23, 1985*

The relationship between variation in structure and variation in antiinflammatory activity was investigated for 125 steroids whose antiinflammatory activity had previously been determined by using the McKenzie–Stoughton human vasoconstrictor assay. Eighty-eight of the compounds were used in the training stages of analysis. A two-class problem was developed by classifying the compounds as low-to-no potency (37 compounds) or potent-to-very potent (51 compounds) on the basis of their activity relative to that of hydrocortisone butyrate. Thirty-eight different structural variations occurred at six different sites on the steroid nucleus. These variations were coded by a total of 10 descriptors—three indicator descriptors and seven descriptors that coded for the lipophilicity of the substituents at specific sites of variation. Linear discriminant analysis, principal components plots, *K* nearest neighbor analysis, and statistical measurements of class separation all confirmed that the more potent compounds existed in a region of the data space different from the less potent compounds. This structure–activity relationship was applied to the prediction of the activities of 37 compounds that were not used in the preliminary analysis with good results.

Steroids play a part in many biological processes and have many therapeutic applications. Over several decades, much effort has been applied to the development of steroids that have both strong antiinflammatory activity and few negative side effects. Variations at several sites on the steroid nucleus yield compounds of greatly differing activities. The biological importance of these compounds, their structural similarity, and the large amount of available data make this class of compounds attractive for structure–activity relationship (SAR) analysis. SAR studies have been performed by using Hansch analysis,[1-3] Free–Wilson analyses,[4-5] conformational analysis,[6] and quantum mechanics.[7] All of these have dealt with small sets of compounds of little variation. Only one study that we know of has used pattern recognition (PR) techniques. This study by Bodor et al.[8] dealt with a larger and more diverse set of compounds than any previous study.

The study of Bodor et al. was performed on 122 steroids with the common nucleus shown in Figure 1. Of these compounds, 74 were considered to be potent antiinflammatory agents and 48 were considered to be nonpotent on the basis of their activities relative to hydrocortisone butyrate. Variations in the steroid nucleus occurred at six sites and are listed in Table I. These variations were described with use of a set of 33 indicator variables that coded for the presence or absence of these variations. Through feature selection methods, 11 of the descriptors were identified as having no effect on classification results. This resulted in a final set of 22 descriptors. With use of these 22 descriptors, a linear learning machine generated a linear discriminant function (LDF) that correctly classified all 122 of the compounds.

The McKenzie–Stoughton human vasoconstrictor test[9] was used to determine the antiinflammatory activity of the compounds. In this test, the vasoconstricting potential of a compound is assessed by applying it in alcoholic solution to the skin of a human forearm, occluding the test area for 16 h, and estimating the resulting level of blanching of the skin. The degree of blanching is positively correlated with antiinflammatory activity. The data provided by this assay are well-suited to analysis by pattern recognition techniques. While a quantitative index of activity is reported for this test, the raw data consist of subjective human evaluations of levels of blanching relative to a standard, usually fluocinolone acetonide, **5**. This results in wide confidence intervals for the reported quantitative values and variations between laboratories. The activities of the compounds used by Bodor et al. ranged from zero to 1900; some of the ranges for individual compounds were several hundreds of units. Furthermore, in order to include as many data as possible in their study, Bodor et al. included some compounds that had only relative activities reported or that were assayed with use of a different standard and sometimes in slightly different systems. In such cases, it is difficult to assign a consistent quantitative index of activity; however, some confidence can be given to general ranking of activities. Pattern recognition techniques deal well with such semiquantitative data and can make use of such information when other methods are not applicable. Bodor et al. arbitrarily separated the data into active and inactive classes by specifying an activity cutoff of 50, the activity of hydrocortisone butyrate. Those compounds with an activity less than 50 were considered inactive, and those with an activity greater than 50 were considered active. The studies reported here adhere to that cutoff.

[1] Wolff, M. E., et al. *J. Steroid Biochem.* **1975**, *6*, 211.

[2] Wolff, M. E. In *Glucocorticoid Hormone Action*; Baxter, J. D., Rousseau, G. G. Eds.; Springer-Verlag: New York, 1979; p 97.

[3] Wolff, M. E. In *Burger's Medicinal Chemistry*, 4th ed.; Wolff, M. E., Ed.; Wiley: New York, 1980; Part III, p 1273.

[4] Schmit, J. P.; Rousseau, G. G. *J. Steroid Biochem.* **1978**, *9*, 909.

[5] Schmit, J. P.; Rousseau, G. G. *J. Steroid Biochem.* **1978**, *9*, 921.

[6] Kollman, P. A.; Giannini, D, D.; Duax, W. L.; Rothenberg, S.; Wolff, M. E. *J. Am. Chem. Soc.* **1973**, *95*, 2869.

[7] Wolff, M. E., et al. *Quantitative Structure-Activity Relationships*; Akademiai Kiado: Budapest, 1973; p 31.

[8] Bodor, N.; Harget, A. J.; Phillips, E. W. *J. Med. Chem.* **1983**, *26*, 318–328.

[9] McKenzie, A. W.; Atkinson, R. M. *Arch. Dermatol.* **1964**, *89*, 741–746.

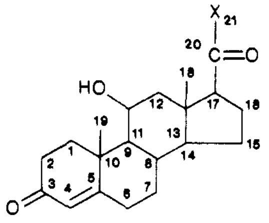


Figure 1. Structural backbone common to all 125 compounds examined in this study.

Table I. Structural Variations and Their Corresponding Descriptor Values

index	position	structural variation	descriptor value	descriptor type
1	1,2	saturation	1	indicator
2	6	hydro	0.23	log *P*
		fluoro	-0.38	
		methyl	0.89	
3	9	hydro	0.23	log *P*
		fluoro	-0.38	
		chloro	0.06	
4	16,17	acetonide	1	indicator
5	16	α or β methyl	1	indicator
		hydroxyl	not coded	
6	17	hydro	0.23	log *P*
		methyl	0.89	
		hydroxyl	-1.64	
7	17	OCOR, R =		
		methyl	0.89	log *P*
		ethyl	1.55	
		propyl	2.09	
		n-butyl	2.63	
		isopropyl	2.08	
		tert-butyl	2.50	
8	21	hydroxyl	-3.29	log *P*
		methyl	0.89	
		ethyl	1.55	
		chloromethyl	0.72	
		methoxymethyl	-0.11	
		hydroxymethyl	-0.98	
		phosphomethyl	-1.17	
		chloromethyl-carboxylate	-0.77	
9	21	OR, R =		
		methyl	0.89	log *P*
		ethyl	1.55	
		propyl	2.09	
		(methylthio)methyl	0.64	
		chloromethyl	0.72	
10	21	CH_2OCOR, R =		
		methyl	0.89	log *P*
		ethyl	1.55	
		propyl	2.09	
		butyl	2.63	
		iospropyl	2.08	
		tert-butyl	2.50	

The nonparametric linear discriminant analysis that was performed by Bodor et al. followed the then-established guidelines concerning the ratio of descriptors to observations that is permitted to avoid fortuitous 100% correct classifications when using these pattern recognition techniques.[10,11] Recently, we have investigated the reliability of the levels of correct classifications provided by these pattern recognition methods and have found that they may be overly optimistic in indicating the presence of a relationship between structure and activity.[12-14] The results of these studies show that for 122 data points divided 48/74 between two classes, a 22-dimensional data space can support correct classifications due to chance of between 85% and 90%. When the data space used by Bodor et al. was simulated in 10 different trials, as it was for several other SAR studies using pattern recognition,[13] rates of correct classifications ranged between 93% and 96%.

These results indicate that much of the linear separation of the two classes shown in the study of Bodor et al. could have been due to chance. This is not to say that the results are meaningless; but, as we suggested previously,[12] the results provided by nonparametric linear discriminant functions (NLDFs) should be verified by other methods when the level of correct classifications due to chance becomes high.

Bodor et al. used a Free-Wilson-like approach to represent the steroids. In Free–Wilson analysis, each variation in structure is assumed to elicit a constant variation in activity regardless of other structural variations. In fact, this assumption does not hold within this data set. For example, in several cases, unsaturation at the 1,2-position yielded no change in activity, but such a change in compound **78** to yield **92** caused an increase in vasoconstriction from 360 to 720. Another problem with the Free–Wilson approach is that it deals only in substructural variations. Information concerning the physical properties that affect the activities is not used, and no extrapolation or interpolation can be made beyond the substituent types present in the data set. The major problem with this approach, for this particular data set, at least, is that over 30 descriptors must be used to represent the 122 compounds. For the data set of 122 compounds, this will result in high level of correct classifications due to chance for linear discriminant function analysis. Reduction of this descriptor set, however, would remove some substituents from the analysis. For example, in the final descriptor set used by Bodor et al., 11 of the variations were removed from the analysis to yield the final set of 22 descriptors. Chlorination of the 9-position and methylation of the 6-position were not represented. Esterification with isovaleric ester at the 21- and 17-positions was not represented. These features were assigned no importance, even though these variations have previously been found to influence activity. An implication of this is that acetyl, propanoyl, and butanoyl esters at the 21- and 17-positions affect activity but isovaleryl esters do not.

In this paper, the data used in that previous study are reexamined with use of several multivariate methods, including linear discriminants, in an effort to further establish and verify the SAR between the structures of these compounds and the results of the McKenzie–Stoughton test. Rather than applying the Free–Wilson formalism, we attempted to code for the structural variations in a more physically meaningful way.

Data. Bodor et al. used all 122 compounds drawn from the literature in their SAR study. The activities of only 88 of these were reported in such a way that they could be unambiguously ranked relative to hydrocortisone butyrate, which had an activity of 50. Of these 88, 51 had activities greater than 50, and so were considered to be active, and 37 had activities less than 50, and so were considered to be inactive. Of the remaining 34, some were reported only as having activities less than that of betamethasone 17-valerate, **26** (activity = 360), or fluocinolone

[10] Stuper, A. J.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.* **1976**, *4*, 238.

[11] Whalen-Pedersen, E. K.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 264.

[12] Stouch, T. R.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 45–50.

[13] Stouch, T. R.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 92–98.

[14] Stouch, T. R.; Jurs, P. C. *Quant. Struct.-Act. Relat.* **1986**, *5*, 57–61.

acetonide, **5** (activity = 100). In the studies reported here, only the unambiguous 88 compounds were used to investigate the SAR so that the analysis would not be complicated by potentially erroneous data. This also excluded a set of compounds with which to perform an unbiased test of the predictive ability of the results of the pattern recognition analyses. All the compounds are listed in Table II; those excluded from analysis are indicated. The numbering and classifications for the compounds are the same as those used by Bodor et al. References to the experimental data can be found in the work of Bodor et al.[8]

Descriptor Development. We wanted to represent the chemical structures by a concise numerical description of the physicochemical variations caused by the structural variations of the compounds. We sought to do this, and at the same time reduce the dimensionality of the problem, by representing the variations at each site of substitution, as is performed for a Hansch analysis. While there are many different substitutions at any one position, they will all exert their influence through physicochemical properties such as steric, electrostatic, and hydrophobic effects.

The structural variations were coded as follows. Indicator variables were used to represent the presence of saturation at the 1,2-position, the presence of an acetonide linkage between sites 16 and 17, and methylation at site 16. Since these variations were binary in nature, they were numerically encoded as "1" if one of the variations was present and "0" if it was not present. The rest of the sites were coded by the log P of the substituents at that site, including hydrogen substitution. In all cases, these values were calculated by the method of Hansch and Leo.[15]

We chose to use the log P of the substituents for several reasons. First, the substituents at several of the sites vary widely. By definition, this descriptor will contain information regarding chain length, branching, and size and the presence of heteroatoms. Use of this descriptor allows us to code all of the variations at any one site whereas other descriptors would not be as general. Second, log P is a physically meaningful quantity. Use of molecular connectivity indices may give similar information, but they have no physical basis. Third, this descriptor has been found to be useful in many SAR studies.

The ester linkages at the 17- and 21-positions could be hydrolyzed during metabolism of these compounds. This has been examined by previous SAR studies.[16] Because of this, these linkages were coded separately from the nonhydrolyzable substitutions. In order to maximize the variance of those descriptors, the esterifications were coded by the log P of the side chain of the esterifying acid rather than that of the entire acid. Since all the side chains were simple hydrocarbons and since log P for simple hydrocarbons can be calculated from a simple additivity scheme, this is equivalent to removing a constant factor from the log P of each acid moiety.

This representation coded the compounds with 10 descriptors—three indicator descriptors and seven descriptors coding for the log P at the sites of varying substitution (positions 6, 9, 17, and 21) or esterification (positions 17 and 21). Each compound could be thought of as a point in a 10-dimensional descriptor space, where each axis is defined by one of the 10 descriptors. All descriptors were autoscaled prior to the data analysis. A list of these descriptors and their numerical values is presented in Table I, and the descriptor values for one of the compounds are listed in Table III.

Data Analysis. Our approach to this problem was to use a variety of techniques in order to probe this 10-dimensional data space and establish biological activity trends within it. The methods that we used in these studies are well-described in other sources such as pattern recognition texts[17-19] or multivariate statistical texts,[20,21] and we will furnish only brief descriptions here.

Linear discriminants can be generated statistically or in a pattern recognition sense with use of a linear learning machine[17-19] or related least-squares techniques.[22,23] Linear discriminants are vectors of weights for the descriptors. These vectors define a hyperplane that separates the data space into distinct regions. For an SAR study, these regions would contain compounds of differing activity. For our current study, we would hope that those compounds with potent antiinflammatory activity would lie in a different region of the data space than those compounds that are nonpotent and that a linear discriminant would partition the data space accordingly.

A linear discriminant generated by an adaptive least-squares method of linear discriminant generation[22] could classify all but four of the 88 training set compounds. However, the ratio of descriptors to observations for this study was approximately 1/9. According to our previous studies,[12] the levels of correct classifications due to chance for this ratio could be as high as 85%. Several other multivariate methods were used to verify that the LDF results were significant.

If linear discriminant analysis is to provide meaningful results, the classes that compose the problem must be separate in the data space. One method for assessing this separation is a comparison of the means of the two classes. If the means are identical, the classes overlap completely. While one class may be more diffuse than the other, linear discriminant analysis will provide nothing but fortuitous results. If the classes are separate, then, in many cases, this could be determined by a comparison of the means of the classes.

The Mahalanobis distance between two vectors is calculated by eq 1, where $\bar{x}_i$ is the mean vector of class i, S is the common variance–covariance matrix, and D is the Mahalanobis distance. This metric provides a distance

$$D = (\bar{x}_1 - \bar{x}_2)^T S^{-1} (\bar{x}_1 - \bar{x}_2) \qquad (1)$$

measure between two vectors that is free of complications caused by correlation or multicorrelation between the descriptors. Equation 2 relates the Mahalanobis distance to an F statistic, providing a measure of significance with which to interpret the results. In this equation, n_i is the

$$J = \frac{n_1 n_2}{n}\left(\frac{n - 1 - d}{(n - 2)d}\right) D^2 \qquad (2)$$

number of observations in class i, d is the number of de-

[15] Hansch, C.; Leo, A. *Substituent Constants for Correlation Analysis in Chemistry and Biology*; Wiley: New York, 1979.

[16] Phillipps, G. H. In *Mechanisms of Topical Corticosteroid Activity*; Wilson, L., Marks, R., Eds.; Churchill Livingston: New York, 1976; pp 1–14.

[17] Nilsson, N. J. *Learning Machines*; McGraw-Hill: New York, 1965.

[18] Tou, J. T.; Gonzalez, R. C. *Pattern Recogntion Principles*; Addison-Wesley: Reading, MA, 1974.

[19] Varmuza, K. *Pattern Recognition in Chemistry*; Springer-Verlag: New York, 1980.

[20] Hand, D. J. *Discrimination and Classification*; Wiley: New York, 1981.

[21] Johnson, R. A.; Wichern, D. W. *Applied Multivariate Analysis*; Prentice-Hall: Englewood Cliffs, NJ, 1982.

[22] Moriguchi, I.; Komatsu, K.; Matsushita, Y. *J. Med. Chem.* **1980**, *23*, 20.

[23] Pietrantonio, L.; Jurs, P. C. *Pattern Recognition* **1972**, *4*, 391.

Table II. List of Compounds

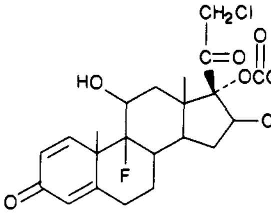
no.	name	class[a]
1	clobetasol 17-propionate	P
2	dexamethasone	NP
3	beclomethasone 17,21-dipropionate	P
4	hydrocortisone 21-acetate	NP
5	fluocinolone acetonide	P
6	hydrocortisone	NP
7	9α-fluorohydrocortisone	NP
8	prednisolone 21-acetate	NP
9	halcinonide	P
10	9α-fluoro-21-chloro-11β-hydroxy-16β-methylpregna-4,4-diene-3,20-dione 17-butyrate	P
11	corticosterone	NP
12	prednisolone	NP
13	hydrocortisone 17-butyrate	P
14	fluocinonide	P
15	ethyl 17α-acetoxy-9-fluoro-11β-hydroxy-16β-methyl-3-oxo-androsta-1,4-diene-17β-carboxylate	NP
16	propyl 9-fluoro-11β,17α-dihydroxy-16β-methyl-3-oxoandrosta-1,4-diene-17β-carboxylate	NP
*17[f]	11β-hydroxy-16α,17α-dihydroxy-16β-methylpregna-1,4-diene-3,20-dione 21-acetate	P
*18	diflucortolone 21-valerate	P
*19	fluocortolone	P
*20	flumethasone 21-pivalate	P
21	dexamethasone 21-phosphate	NP
*22	flurandrenolone	NP
*23	flurandrenolone acetonide	P
24	triamcinolone acetonide	P
25	betamethasone 17-isobutyrate	P
26	betamethasone 17-valerate	P
27	betamethasone 21-acetate	NP
28	betamethasone	NP
29	methyl prednisoloneacetate	NP
*30	9α-fluoro-16α,17-dimethylpregna-1,4-diene-3,20-dione 21-propionate	P
*31	(methylthio)methyl 17α-(pentanoyloxy)-11β-hydroxy-3-oxoandrost-4-ene-17β-carboxylate	NP
*32	9α-fluoro-21-chloro-11β,16α,17α-trihydroxypregna-1,4-diene-3,20-dione 16,17-acetonide	P
33	9α-fluoro-11β,17α-dihydroxy-16β-methyl-3-oxo-androsta-1,4-diene-17β-carboxylic acid	NP
34	methyl 17α-acetoxy-9α-fluoro-11β-hydroxy-16β-methyl-3-oxoandrost-1,4-diene-17β-carboxylate	P
35	betamethasone 21-isobutyrate	P
36	chloromethyl 17α-(propanoyloxy)-9-fluoro-11β-hydroxy-16β-methyl-3-oxoandrosta-1,4-diene-17β-carboxylate	P
*37	paramethasone	NP
38	betamethasone 21-butyrate	P
*39	chloromethyl 17α-(propanoyloxy)-11β-hydroxy-3,20-dioxopregn-4-en-21-oate	P
*40	11β-hydroxy-16α,17α,21-trimethylpregna-1,4-diene-3,20-dione	P
41	methyl 17α-(butanoyloxy)-9-fluoro-11β-hydroxy-16β-methyl-3-oxoandrosta-1,4-diene-17β-carboxylate	P
42	betamethasone 21-valerate	NP
43	betamethasone 17-acetate	P
44	6α,9α-difluoroprednisolone 17-isobutyrate	P
45	6α,9α-difluorohydrocortisone 17-valerate	P
46	6α,9α-difluoroprednisolone 17-propionate	P
47	6α,9α-difluoroprednisolone 17,21-dibutyrate	P
48	triamcinolone	NP
49	dexamethasone 21-acetate	NP
50	6α,9α-difluoroprednisolone 17-pripionate 21-trimethylacetate	P
51	prednisolone 17-valerate	NP
52	ethyl 17α-(propanoyloxy)-9-fluoro-11β-hydroxy-16β-methyl-3-oxoandrosta-1,4-diene-17β-carboxylate	P
53	6α,9α-difluoro-21-deoxyprednisolone 17-acetate	P
54	beclomethasone	NP
*55	paramethasone 21-acetate	P
56	6α,9α-difluoroprednisolone 17,21-diacetate	P
57	9α-fluoro-21-chloro-11β-hydroxy-16β-methyl-pregna-1,4-diene-3,20-dione 17-isobutyrate	P
58	prednisolone phosphate	NP
59	propyl 17α-(propanoyloxy)-9-fluoro-11β-hydroxy-16β-methyl-3-oxoandrosta-1,4-diene-17β-carboxylate	NP
60	17α-acetoxy-9-fluoro-11β-hydroxy-16β-methyl-3-oxo-androsta-1,4-diene-17β-carboxylic acid	NP
61	methyl 17α-(propanoyloxy)-9-fluoro-11β-hydroxy-16β-methyl-3-oxoandrosta-1,4-diene-17β-carboxylate	P
62	betamethasone 17-butyrate	P
63	betamethasone 21-propionate	NP
64	hydrocortisone 17-valerate	NP
*65	chlorcortolone	P
*66	desonide	P
67	9α-fluorohydrocortisone acetate	NP
68	methyl prednisolone	NP
*69	(methylthio)methyl 11β,17α-dihydroxy-3-oxoandrost-4-ene-17β-carboxylate	NP
*70	chloromethyl 17α-(propanoyloxy)-11β-hydroxy-3-oxoandrost-4-ene-17β-carboxylate	P
*71	9α-fluoro-11β,21-dihydroxy-16α,17α-dimethylpregna-1,4-diene-3,20-dione	P
*72	chloromethyl 6α,9α-difluoro-11β-hydroxy-3,20-dioxopregna-1,4-dien-21-oate 16,17-acetonide	P
73	17α-(propanoyloxy)-9-fluoro-11β-hydroxy-16β-methyl-3-oxoandrosta-1,4-diene-17β-carboxylic acid	NP
74	6α,9α-difluoroprednisolone 17-isobutyrate 21-acetate	P
75	6α,9α-difluoroprednisolone 17-acetate 21-isobutyrate	P
76	6α,9α-difluoroprednisolone 17-acetate	P
*77	6α,9α-difluoroprednisolone	NP
78	6α,9α-difluoro-11β-hydroxypregn-4-ene-3,20-dione 17-valerate 21-acetate	P
79	ethyl 9-fluoro-11β-hydroxy-16β-methyl-3-oxo-17α-(butanoyloxy)androsta-1,4-diene-17β-carboxylate	NP
*80	deoxymethasone	P
81	21-deoxycortisol	NP
82	betamethasone 17,21-dipropionate	P
83	propyl 9-fluoro-11β-hydroxy-16β-methyl-3-oxo-17α-(butanoyloxy)androsta-1,4-diene-17β-carboxylate	NP
84	21-deoxybetamethasone 17-propionate	P
*85	6α,9α-difluoroprednisolone 17-acetate 21-butyrate	P
86	6α,9α-difluoroprednisolone 17-propionate 21-isobutyrate	P
*87	difluorocortolone trimethylacetate	P
88	6α,9α-difluoroprednisolone 17-valerate	P
89	6α,9α-difluoroprednisolone 17-butyrate	P
90	6α,9α-difluoro-21-deoxyprednisolone 17-propionate	P
91	fluprednisolone	NP
92	6α,9α-difluoroprednisolone 17-valerate 21-acetate	P
93	flurandrenolone acetate	NP
94	methyl 9-fluoro-11β,17α-dihydroxy-16β-methyl-3-oxoandrosta-1,4-diene-17β-carboxylate	NP
95	9-fluoro-11β-hydroxy-16β-methyl-3-oxo-17α-(pentanoyloxy)androsta-1,4-diene-17β-carboxylic acid	NP
*96	21-chloro-11β,16α,17α-trihydroxypregna-1,4-diene-3,20-dione 16,17-acetonide	P
97	6α,9α-difluoroprednisolone 17-butyrate 21-acetate	P
98	6α,9α-difluoroprednisolone 17-butyrate 21-propionate	P
*99	hydrocortisone 17-acetate	NP
100	medrysone	NP
*101	chloromethyl 11β,17α-dihydroxy-3-oxoandrost-4-ene-17β-carboxylate	NP
102	betamethasone 17-propionate	P
103	hydrocortisone phosphate	NP
*104	(methylthio)methyl 9α-fluoro-17α-(pentanoyloxy)-16β-methyl-11β-hydroxy-3-oxoandrosta-1,4-diene-17β-carboxylate	NP
105	6α,9α-difluoroprednisolone 17,21-dipropionate	P
106	methyl 17α-(pentanoyloxy)-9-fluoro-11β-hydroxy-16β-methyl-3-oxoandrosta-1,4-diene-17β-carboxylate	NP
*107	fluocinolone	NP
108	6α,9α-difluoroprednisolone 17-propionate 21-acetate	P
*109	21-chloro-11β,16α,17α-trihydroxypregn-4-ene-3,20-dione 16,17-acetonide	P
*110	9α-fluoro-21-chloro-11β,17α-dihydroxy-16β-methyl-pregna-1,4-diene-3,20-dione	NP
111	6α,9α-difluoroprednisolone 17-butyrate 21-isobutyrate	P

Table II (Continued)

no.	name	class[a]
112	6α,9α-difluoroprednisolone 17-propionate 21-butyrate	P
113	beclomethasone 17-propionate	P
114	6α,9α-difluorodeoxyprednisolone 17-butyrate	P
*115	desfluochlorocortolone trimethylacetate	P
*116	chloromethyl 17α-(pentanoyloxy)-11β-hydroxy-3,20-dioxopregn-4-en-21-oate	P
*117	prednisolone 17-valerate 21-acetate	P
118	6α,9α-difluoroprednisolone 17-isobutyrate 21-propionate	P
119	6α,9α-difluoroprednisolone 17,21-diisobutyrate	P
*120	betamethasone phosphate	NP
121	9α-fluoro-21-chloro-11β-hydroxy-16β-methylpregna-1,4-diene-3,20-dione 17-valerate	P
122	9α-fluoro-21-chloro-11β-hydroxy-16β-methylpregna-1,4-diene-3,20-dione 17-acetate	P
*123	6α,9α-difluoroprednisolone 17,21-dibutyrate	P[b,c]
*124	dexamethasone 21-acetate	NP[b,d]
*125	beclomethasone 17-propionate	P[b,e]

[a] P = potent, NP = nonpotent. [b] Not included in the analysis of Bodor et al. [c] Gardi, R.; Vitali, R.; Falconi, G.; Ercoli, A. *J. Med. Chem.* **1972**, *15*, 556. [d] Lorenzetti, O. J. *Curr. Ther. Res.* **1979**, *25*, 92. [e] Harris, D. M. *J. Steroid Biochem.* **1975**, *6*, 711. [f] * means not included in the 88-compound training set.

Table III. Values of Table I Descriptors for Clobetasol 17-Propionate (1)



descriptor index	variation	descriptor value	descriptor type
1	1,2-unsaturation	0	indicator
2	6-hydro	0.23	log *P*
3	9-fluoro	−0.38	log *P*
4	no 16,17-acetonide linkage	0	indicator
5	16-methyl	1	indicator
6	17-esterification	0.0	log *P*
7	17-$OCOCH_2CH_3$	1.55	log *P*
8	21-CH_2Cl	0.72	log *P*
9	no 21-esterification	0.0	log *P*
10	no 21-esterification	0.0	log *P*

Table IV. KNN Results for the First 10-Descriptor Set

	percentage correctly classified		
no. NN[a]	overall	active	inactive
1	86	96	73
3	90	92	86
5	90	92	86
7	85	90	78

[a] Number of nearest neighbors included in classification.

scriptors used, and D is the Mahalanobis distance. J is compared with the F distribution with d and $(n - 1 - d)$ degrees of freedom.[20] For the 10-dimensional data space described above, the Mahalanobis distance between the means is 19.2. The J value for this problem is 36.1, and the tabulated F statistic is 1.9. This shows that the means are different at the 95% probability level. The F statistics were tabulated for normally distributed, continuous data, and so we do not expect to interpret these results strictly. It is further evidence, however, that the two classes that define this problem are well-separated in this 10-dimensional space.

The K nearest neighbor (KNN) method classifies a pattern on the basis of the classes of neighboring patterns. For example, first nearest-neighbor classification assigns a compound to the class of that compound that is nearest to it in the descriptor space. Any odd number of nearest neighbors can be used; a vote can be taken if there are disagreements between neighbors. As in all pattern recognition methods, this method makes the assumption that compounds of like activity will have similar descriptor values and so will exist in similar regions of the descriptor space. If this assumption holds, and if the classes are well-separated in the data space, then the nearest-neighbor classification success rate will be high. The results for this data space are tabulated in Table IV. The Euclidean metric was used to calculate the distances between the patterns. Monte Carlo trials using 88 patterns with the same class distribution but randomly assigned classes gave classification levels less than 61%, far lower than the 85–90% correct classifications reported in Table IV.

Additional evidence for data structure was provided by principal components projection of the data space. Principal components analysis provides a new set of axes, which are linear combinations of the original axes. These new axes are mutually orthogonal and are calculated such that each successive axis contains a successively smaller portion of the variance in the data. With this method, much of the variance of a high-dimensional space can often be projected into a lower dimensional space and displayed for visual examination. Often, this allows relationships within high-dimensional data to be examined visually. Figure 2 is a plot of the data projected into the space of the first two principal components of the data. Even though these two vectors account for only 54% of the total variance of the data, much class structure is evident.

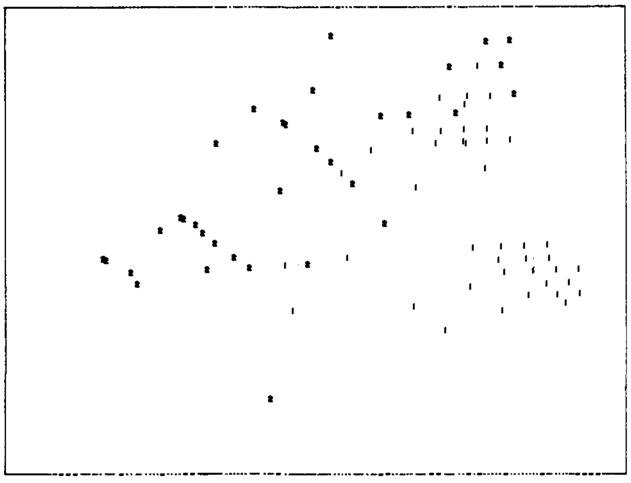


Figure 2. Data plotted in the first two principal components of the first set of 10 descriptors. 1's are active, 2's are inactive.

Since KNN analysis and principal components (PC) plots both provided evidence of a real difference in positioning of the two classes in the data space, the classification results provided by the LDF were assumed to be due to real data structure and not to chance, and the four compounds that were misclassified in the LDF analysis (**93**, **13**, **52**, **42**) were examined. The activities of these compounds were all close to 50, the activity that was chosen as the cutoff between the active and inactive classes. This, in itself, provides some evidence of data structure. Compounds with activities close to the cutoff would presumably lie somewhere between the most and least active compounds in the data space and so would be expected to be

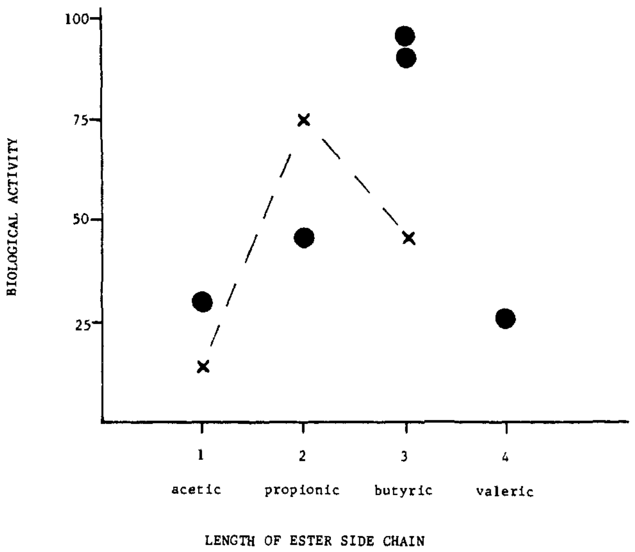


Figure 3. Plot of the nonlinear relationship between the activity and the length of the side chain of the acid esterifying the 17 (X) and 21 (●) positions.

close to a discriminant that divided the two classes. On reexamination of the original literature,[24] the classification of compound **93** as inactive was questioned. It was assigned to the inactive class; however, the literature value was listed as 100–300 and that of hydrocortisone butyrate, the cutoff, was listed as 300. These were not reported on the same scale as most of the rest of the data in which hydrocortisone butyrate had an activity of 50. Another misclassified compound, **13**, hydrocortisone butyrate, was that compound whose activity was used as the cutoff between the active and inactive classes.

The other two misclassified compounds yield some valuable information. Compound **52** has an activity of 75, relatively close to 50, the cutoff. Its nearest neighbors in the data space were compounds **79** (activity = 40) and **15** (activity = 16). The only difference between these compounds was in the acid that esterified the 17-position. The 17-position of compound **52** was esterified with propanoic acid, that of compound **79** with butanoic acid, and that of **15** with acetic acid. This indicates a nonlinear relationship between the length of this side chain and activity (Figure 3). Within this homologous series, esterification of the 17-position with propanoic acid yields the highest activity. Esterification with acids with longer or shorter side chains decreases the activity.

Much the same effect was seen for compound **42** (activity = 26). Its 21-position was esterified with valeric acid. In this data space, its nearest neighbors were esterified with propanoic acid (activity = 40), acetic acid (activity ~ 30), butanoic acid (activity = 90), and isobutyric acid (activity = 85). Once again, a nonlinear relationship exists between the length of the side chain and the activity (Figure 3). For this series, a three-carbon side chain on the esterifying acid yielded the highest activity.

The descriptors that coded for the log *P* of the side chains of the 17- and 21-esters were transformed to reflect these nonlinear relationships. For the 17-position, this was done by subtracting the log *P* of the side chain of the propanoic acid (1.55) and squaring the result for each value in the descriptor. For the 21-position, this was done by subtracting the log *P* of the *n*-propyl group (2.09) from the log *P* of the side chain of the acids and squaring that term. These transformed descriptors were substituted for the original descriptors that coded for the ester substitution at sites 17 and 21. These transformations caused the new descriptors to reflect the nonlinear nature of the dependence of activity on chain length. The most active substituent had a value of zero, and any chain length longer or shorter had a negative value.

Table V. KNN Results for the Second 10-Descriptor Set

	percentage correctly classified		
no. NN[a]	overall	active	inactive
1	93	96	89
3	91	92	89
5	91	92	89
7	85	92	76

[a] Number of nearest neighbors included in classification.

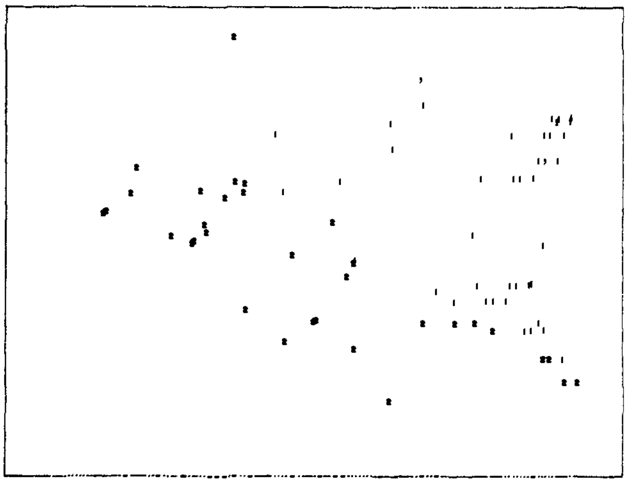


Figure 4. Data plotted in the first two principal components of the second set of 10 descriptors. 1's are active, 2's are inactive.

With use of this modified set of 10 descriptors, a linear discriminant was generated that correctly classified all 88 of the compounds into their assigned potent/nonpotent classes. KNN correct classifications were slightly higher in all cases than for the original set of descriptors (Table V). The first two principal components of this data are plotted in Figure 4 and represent 49% of the total variance of the data. The classes are well-separated in this plot. In fact, most of the discriminatory information in this data is contained in these first two principal components. A linear discriminant, that classified all but four of the 88 compounds could be obtained with these first two principal components. Of these four, one was compound **92**, whose inactive classification is in question. Two of these were compounds **35** and **38**, whose activities (90 and 85, respectively) are very close to the cutoff and whose structures differ from inactive compounds only by the length of the acid esterifying the 21-position. These high levels of correct classifications in the reduced space of the two principal components reinforce the nonradomness of the results for the 10-dimensional space.

Figure 5 is the same plot as Figure 4, but the activities of the compounds are plotted semiquantitatively, the highest as "A" and the lowest as "V". Even finer activity structure can be seen in this plot. Most of the compounds with a very low activity, "V", fall in the same region of the plot. Most of the very active compounds, "B", "C", "D", and "E", also lie in one area of the plot.

Prediction. As noted above, 34 compounds were initially excluded from the analysis because of lack of information about their activities. Further examination of the original literature confirmed the classifications of 11 of these, which are listed in Table VI. The other 23 either had no listed activity information or had activities that

[24] Lorenzetti, O. J. *Curr. Ther. Res.* **1979**, *25*, 92–103.

Table VI. Prediction Results for the 10-Descriptor Set

				KNN[d]				multiple discriminants[g]	
no.	compound[a]	class[b]	1st NN[c] distance	1	3	5	7	trial 1[e]	trial 2[f]
1	*20	1	2.032	2	2	2	2	8	0
2	22	2	3.782	2	1	1	1	0	94
3	X 23								
4	*65	1	2.983	1	1	1	1	100	35
5	80	1	1.233	1	1	1	1	100	100
6	*87	1	2.453	1	1	1	1	0	0
7	*107	2	2.031	2	2	2	2	11	9
8	115	1	2.344	1	1	1	1	44	3
9	*77	2	2.031	2	2	2	2	11	9
10	X 17								
11	18	1	2.433	1	1	1	1	0	0
12	19	1	3.102	2	2	2	1	100	70
13	*30	1	2.603	1	1	1	1	0	0
14	31	2	1.682	2	2	2	2	0	2
15	*32	1	1.912	1	1	1	1	0	0
16	37	2	1.995	2	2	2	2	0	1
17	39	1	0.369	1	2	2	2	0	0
18	40	1	3.194	1	1	1	1	16	5
19	55	1	2.816	2	2	2	2	100	53
20	66	1	2.555	1	1	1	1	100	88
21	69	2	1.617	2	2	2	2	0	0
22	*70	1	1.826	1	2	2	2	35	2
23	71	1	1.675	1	1	1	1	70	56
24	72	1	0.237	1	1	1	1	0	0
25	85	1	0.000	1	1	1	1	0	0
26	96	1	3.191	1	1	1	1	66	2
27	*99	2	0.139	1	2	2	2	6	57
28	101	2	1.744	2	2	2	2	0	0
29	104	2	0.496	2	1	1	1	100	90
30	109	1	2.555	1	1	2	2	99	0
31	*110	2	1.841	2	2	2	2	9	3
32	116	1	0.237	2	2	2	2	100	95
33	117	1	1.987	2	2	2	2	0	0
34	120	2	0.000	2	2	2	2	0	0
35	*123	1	2.775	1	1	1	1	6	0
36	*124	1	0.005	1	1	1	1	0	0
37	*125	2	0.000	2	2	2	2	0	0

[a] * means certain classification; X—no available data—was not used. [b] 1 is active, 2 is inactive. [c] Nearest neighbor. [d] Results reported for 1, 3, 5, and 7 nearest neighbors. [e] 93 included in analysis. [f] 93 excluded from analysis. [g] Number of times misclassified for 100 discriminants generated.

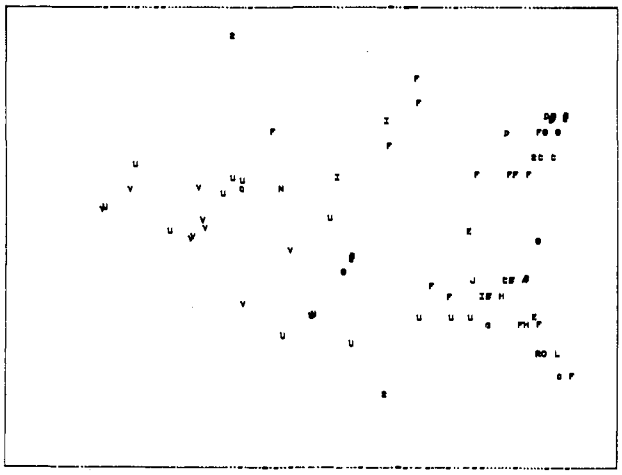


Figure 5. Figure 4 with semiquantitative rather than class activity. "A" is most active, "V" is least active.

were borderline or ambiguous. In addition to the 11, a search of the literature yielded three additional compounds (**123**–**125**) that were similarly assayed. This yielded a set of 14 compounds of unambiguous classification and that were not used to develop either the descriptor set or the discriminant and so could be used to evaluate the predictive ability of the descriptor set.

Prediction of the activity of compounds not contained in the original training set is not a straightforward task. Several questions must first be answered. First, how general is the data in the data set? Is the entire range of substitution represented here? How far can the data in the training set be extrapolated?

The data set was chosen solely by the availability of data in the literature. No experimental design was involved, and there was no effort to represent all possible variations or all possible combinations of variations. It is very unlikely that the training set is a complete representation of all possible compounds of the backbone shown in Figure 1. A simple calculation reinforces this statement. For the descriptors that are used here, if only the extremes of each descriptor were represented, 2^{10} or 1024 compounds would be required. This does not account for all the compounds with intermediate values, however, and even within the set of substituents and variations contained in this data set, 150 000 possible combinations could result. It is unrealistic to assume that these 150 000 compounds have been well-represented by 88 almost randomly chosen compounds. Any of those 150 000 could be represented in this space, however, and a classification could be made based on its discriminant score or nearest-neighbor vote even if it were very different from the other compounds. Figure 6 is a pictorial representation of this problem. Of the entire possible descriptor space, only a portion has been represented by the compounds in our training set. The X's and O's represent training set compounds of different classes, the boxes represent prediction set compounds, and the lines represent two of many discriminants that could correctly classify all of the compounds in the training set.

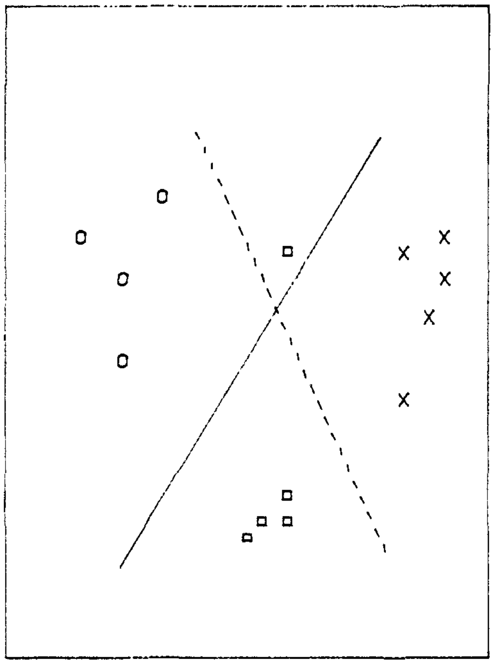


Figure 6. Hypothetical data space. X and O represent compounds of different classes. Squares represent new compounds whose activities are to be predicted.

This prediction set compounds lie in regions that are not populated by training set compounds. Their positions would cause their predicted activity to differ, depending on which discriminant was used. Since their is no information about the region that these compounds are in, any prediction of their activity would be fortuitous.

If the compounds used in this analysis were a sample from a larger statistical population, then perhaps the data space could be described by parameters calculated from that sampling. We question, however, whether a statistical approach is warranted in SAR problems. If no limits are placed on the acceptable substitutions at any given site, it is unlikely that SAR problems can be thought of in a statistical sense. At any given site of substitution, there will be no one substituent that is the mean substituent; rather, there are an infinite range of substitutions. Furthermore, the entire data space could be uniformly populated in the space of the descriptors used in this study. There will not necessarily be a separate and distinct clustering of activity classes in different regions of the space when it is completely populated by all possible compounds. There may be a gradual merging of the classes, or several areas of high or low activity. Since minor structural variations can often cause large changes in biological activity, an area of very high activity could exist immediately adjacent to one of very low activity. Because of this, a statistical approach might not always be the best means of investigating SAR. Instead, methods that could establish and uncover trends between the individual compounds may be required.

Since the compounds in the data set might not be representative of all the compounds for which predictions are to be made, two problems must be faced before prediction can be done. First, nonparametric methods of generating LDFs do not yield unique discriminants. Except for special cases, many different discriminants could be developed for a separable data set. Which is to be used? The second problem is the extent of extrapolation. Any prediction of the activity of a compound that is not included in the data set requires some amount of extrapolation or interpolation of the data in the training set. How similar do predicted compounds have to be to the training set compounds, and how is this similarity to be assessed?

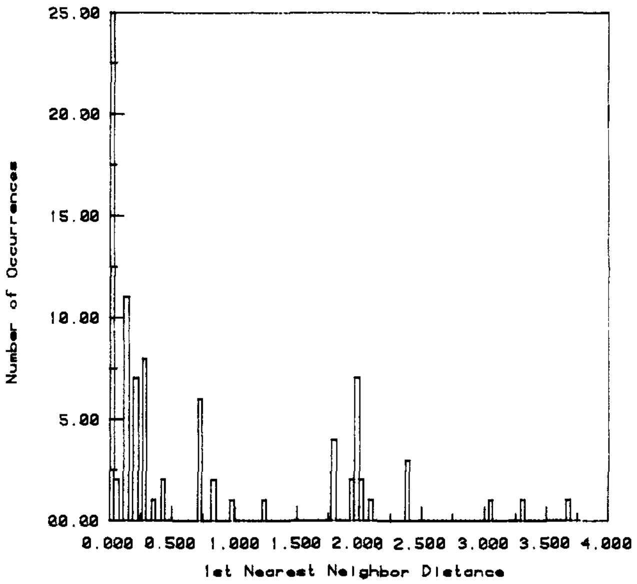


Figure 7. Histogram of the first nearest neighbor distances for the 88 training set compounds in the 10-dimensional space.

One way to develop an unambiguous discriminant is to centralize it between the points from the different classes by specifying that it have a thickness. A dead zone can be specified on either side of the hyperplane defined by the discriminant. If any patterns lie in that dead zone, they are considered to be misclassified, even though they may be on the proper side of the plane. If the size of this dead zone is maximized, this has the effect of forcing the discriminant to lie midway between the two classes. There is no guarantee that this centralized discriminant is the best one, however. In actuality, this method is simply a means for dealing with the lack of data that would define a unique discriminant. Another way to deal with this problem is to generate many different discriminants and apply each to the prediction set. Any compound whose activity is predicted differently by different discriminants should be called into question. Such variation in predicted activity would indicate that this compound would exist in a region of the space that was not well-defined by the training set. In such a case, the prediction of the activity of such a compound might not be justified.

One way to ensure that a prediction compound's structure does not deviate too far from those of the training sets compounds is to predict the activities only of those compounds that have substitutions the same as or similar to those compounds used in the training set. Certainly, there is no justification for predicting the activity of compounds whose structures differ greatly. The method that Bodor et al. used to choose compounds for analysis and visual examination of the data assured such structural homogeneity for these prediction compounds.

Another level-of-similarity check is to use distance measures to at least ensure that the prediction compounds are in the same region of space as the training set compounds. If a compound lies far from the remainder of the data, it is in a region of the data space that is poorly represented by the training set. Many PR methods, such as linear discriminants and KNN analysis, can assign a classification to such a compound, but prediction by extrapolating beyond the boundaries of known data is risky.

We have chosen to examine nearest-neighbor distances as a means of assessing similarity. A histogram of the first nearest-neighbor distances for the 88 training set compounds is shown in Figure 7. Sixty-four of the 88 distances are below 1.0 for the autoscaled data; all but three are below 2.5. The first NN distances for the prediction set compounds are shown in column four of Table VI under the heading 1st NN distance. Comparison of these distances to the histogram in Figure 7 shows that most of the 14 prediction compounds whose activities are fairly certain are, by coincidence, relatively far from their first nearest neighbors. If the activity of these compounds were truly unknown, this should prompt caution in the interpretation of predictive classifications since many of these compounds might not be close enough to training set compounds to allow reliable prediction results.

Three different methods were used for prediction studies: LDFs, KNN, and PC plots. One hundred discriminants were generated to serve both for prescreening the prediction compounds, as noted above, and for their prediction results. Each was used to predict the activity of the compounds that were reserved for prediction, the 14 that had reasonably certain activity as well as the 23 others whose activity was less certain. Column six of Table VI (multiple discriminants, trial 1) shows the results for the 100 predictions. Of the 37 compounds, 12 were predicted differently with the different discriminants. These compounds might be thought of as being in a gray area that was not sufficiently represented by the training set. In our prediction, we were fortunate enough to have the activities available to us for interpretation. In a case of blind prediction, compounds such as this whose predicted activity varies with different discriminants might be flagged for further investigation. Of the other prediction compounds, 17 were consistently predicted correctly and eight were consistently misclassified. Of the 14 compounds of unambiguous classification, one was consistently misclassified, seven had varying activities, and six were classified correctly by all 100 of the discriminants.

As noted earlier, the classification of compound **93** might be in error. This compound was found to be troublesome throughout the analysis. In order to determine its influence on the prediction studies, it was removed from the 88-member training set, 100 more discriminants were generated, and the predictions were redone. The results are shown in column seven of Table VI (multiple discriminants, trial 2). Removal of pattern **93** decreased the total incidence of misclassifications considerably. Only one compound was misclassified by all 100 of the discriminants. The predicted activity for 19 of the compounds varied between discriminants; for eight of the compounds, the number of misclassifications was low. Seventeen compounds were predicted consistently correctly. Of those 14 prediction compounds of unambiguous activity, in all but one case (compound **99**) the number of misclassifications were decreased; eight were consistently predicted correctly, and six had varying activities. Of those six, only one was predicted incorrectly by more than 10 of the 100 discriminants. It appears that compound **93** had a large effect on discriminant development. Both the PC plot of Figure 4 and KNN results showed this compound to be far removed from the rest of the data points. This may account for its high leverage in the discriminant development. These results might indicate that compound **93** was assigned to the incorrect activity class.

The prediction results using KNN analysis are shown in column five of Table VI (KNN). Of the 14 compounds with certain classification, two were misclassified by all levels of nearest-neighbor classification. Both of these were far from their nearest neighbors, however, and so the validity of their predictions would be questioned in a blind study. One of these, compound **70**, was assayed with use of a cream base rather than an alcohol base for application. Such variation in procedure has been shown to affect results,[25] but since compound **70** did not have a borderline activity (activity = 360), it was kept in the prediction set. Of the 23 other prediction compounds, seven were misclassified. Two of these had activities that would make them borderline cases. Compound **22** had an activity of 46, very close to the cutoff of Bodor et al. of 50. Compound **55** had an activity listed only as less than fluocinolone acetonide, which had an activity of 100. Both of these were far from their nearest neighbors.

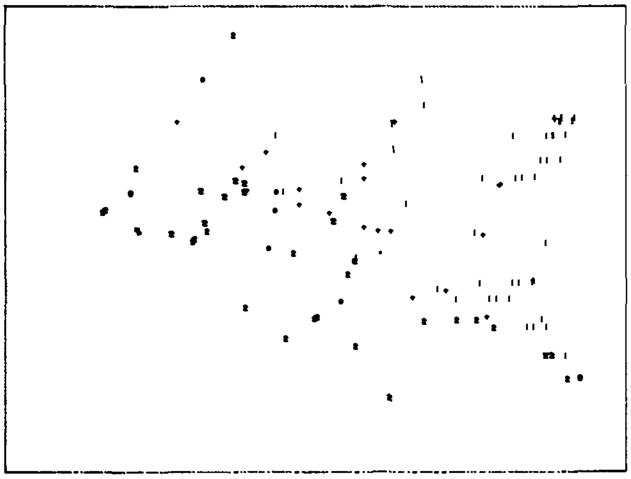


Figure 8. Figure 4 with active prediction compounds plotted as "+" and inactive prediction compounds plotted as "*".

Table VII. KNN Results for the 7-Descriptor Set

	percentage correctly classified		
no. NN[a]	overall	active	inactive
1	97	96	97
3	92	92	92
5	90	92	86
7	85	94	73

[a] Number of nearest neighbors included in classification.

Prediction from the PC plot can be made through visual inspection of Figure 8, which is Figure 4 with the prediction compounds plotted as "+", active, and "*", inactive. Of the 14 compounds, all but two (**70** and **99**) could be unambiguously and correctly classified by their location within the plot. Many of the remaining 23 prediction compounds lay somewhere between the areas occupied by the active and inactive classes. This corresponds well to the results from the LDF predictions which suggest that many of the prediction compounds are in areas that are ill-defined. Since many of those compounds have borderline or ambiguous activities, this plot provides further evidence of the existence of structure in the data.

Feature Section. The variance method of feature selection has been described previously.[26,27] This is a method for removing the descriptors that provide little or no discriminatory information. It was used to assess the discriminatory value of the 10 descriptors.

Three variables were identified as having no effect on the classification results: the indicator variable for the

[25] Barry, B. W.; Brace, A. R. *J. Invest. Dermatol.* **1975**, *64*, 418–422.

[26] Stuper, A. J.; Brugger, W. E.; Jurs, P. C. *Computer Assisted Studies of Chemical Structure and Biological Function*; Wiley-Interscience: New York, 1979.

[27] Zander, G. S.; Stuper, A. J.; Jurs, P. C. *Anal. Chem.* **1975**, *47*, 1085.

Table VIII. Prediction Results for the 7-Descriptor Set

				KNN[d]				
no.	compound[a]	class[b]	1st NN[c] distance	1	3	5	7	multiple discriminants[e,f]
1	*20	1	0.070	2	2	2	2	97
2	22	2	0.303	2	2	2	2	0
3	X 23							
4	*65	1	2.118	1	1	2	2	100
5	80	1	1.232	1	1	1	1	100
6	*87	1	2.142	1	1	1	1	0
7	*107	2	0.000	2	2	2	2	0
8	115	1	2.257	1	1	1	1	0
9	*77	2	0.000	2	2	2	2	0
10	X 17							
11	18	1	2.119	1	1	1	1	0
12	19	1	1.995	2	2	2	2	100
13	*30	1	2.403	1	1	1	1	0
14	31	2	1.682	2	2	2	2	0
15	*32	1	0.000	1	1	1	1	0
16	37	2	0.712	2	2	2	2	0
17	39	1	0.368	1	2	2	1	0
18	40	1	2.299	2	1	1	1	0
19	55	1	1.995	2	2	2	2	100
20	66	1	0.303	2	2	2	2	100
21	69	2	1.276	2	2	2	2	0
22	*70	1	1.826	1	2	2	2	56
23	71	1	1.674	1	1	1	1	0
24	72	1	0.237	1	1	1	1	0
25	85	1	0.000	1	1	1	1	0
26	96	1	0.358	2	2	2	2	100
27	*99	2	0.140	1	2	2	1	28
28	101	1	1.434	2	2	2	2	0
29	104	2	0.496	2	1	1	1	100
30	109	1	0.358	2	2	2	2	100
31	*110	2	1.841	2	2	2	2	0
32	116	1	0.237	2	2	2	2	100
33	*117	1	1.195	2	2	2	2	0
34	120	2	0.000	2	2	2	2	0
35	*123	1	2.775	1	2	2	2	0
36	*124	1	0.005	1	1	1	1	0
37	*125	2	0.000	2	2	2	2	0

[a] * means certain classification; X—no available data—was not used. [b] 1 is active, 2 is inactive. [c] Nearest neighbor. [d] Results reported for 1, 3, 5, and 7 nearest neighbors. [e] **93** excluded from analysis. [f] Number of times misclassified for 100 discriminants generated.

acetonide linkage, the log *P* at the 6-position, and the indicator variable for the 1,2-saturation. When these were removed from the analysis, the reduced set of seven variables could support a discriminant that correctly classified all 88 of the training set compounds. KNN results are listed in Table VII and are somewhat higher than for the full set of 10 descriptors. A PC plot is shown in Figure 9; good separation is still obvious.

This seven-dimensional data space was also used to predict the activity of the compounds that were excluded from the training set. As above, 100 discriminants were generated with use of the 87-member training set (compound **93** excluded). The results for the prediction are shown in Table VIII. Of the 14 compounds of certain classification, one was misclassified, 10 were classified correctly, and three had varying predicted activity. For the remaining 23 compounds, nine were misclassified by all 100 discriminants and 14 were classified correctly by all the discriminants. This reduced data space seems to be more restricted than the 10-dimensional space. Fewer of the prediction compounds had activities that varied between the different discriminants. Also, more compounds were misclassified. KNN results for the misclassified compounds are also listed in Table VIII. A histogram of the first nearest-neighbor distances of the 88 training set compounds is shown in Figure 10.

Much information is lost when the three descriptors are removed from the 10-descriptor set. Many unique compounds are equivalent in the reduced seven-dimensional data space. Examination of the raw data shows that all the structural variations affect activity; however, within the training set they are not responsible for switching activity between the two classes. Close examination of the structures and class assignments verified this. There were no instances of the removed variations causing a change in activity of a compound from one class to another.

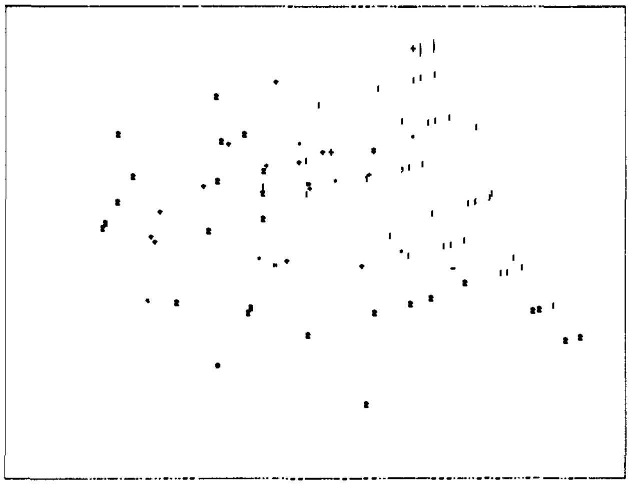


Figure 9. Data plotted in the first two principal components of the set of seven descriptors. 1's are active, 2's are inactive.

Many of the 37 prediction compounds were found to have borderline or ambiguous activities or were found to be in positions in the data space that were ill-defined. These contain new information, which was not present in

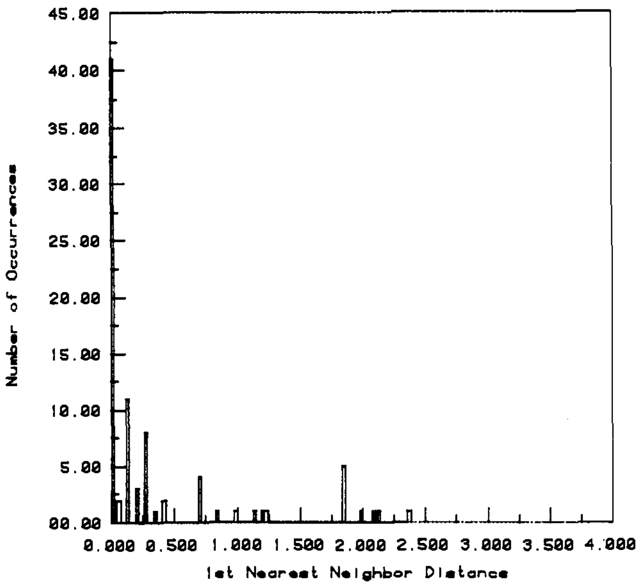


Figure 10. Histogram of the first nearest neighbor distances for the 88 training set compounds in the seven-dimensional space.

the 88-compound training set. In order to assess the effect of this new information on the two different descriptor sets, the entire set of 122 compounds was submitted to the LDF analysis for both the 10- and seven-descriptor sets. With the 10 descriptors, three compounds were misclassified. Two of these were **22** and **93**, which had questionable class assignments. (The activity of **22** was listed as 46, very close to the cutoff of 50.) The other was **116**, which did not have a verifiable activity. The best NLDF results for the seven-descriptor study misclassified 10 compounds. This increased misclassification may be due to the loss of the information from the three removed descriptors.

Conclusions

The work reported here illustrates that there is structure within the data space defined for the steroids by the 10 indicator and log *P* descriptors. Several different methods have shown that the potent and nonpotent compounds occupy different regions of the data space. The classes are sufficiently separated to support a linear discriminant that separates the two classes of compounds as defined by Bodor et al. The KNN method also provided a high correct classification success rate, and principal components plots showed structure not only between the classes but also within the classes. The prediction studies show that this data space could be useful for predicting the activity of new compounds that passed the prescreening criteria mentioned previously.

While the two-class representation is a useful means for exploring variations in activity of such data, it should not be interpreted too strictly. The activities of the compounds are continuous, even though the quantitative data have large errors. This is supported by the fact that many of the compounds that were consistently troublesome for the LDF methods were those with activities close to the class cutoff. The principal components plots do not rely on the two-class representation and can be useful for looking at finer structure within the data.

Future directions for this work could include expansion of the training set to include a broader range of structural variations. These could include both new substituents at the sites that were coded in this study as well as variations at other sites. Compounds could be included that would broaden the range of the variations in activities that were caused by those descriptors that were eliminated by the feature selection. Other structural descriptors could also be used to represent these structures.

While this particular data set seems to be well-described by the indicator and log *P* descriptors used here, this might be due to a limited variation of the substituents at some sites or to high correlations of these descriptors with properties other than those that are directly coded. These compounds might be described more precisely by other descriptors that would directly code for steric and/or electronic properties. New descriptors might be necessary if compounds containing new substituents are included in the analysis.

The work reported here demonstrates the advantages of physicochemical descriptors over substructural descriptors. First, far fewer descriptors were required for this study than for that of Bodor et al. Second, reduction of the data space through feature selection leads to chemically meaningful elimination of descriptors which can be used to investigate the generality of the data set. Third, variations within the sites of substution beyond the specific variations contained in the training set can be accommodated without adding new descriptors and redoing the entire study as would be required with a substructure-based approach.

These studies have shown that several different multivariate methods can be used in tandem to investigate a data space and verify the results of an LDF. Some of these methods can be used even when the classes are not linearly separable.

Through this work it has been suggested that a data space composed of structural data does not necessarily have to consist of separate clusters or a well-defined ordering of activities of compounds. For example, for the variables that were used in this work, the entire data space could be uniformly occupied by points that would represent chemical compounds, and no distinct clustering of the different classes need necessarily be present, even though such seems to be the case in this study. This attitude could affect the approach that is taken for the analysis of such problems and may suggest the need to develop new methods with which to more carefully investigate a data space that is to be used for SAR analysis.

This study has also been used as a medium to suggest the use of a multistage prescreen of compounds prior to use of LDFs for prediction of the activity of unknowns. Similarity of a predicted compound to the compounds in the training set should be ascertained prior to interpretation of any prediction results. Use of multiple discriminants can help to identify those compounds that may be similar to those in the training set but whose structural variations place them in a region of space that was not sufficiently represented by the training set.

Acknowledgment. We thank Maria Czerwinski for assistance with the early stages of this work. This research was supported by the United States Environmental Protection Agency under Cooperative Research Agreement No. CR 807831. The contents do not necessarily reflect the views of the Agency and no official endorsement should be inferred. The PRIME 750 computer used in these studies was purchased, in part, with the support of the National Science Foundation.

Registry No. **1**, 25122-46-7; **2**, 50-02-2; **3**, 5534-09-8; **4**, 50-03-3; **5**, 67-73-2; **6**, 50-23-7; **7**, 127-31-1; **8**, 52-21-1; **9**, 3093-35-4; **10**, 25122-47-8; **11**, 50-22-6; **12**, 50-24-8; **13**, 13609-67-1; **14**, 356-12-7; **15**, 37926-91-3; **16**, 37926-78-6; **17**, 913-42-8; **18**, 59198-70-8; **19**, 152-97-6; **20**, 2002-29-1; **21**, 312-93-6; **22**, 2022-55-1; **23**, 1524-88-5; **24**, 76-25-5; **25**, 5635-85-8; **26**, 2152-44-5; **27**, 987-24-6; **28**, 378-44-9; **29**, 53-36-1; **30**, 84099-84-3; **31**, 84099-85-4; **32**, 4524-39-4; **33**, 37926-75-3; **34**, 37926-79-7; **35**, 52619-01-9; **36**, 52510-15-3; **37**, 53-33-8; **38**, 56933-60-9; **39**, 84108-26-9; **40**, 49697-38-3; **41**, 37926-90-2; **42**, 2240-28-0; **43**, 5534-12-3; 44, 23640-98-4; **45**, 29379-85-9; **46**, 23640-97-3; **47**, 23641-10-3; 48, 124-94-7; **49**, 1177-87-3; **50**, 37792-94-2; **51**, 15180-00-4; **52**, 37926-92-4; **53**, 2693-06-3; **54**, 4419-39-0; **55**, 1597-82-6; **56**, 23641-05-6; **57**, 25122-48-9; **58**, 302-25-0; **59**, 37926-95-7; **60**, 37927-21-2; **61**, 37926-89-9; **62**, 5534-14-5; **63**, 75883-07-7; 64, 57524-89-7; **65**, 103836-56-2; **66**, 638-94-8; **67**, 514-36-3; **68**, 83-43-2; 69, 84099-87-6; **70**, 52510-28-8; **71**, 25092-07-3; **72**, 84099-88-7; **73**, 37927-23-4; 74, 23641-03-4; **75**, 37792-93-1; **76**, 23674-85-3; **77**, 806-29-1; **78**, 37792-98-6; **79**, 37926-93-5; **80**, 382-67-2; 81, 641-77-0; **82**, 5593-20-4; **83**, 37926-96-8; 84, 4351-48-8; **85**, 23641-04-5; **86**, 37872-65-4; **87**, 103836-58-4; 88, 29379-86-0; **89**, 23640-96-2; **90**, 42363-29-1; **91**, 53-34-9; **92**, 28971-62-2; **93**, 2802-11-1; **94**, 37926-76-4; **95**, 84099-92-3; **96**, 749-69-9; **97**, 23674-86-4; **98**, 25450-34-4; **99**, 16463-74-4; **100**, 2668-66-8; **101**, 82034-75-1; **102**, 5534-13-4; **103**, 3863-59-0; **104**, 84099-93-4; **105**, 23641-07-8; **106**, 84099-94-5; **107**, 807-38-5; **108**, 23641-01-2; **109**, 630-44-4; **110**, 25122-41-2; **111**, 23641-09-0; **112**, 23641-02-3; **113**, 5534-18-9; **114**, 42363-32-6; **116**, 103851-06-5; **117**, 72064-79-0; 118, 23641-08-9; **119**, 37792-97-5; **120**, 360-63-4; **121**, 61774-17-2; **122**, 25122-45-6; **123**, 23641-10-3.

Probes for Narcotic Receptor Mediated Phenomena. 13.[1] Potential Irreversible Narcotic Antagonist-Based Ligands Derived from 6,14-*endo*-Ethenotetrahydrooripavine with 7-(Methoxyfumaroyl)amino, (Bromoacetyl)amino, or Isothiocyanate Electrophiles: Chemistry, Biochemistry, and Pharmacology

Ralph A. Lessor,[†] Balbir S. Bajwa,[†,‡] Kenner C. Rice,[†] Arthur E. Jacobson,*[†] Richard A. Streaty,[§] Werner A. Klee,[§] Charles B. Smith,[⊥] Mario D. Aceto,[∥] Everette L. May,[∥] and Louis S. Harris[∥]

Section on Medicinal Chemistry, Laboratory of Chemistry, National Institute of Diabetes, and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, Maryland 20892, Laboratory of General and Comparative Biochemistry, National Institute of Mental Health, Bethesda, Maryland 20892, Department of Pharmacology, The University of Michigan, Ann Arbor, Michigan 48109, and Department of Pharmacology, Medical College of Virginia, Virginia Commonwealth University, Richmond, Virginia 23298. *Received July 1, 1985*

N-Allyl-, *N*-(cyclopropylmethyl)-, and *N*-propyl-*endo*-ethenotetrahydronororipavines (N-substituted 6,14-*endo*-etheno-4,5-epoxy-3-hydroxy-6-methoxymorphinans) were synthesized with potential acylating or alkylating moieties at the C-7 position (isothiocyanato, (bromoacetyl)amino, and (methoxyfumaroyl)amino) and examined in vivo for their narcotic agonist and antagonist activities and for their ability to interact with opioid receptors in vitro. The *N*-(cyclopropylmethyl)-substituted compounds were found to have the highest affinity for opioid receptors among these N-substituted compounds, although all of them were found to be reasonably potent narcotic antagonists in the mouse tail flick vs. morphine assay. Their in vivo potency ranged from $^1/_8$ to 4 times that of nalorphine on intravenous injection in mice. Rat brain membrane binding studies indicated that the compounds interacted with opioid receptors with potencies that ranged from 0.5 times that of morphine (**8c**, **9c**, and **10c**) to 0.017 that of morphine (**8b**). Among the compounds studied here, only the previously reported isothiocyanato compound (**10c**) and (methoxyfumaroyl)amino compound (**8c**) interacted irreversibly and selectively with μ or δ opioid receptors, respectively, in assays using NG108-15 neuroblastoma–glioma hybrid cells and/or in a rat brain membrane preparation. Both **8c** and **10c** were found to interact irreversibly, to a limited extent, with κ opioid sites in rat brain membranes in which the μ and δ opioid receptors were depleted by interaction with the μ-selective irreversible ligand BIT and the δ-selective irreversible ligand FIT. Neither compound showed irreversible actions in the electrically stimulated mouse vas deferens preparation.

Irreversible ligands[2] for specific opioid receptors[3] are valuable tools for a number of purposes. For example, we have previously reported the characterization of a covalently labeled glycopeptide subunit of the δ opioid receptor using FIT, an opioid agonist which specifically acylates this receptor class.[4a] Using a more potent acylating analogue, we have now purified this subunit to apparent homogeneity.[4b] Specific covalent modifying agents can also be utilized for the production of antibodies to drugs, and these can lead to antiidiotypic antibodies to the receptor.[5] Affinity columns for purification of receptors can be prepared with these selective modifying agents.[6] Autoradiographic mapping of receptor subtypes in brain sections[7] and the determination of the effect of receptor occupancy in individual neurons using electrophysiological techniques[8] can also be carried out using specific affinity ligands. For these reasons, we have been engaged in a program to identify a number of different affinity ligands that would be specific for each of the known, or purported, types of opioid receptors.

[†] National Institute of Diabetes, and Digestive and Kidney Diseases.

[‡] Deceased June 7, 1983.

[§] National Institute of Mental Health.

[⊥] The University of Michigan.

[∥] Virginia Commonwealth University.

[1] Part 12: Burke, T. R., Jr.; Jacobson, A. E.; Rice, K. C.; Silverton, J. V.; Simonds, W. F.; Streaty, R. A.; Klee, W. A. *J. Med. Chem.* **1986**, *29*, 1087.

[2] (a) Takemori, A. E.; Larson, D. L.; Portoghese, P. S. *Eur. J. Pharmacol.* 1981, *70*, 445. (b) Marynoff, B. E.; Simon, E. J.; Gioannini, T.; Govisser, H. *J. Med. Chem.* **1982**, *25*, 913 (see ref 7 and 10–12 therein). (c) Kolb, V. M.; Gober, J. R. *Life Sci.* **1983**, *33*, 419. (d) Sayre, L. M.; Larson, D. L.; Fries, D. S.; Takemori, A. E.; Portoghese, P. S. *J. Med. Chem.* **1983**, *26*, 1229. (e) Sayre, L. M.; Takemori, A. E.; Portoghese, P. S. *J. Med. Chem.* **1983**, *26*, 503. (f) Archer, S.; Seyed-Mozaffari, A.; Osei-Gyimak, P.; Bidlack, J. M.; Abood, L. G. *J. Med. Chem.* **1983**, *26*, 1775. (g) Hallermayer, K.; Harmening, C.; Merz, H.; Hamprecht, B. *J. Neurochem.* **1983**, *41*, 1761. (h) Kolb, V. M.; Hua, D. H. *J. Org. Chem.* **1984**, *49*, 3824. (i) Fang, S.; Bell, K. H.; Portoghese, P. S. *J. Med. Chem.* **1984**, *27*, 1090.

[3] (a) Martin, W. R. *Pharmacol. Rev.* **1967**, *19*, 463. (b) Iwamoto, E. T.; Martin, W. R. *Med. Res. Rev.* **1981**, *1*(4), 411. (c) Lord, J. A. H.; Waterfield, A. A.; Hughes, J.; Kosterlitz, H. W. In *Opiates and Endogeneous Opioid Peptides*; Kosterlitz, H. W., Ed.; North-Holland: Amsterdam, 1976; pp 275–280.

[4] (a) Klee, W. A.; Simonds, W. F.; Sweat, F. W.; Burke, T. R.; Jacobson, A. E.; Rice, K. C. *FEBS Lett.* **1982**, *150*, 125. (b) Simonds, W. F.; Burke, T. R., Jr.; Rice, K. C.; Jacobson, A. E.; Klee, W. A. *Proc. Natl. Acad. Sci. U.S.A.* **1985**, *82*, 4974.

0022-2623/86/1829-2136$01.50/0