JOURNAL OF PHARMACEUTICAL SCIENCES

JUNE 1978
VOLUME 67 NUMBER 6

RESEARCH ARTICLES

Structure–Activity Studies of Barbiturates Using Pattern Recognition Techniques

A. J. STUPER * **and P. C. JURS** ˣ

Received August 31, 1977, from the *Department of Chemistry, Pennsylvania State University, University Park, PA 16802.* Accepted for publication September 29, 1977. *Present address: Research Laboratories, Rohm and Haas Co., Spring House, PA 19477.

Abstract □ The relationship between molecular structure and duration of depressant effect for barbiturates was investigated. A data set of 160 5,5′-disubstituted barbiturates with various acyclic substituents was coded using 47 numerical descriptors including fragments, substructures, environmental descriptors, and molecular connectivity indexes. All descriptors were derived directly from the connection tables of the barbiturates. Using an interactive error-correction feedback algorithm, linear discriminant functions were developed that could dichotomize the data set with respect to several thresholds separating longer from shorter acting compounds. Feature selection was used to focus on the relatively few structural descriptors sufficient to support linear separability. For three specific thresholds, nine, 11, and nine descriptors were sufficient. The importance of these descriptors and the utility of the technique are discussed. Predictive abilities of approximately 94% were obtained for known barbiturates of the same general molecular types.

Keyphrases □ Structure–activity relationships—various barbiturates, prediction of duration of depressant effects using pattern recognition system based on structural descriptors □ Barbiturates, various—prediction of duration of depressant effects using pattern recognition system based on structural descriptors □ Pattern recognition system—based on structural descriptors, applied to prediction of duration of depressant effects of various barbiturates

The study of systematic alteration of molecular structure and its effect on biological activity has been of interest since the first drug was discovered. The pursuit of compounds with enhanced activity often requires choices between several possible alternatives made by reliance on the intuition and experience of the chemist. Although intuitive procedures have produced many new and useful compounds, they are not necessarily the optimal method of enhancing biological activity. Recently, there have been attempts to quantify this procedure through methods that predict a compound's action based on the results obtained for compounds of similar structure. The information provided by such methods can be used as an indication of whether a particular alteration holds promise. The best known of these procedures is Hansch analysis (1, 2).

Hansch analysis provides a means of relating the change in the level of biological activity with changes in the physical and chemical properties of a series of drug molecules. This analysis is accomplished by fitting the relative biological responses, A_i, to an equation of the form:

$$\log (A_i) = a(\log P)^2 + b \log P + \rho\sigma + cE_s + d \qquad \text{(Eq. 1)}$$

where P is the octanol–water partition coefficient, σ is the Hammett constant for the substituents under study, E_s is the Taft steric parameter, and a, b, c, d, and ρ are constants determined by multiple linear regression.

Recent reports indicated an alternative method of elucidating structure–activity relationships. Hansch *et al.* (3) discussed application of hierarchial clustering techniques to substituent selection. Ting *et al.* (4) reported correlations between the low-resolution mass spectra of 66 drugs and their activity as sedatives or tranquilizers.

Applications of pattern recognition to investigations of structure–activity relations have been reported by using substructural parameters as descriptors of biological action (5–7). Other examples of using structurally derived parameters in studies involving pattern recognition also exist (8, 9). Several methods of pattern recognition were applied to a set of compounds of accepted therapeutic utility, and the application of pattern recognition methods to structure–activity studies of pharmaceuticals was discussed (10, 11). While objections to some methods used to describe the data sets have arisen (12–14), there is almost universal agreement that a compound's activity is related to its structure.

The present study demonstrates the application of an interactive pattern recognition system to the development of rules that predict the duration time of barbiturate action on the basis of information provided solely from the molecular structure. The results of classification are used further to deduce which of the given parameters are most effective in the determination of these rules. Also, the potential of the method for augmentation of chemical intuition is discussed.

Table I—List of Compounds Forming the Data Set

Compound	R_2	Duration, min	Compound	R_2	Duration, min
		$R_1 = CH_3$			
1	$(CH_3)_3CCH$	580	2	$CH_3(CH_2)_5$	260
3	$CH_3(CH_2)_3CH(CH_3)$	227	4	$CH_3(CH_2)_3CH(CH_3CH_2)CH_2$	223
5	$H_2C{=}C(CH_3)$	60	6	$CH_3CH{=}C(CH_3)$	120
7	$CH_3CH_2HC{=}C(CH_3)$	60	8	$CH_3HC{=}C(CH_3CH_2)$	60
9	$CH_3(CH_2)_2HC{=}C(CH_3)$	60	10	$(CH_3)_2CHC{=}C(CH_3)$	36
11	$CH_3(CH_2)_3HC{=}C(CH_3)$	24	12	$CH_3CH_2SCH_2$	330
13	$CH_3(CH_2)_3SCH_2$	150			
		$R_1 = CH_3CH_2$			
14	CH_3CH_2	1400	15	$CH_3CH_2CH_2$	1140
16	$CH_3CH(CH_3)$	1520	17	$CH_3CH_2CH_2CH_2$	450
18	$CH_3CH(CH_3)CH_2$	540	19	$CH_3CH_2CH(CH_3)$	600
20	$CH_3CH_2CH_2CH_2CH_2$	220	21	$CH_3CH_2CH(CH_3)CH_2$	190
22	$(CH_3)_3CCH_2$	200	23	$CH_3CH_2CH(CH_3CH_2)$	300
24	$CH_3(CH_2)_5$	45	25	$CH_3(CH_2)_2CH(CH_3)CH_2$	210
26	$CH_3CH_2C(CH_3)_2CH_2$	60	27	$CH_3(CH_2)_3CH(CH_3)$	90
28	$CH_3CH_2CH(CH_3CH_2)CH$	300	29	$CH_3(CH_2)_6$	120
30	$(CH_3)_2CHCH_2CH(CH_3)CH_2$	54	31	$(CH_3)_2CH(CH_2)_2CH(CH_3)$	50
32	$CH_3CH_2CH(CH_3)CH_2CH(CH_3)$	74	33	$CH_3(CH_2)_2CH(CH_3CH_2CH_2)$	81
34	$CH_3(CH_2)_2CH(CH_3)CH_2CH_2CH_2$	60	35	$CH_3CH_2CH(CH_3)CH_2CH(CH_3)CH_2$	60
36	$CH_3(CH_2)_3CH(CH_3CH_2)CH_2$	75	37	$CH_3(CH_2)_4CH(CH_3CH_2)$	60
38	$CH_3(CH_2)_5CH(CH_3)$	150	39	$CH_3(CH_2)_2CH(CH_3)CH(CH_3CH_2)CH_2$	240
40	$(CH_3)_2CH(CH_2)_2CH(CH_3CH_2)CH_2$	120	41	$H_2C{=}CH$	288
42	$H_2C{=}C(CH_3)$	150	43	$CH_3CH_2HC{=}CH$	18
44	$CH_3HC{=}C(CH_3)$	180	45	$(CH_3)_2C{=}CH$	240
46	$CH_3(CH_2)_2HC{=}CH$	96	47	$CH_3CH_2HC{=}C(CH_3)$	24
48	$(CH_3)_2CHHC{=}CH$	12	49	$CH_3HC{=}C(CH_3CH_2)$	42
50	$CH_3(CH_2)_2HC{=}C(CH_3)$	72	51	$CH_3(CH_2)_3HC{=}C(CH_3)$	6
52	$CH_3CH_2HC{=}C(CH_3CH_2CH_2)$	6	53	$H_2C{=}CHCH(CH_3)$	720
54	$H_2C{=}C(CH_3)CH_2$	326	55	$CH_3HC{=}CHCH_2$	372
56	$CH_3CH_2OCH(CH_3)$	460	57	$CH_3(CH_2)_2OCH(CH_3)$	150
58	$CH_3(CH_2)_3OCH(CH_3)$	150	59	$(CH_3)_3CCH_2OCH(CH_3)$	75
60	$CH_3CH_2OC(H_2C)$	200	61	$(CH_3)_3COC(H_2C)$	63
62	$CH_3(CH_2)_2SCH_2$	59	63	$(CH_3)_2CHSCH_2$	139
64	$H_2C{=}CHCH_2SCH_2$	117	65	$CH_3(CH_2)_3SCH_2$	66
66	$CH_3(CH_2)_4SCH_2$	75	67	$(CH_3)_3CCH_2SCH_2$	37
68	$CH_3(CH_2)_2CH(CH_3)SCH_2$	62	69	$CH_3(CH_2)_5SCH_2$	15
70	$(CH_3CH_2)_2CHCH_2SCH_2$	22	71	$CH_3CH_2SCH(CH_3CHCH_3)$	12
72	$CH_3(CH_2)_3SCH(CH_3)$	34	73	$CH_3(CH_2)_3SCH(CH_3CH_2)$	52
74	$CH_3(CH_2)_4SCH(CH_3)$	28	75	$(CH_3)_3CCH_2SCH(CH_3)$	41
76	$CH_3(CH_2)_2CH(CH_3)$	180			
		$R_1 = CH_3CH_2CH_2$			
77	$CH_3CH_2CH_2CH_2CH_2$	4	78	$CH_3CH_2CH(CH_3)CH_2$	165
79	$CH_3(CH_2)_5$	1	80	$CH_3(CH_2)_6$	15
81	$CH_3HC{=}CH$	60	82	$CH_3CH_2HC{=}CH$	18
83	$(CH_3)_2CHHC{=}CH$	18	84	$H_2C{=}C(CH_3)$	168
85	$CH_3HC{=}C(CH_3)$	30	86	$CH_3CH_2HC{=}C(CH_3)$	18
87	$CH_3HC{=}C(CH_3CH_2)$	24	88	$H_2C{=}CHCH(CH_3)$	420
89	$H_2C{=}C(CH_3)CH_2$	300	90	$CH_3HC{=}CHCH_2$	120
91	$CH_3CH_2OCH(CH_3)$	162	92	$CH_3CH_2SCH_2$	150
93	$CH_3(CH_2)_3SCH_2$	76	94	$CH_3(CH_2)_3SCH(CH_3)$	35
95	$(CH_3)_2CHCH_2SCH(CH_3)$	45			
		$R_1 = (CH_3)_2CH$			
96	$(CH_3)_2CHCH_2$	25	97	$CH_3HC{=}CH$	36
98	$CH_3CH_2HC{=}CH$	36	99	$CH_3(CH_2)_2HC{=}CH$	18
100	$(CH_3)_2CHHC{=}CH$	12	101	$CH_3CH_2HC{=}C(CH_3)$	18
102	$CH_3C{=}C(CH_3CH_2)$	18	103	$H_2C{=}CHCH(CH_3)$	210
104	$CH_3HC{=}CHCH_2$	200	105	$CH_3CH_2SCH_2$	86
106	$CH_3(CH_2)_3SCH_2$	38			
		$R_1 = CH_3(CH_2)_3$			
107	$CH_3CH_2CH(CH_3)$	16	108	$(CH_3)_3C$	1
109	$CH_3HC{=}CH$	12	110	$CH_3CH_2HC{=}CH$	18
111	$H_2C{=}C(CH_3)$	90	112	$CH_2HC{=}C(CH_3)$	60
113	$H_2C{=}CHCH(CH_3)$	110	114	$CH_3HC{=}CHCH_2$	40
115	$(CH_3)_2C{=}CHCH_2$	30	116	$CH_3CH_2OCH(CH_3)$	120
		$R_1 = CH_3(CH_2)_3$			
117	$CH_3CH_2SCH_2$	74	118	$CH_3(CH_2)_3SCH_2$	95
		$R_1 = H_2C{=}CH$			
119	$CH_3CH_2CH_2CH_2$	288	120	$(CH_3)_3CCH$	192
		$R_1 = H_2C{=}CH(CH_3)$			
121	$H_2C{=}CHCH_2$	102	122	$(CH_3)_2CHCH_2$	90
123	$CH_3CH_2CH_2CH_2$	30	124	$(CH_3)_3CCH$	18
		$R_1 = CH_3HC{=}C(CH_3)$			
125	$H_2C{=}CHCH_2$	30			

Table I—(*Continued*)

Compound	R_2	Duration, min	Compound	R_2	Duration, min
		$R_1 = H_2C{=}CHCH_2$			
126	$CH_3(CH_2)_3CH(CH_3)$	108	127	$H_2CHCH(CH_3)$	456
128	$CH_3CH_2OCH(CH_3)$	300	129	$CH_3CH_2OC(CH_2)$	300
130	$CH_3(CH_2)_2OCH(CH_3)$	204	131	$(CH_3)_3CCH_2OCH_2$	900
132	$H_2C{=}C(CH_3)CH_2$	380	133	$CH_3CH_2SCH_2$	164
134	$CH_3(CH_2)_2SCH_2$	117	135	$CH_3(CH_2)_3SCH_2$	123
136	$CH_3(CH_2)_3SCH(CH_3)$	34	137	$(CH_3)_2CHCH_2$	162
138	$(CH_3)_3CCH_2$	96	139	$H_2C{=}CHCH_2$	880
140	$(CH_3)_2CH$	720	141	$CH_3(CH_2)_2CH(CH_3)$	150
		$R_1 = CH_3HC{=}CHCH_2$			
142	$(CH_3)_3CCH$	40	143	$CH_3(CH_2)_2CH(CH_3)$	66
144	$CH_3CH_2CH(CH_3)$	120	145	$(CH_3)_3CHCH_2$	45
		$R_1 = (CH_3)_2C{=}CHCH_2$			
146	$(CH_3)_2C{=}CHCH_2$	70	147	$CH_3CH_2CH(CH_3)$	120
		$R_1 = CH_3CH_2OCH(CH_3)$			
148	$(CH_3)_3CCH_2$	102	149	$CH_3(CH_2)_2CH(CH_3)$	108
		$R_1 = CH_3(CH_2)_2OCH(CH_3)$			
150	$CH_3CH_2CH_2CH(CH_3)$	300			
		$R_1 = CH_3SCH_2$			
151	$(CH_3)_2CHCH_2$	108			
		$R_1 = CH_3CH_2SCH_2$			
152	CH_3CH_2	143	153	$(CH_3)_2CHCH_2$	81
154	$CH_3CH_2CH(CH_3)$	61	155	$(CH_3)_3CCH_2$	8
156	$CH_3(CH_2)_2CH(CH_3)$	35			
		$R_1 = CH_3CH_2SCH(CH_3)$			
157	$CH_3(CH_2)_5$	12			
		$R_1 = H_2C{=}CHCH_2SCH(CH_3)$			
158	$(CH_3)_2CHCH_2$	28			
		$R_1 = CH_3(CH_2)_3SCH_2$			
159	$(CH_3)_2CHCH_2$	78	160	$CH_3CH_2CH(CH_3)$	69

METHOD OF APPROACH

The assumptions underlying the pattern recognition approach to structure–activity studies are similar to those of Hansch analysis. The factors that govern the activity of a compound are viewed as combinations of a molecule's electronic, steric, and lipophilic properties. It is felt, however, that the structure of the molecule is the overriding factor in the determination of a compound's physical properties and, therefore, its biological activity. Thus, it should be possible to decompose a structure into a set of descriptors that provide information correlating to biological activity.

This approach involves two problems: (*a*) how to develop parameters that relate structural modifications to changes in biological activity, and (*b*) what type of method to use in defining these relationships once sufficiently informative parameters are found.

Several types of descriptors that can be derived directly from the molecule structure are available. Methods for their development were reported previously (15).

In the present study, binary pattern classifiers were employed to develop a relationship between structure and biological action. Use of these classifiers in the form of a linear learning machine was described in detail elsewhere (16–18). Implementation involves representing the ith molecular structure as an n dimensional vector, $X_i = (x_1, x_2, x_3, \ldots, x_n)$, such that each component, x_j, is the value of one structural descriptor. Thus, each compound is represented as a point in n space whose position is determined by its structural descriptors. The assumption is that compounds of similar activity will cluster in the same general region of space. Earlier studies indicated that this clustering occurs for parameters described in Ref. 15, as well as those used in Hansch analysis (7, 8).

If two clusters can be separated from each other by a linear surface (hyperplane), they are said to be linearly separable. In practice, discrimination between clusters is made by calculating the dot products of the data vectors with a weight vector, W, normal to the surface of the hyperplane. All data vectors on one side of the plane will have a positive dot product; all those on the opposite side will have a negative dot product. A separating plane can be developed by choosing an initial weight vector and iteratively correcting it until all members of a cluster have the same dot product sign. Once such a surface is developed, it can be used to predict the cluster to which an unknown belongs.

Since any member of the data set not used to develop the discriminant surface is effectively an unknown, the data set itself can be used to estimate the predictive ability of the discriminant. This step is accomplished by dividing the data into two sets, one to develop the discriminant (training set) and one to test the ability of the discriminant to classify unknowns (prediction set) correctly.

The best measure of predictive ability is obtained by leaving out one compound and using the remaining compounds as the training set. The surface developed from the training set is used to predict the cluster into which the remaining (and, therefore, unknown) compound belongs. This procedure is continued until each member of the data set has been left out of the training set once. The predictive ability is the number of correct classifications divided by the total number of classifications. For a finite set, this method is considered the most unbiased estimator of predictive ability (19, 20). Approximations to this measurement can be made by repeating the process several times using a larger prediction set.

Since no assumptions regarding the distribution of the data are required, the use of a linear surface to discriminate between several classes (clusters) present in the data is a quite general approach. Such methods are termed nonparametric. Discussions of the capabilities and limitations of nonparametric methods of discriminant development can be found elsewhere (16–18).

The linear learning machine, as well as the descriptor development routines used in this study, is currently implemented in a general interactive pattern recognition package called ADAPT[1], coded in FORTRAN IV. A detailed discussion of the architecture of this system was given elsewhere (21).

DATA SET

The set of compounds consisted of 160 5,5′-disubstituted barbiturates (Table I) selected from a standard reference (22). These compounds range in molecular weight from 172 to 276 and have durations of action ranging from 10 to 1600 min. Administration was either intraperitoneal or subcutaneous, using mice, rats, or rabbits. The fact that this data set is heterogeneous in mode of administration and test species but can still be dealt with using pattern recognition methods illustrates one strength of the approach. The methods employed in pattern recognition will often allow study of incomplete, ill-defined, or otherwise imperfect sets of compounds, while many other more rigorous methods demand better quality data sets. The success enjoyed in analyzing this barbiturate data

[1] Executed on the Pennsylvania State University Department of Chemistry MODCOMP II/25 computer.

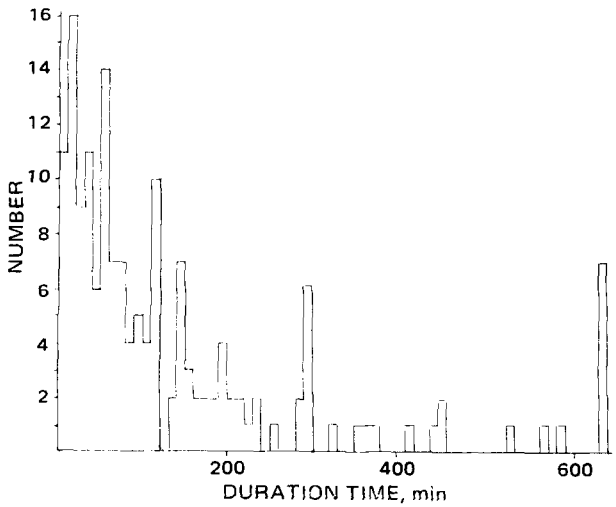

Figure 1—*Histogram of barbiturate duration times.*

set is meant to indicate how pattern recognition methods can be used; it is not meant to point out how to synthesize new barbiturates.

The compounds were grouped into classes according to the duration of depressant effect. These classes were formed by dividing the duration time, expressed in minutes, by 10. The resulting class designation was rounded up if the remainder was five or greater and down otherwise. Thus, a compound whose duration time was 227 min was placed in Class 23, whereas a compound having a duration time of 223 min was placed in Class 22. Compounds with a duration greater than 650 min were placed in Class 65. This approach resulted in a total of 65 different classes, distributed as shown in Fig. 1.

Four types of descriptors were employed for these studies; numeric fragment descriptors, substructural descriptors, environmental descriptors, and molecular connectivity descriptors (Table II). The descriptors were generated using the automated descriptor packages described previously.

While the nature of the atom, bond, and substructural descriptors is obvious, further comment is necessary regarding the environment and molecular connectivity descriptors. The environment descriptor takes into account how different parts of the molecule are connected by providing a measure of the local environment of a single atom fragment. This analysis is accomplished by combining the fragment's first and second nearest neighbors and their bonds into a single parameter that reflects the chemical environment around the fragment. Three types of environment descriptors are employed: the bond environment descriptor (BED), which uses only the number of bonds to calculate a descriptor value; the weighted environment descriptor (WED), which uses the type of bond in the calculation; and the augmented environmental descriptor (AED), which uses both the type of atom and type of bond in the calculation. Further discussion of these descriptors is given in Ref. 15.

The molecular connectivity descriptor provides a measure of the connectivity for the entire molecule. The concept was developed by Randic (23) and used in structure–activity studies (24, 25). A number of correlations between the molecular connectivity and several different physical parameters were described (23–29). The connectivity index is calculated directly from the connection table representation of the molecules as described in Ref. 25. In the present study, the simple index, the index corrected for rings, and the square of these indexes were used as descriptors. Ring correction was accomplished by subtracting from the simple index a value equal to the average of the contributions from all bonds contained in a ring. The descriptors were then multiplied by 10 and truncated to integer values.

Thus, the data set consists of 160 compounds, each coded with 47 descriptors. In no case does any one descriptor, or any binary combination of descriptors, contain sufficient information to classify the data successfully. Preprocessing of the raw data prior to classification consisted of autoscaling so that each descriptor had an average of zero and a standard deviation of 127. This method allowed the data to be truncated to integer values with a negligible loss of precision (recalculation after truncation yielded a standard deviation of 127 and a mean of 0 ± 0.17).

The learning machine requires that a constant-valued descriptor be added to the data set. A value of 250 was used because it provided for fast training and high predictive abilities. This parameter is discussed further in Ref. 30.

Table II—Molecular Structure Descriptors

Atom and Bond Descriptors			
1	Number of atoms	2	Number of bonds
3	Number of carbon atoms	4	Number of nitrogen atoms
5	Number of oxygen atoms	6	Number of single bonds
7	Number of double bonds	8	Length[a]

	Environment Descriptors		
	Atom Centered Fragment	**General[b]**	**Cyclic**
9–11	CH_3	A–C	—
12–14	CH_2	A–C	—
15–17	CH	A–C	—
18–23	C	A–C	A–C
24–26	O=	A–C	—
27–29	HC=	A–C	—
30–35	>C=	A–C	A–C

Substructural Descriptors					
36	CH_3CH_2	37	$CH(CH_3)CH_2$	38	CH_3
39	CH_2	40	CH_2CH_2	41	$CH_3CH_2CH_2$
42	CH	43	HC=		
Molecular Connectivity Descriptors[c]					
44	MC1	45	MC2	46	MC3
47	MC4				

[a] Length = 4* (number of single bonds) + 2* (number of double bonds). [b] A is the bond environment descriptor, B is the weighted environment descriptor, and C is the augmented environmental descriptor. [c] MC1 is the simple index, MC2 is the ring corrected index, MC3 is the square of the simple index, and MC4 is the square of the ring corrected index.

RESULTS

The duration of the barbiturate depressant effect is highly dependent on the conditions under which a compound is tested. The data compiled for this study represent a series of studies on different animals at different laboratories, so a large degree of variation within the data is expected. However, several series of compounds were tested as a group and, therefore, trends in the duration correlating to structural alterations may exist.

To account for these variations, any one classifier will develop a discriminant that answers the question: "Is the duration time less than x minutes?" Compounds within 30 min of this duration time are not used to develop the discriminant. With the class designations formed as noted previously, there are 61 possible ways of forming two clusters (longer or shorter than duration x) such that a gap of three classes lies between the clusters. Initial studies showed that it was possible to develop discriminants for each of these 61 cluster sets; however, only three such sets will be used to demonstrate the method.

Set I assigned all members in Classes 1–10 to the short duration cluster and 14–65 to the long duration cluster, Set II assigned classes 1–20 to the short duration cluster and 24–65 to the long duration cluster, and Set III assigned Classes 1–24 to the short duration cluster and 28–65 to the long duration cluster. With these three sets, discriminants can be developed to classify compounds as having a duration less than 100 min, less than 200 min, or less than 240 min. Compounds belonging to a class of longer duration would not be assigned to any of these duration regions.

One method of assessing the reliability of these discriminant functions is to subdivide each of the three sets such that each successive group contains more members in the prediction set and fewer members in the training set. These groups can be used to estimate the predictive ability and to determine which descriptors support the discriminant function's ability to separate the cluster of short duration barbiturates from those of long duration. If, within each set of clusters, the descriptors selected vary significantly and the predictive abilities are quite different, it would be clear that no clusters actually existed and that no relation between the structure and duration was found.

Results for the predictive ability tests and for feature selection using Descriptors 1–43 are shown in Table III. The portion of the data placed into the prediction set is indicated at the top of each column. Equal percentages of both the long and short duration clusters were taken to form these sets. The remaining members were placed into the training set. Ten such sets were formed for each percentage group. The highest predicting of these sets was used to select the features responsible for the discriminant's ability to classify the data.

Table III—Comparison of the Descriptors Retained for Each Cluster Set

	Set I				Set II				Set III			
Descriptor	Total	10%	15%	20%	Total	10%	15%	20%	Total	10%	15%	20%
1									×			
2												×
3	×											
5		×	×	×	×	×		×	×	×	×	×
6	×	×	×				×	×	×		×	
7	×	×	×	×	×	×	×	×		×	×	×
8					×				×			
9			×	×								
10		×	×				×					
11					×	×			×			
12	×											
14			×									
15	×	×	×	×	×	×	×	×	×			
16					×	×	×	×	×		×	
17								×				
19									×			
20					×							
21						×						
23					×		×		×			
27	×	×	×	×							×	
28		×	×							×		×
30								×	×		×	
32	×	×		×				×	×	×	×	×
33		×			×	×	×					
34			×	×	×		×		×	×	×	×
35	×			×		×	×	×	×	×	×	×
36	×	×	×	×								
37								×		×	×	×
38				×					×	×	×	×
39	×	×		×		×	×					
40					×			×	×	×	×	×
41			×									
42					×	×	×	×				
43	×		×		×	×	×	×				
Reference[a]												
Initial	—	100	95.5	96.6	—	100	100	96.8	—	100	95.4	96.9
Final	—	100	100	100	—	100	100	93.6	—	100	95.4	96.9
Total set[b]												
Initial	—	92.0	88.2	89.0	—	91.9	90.4	91.0	—	91.9	93.3	93.4
Final	—	92.7	91.8	92.4	—	94.4	93.0	92.9	—	95.0	95.4	95.3

[a] Predictive ability for feature selection reference set. [b] Average predictive ability for the 10 prediction sets within each percentage group.

Feature selection was accomplished using the variance feature selection method (30). The retained descriptors are indicated by an ×. The predictive ability is the average for all 10 sets before and after the feature selection process. Total refers to the results of feature selection using all members of each cluster. Reference refers to the results for the one prediction and training set used in feature selection. The members of this prediction set were never used to develop the discriminant function and, therefore, represent total unknowns.

The molecular connectivity descriptors (Descriptors 44–47) were not included in these initial studies to keep the ratio of compounds to descriptors above 3:1. This step is necessary to ensure that a nontrivial discriminant function is developed (31). To include Descriptors 44–47, a reduced set of the first 43 descriptors was chosen. This was accomplished for each set by pooling those descriptors from Table III that were selected three or more times. Using these as the initial descriptors, each set of clusters was feature selected using the variance method. The resulting descriptors represent a minimum set; that is, if any selected descriptors are excluded from the training process, a linear discriminant function that separates the data cannot be developed. Descriptors 44–47 were then added to these reduced sets, and each was once again subject to variance feature selection. The descriptors ultimately selected for each set of clusters is shown in Table IV. Predictive abilities were estimated using the leave-one-out procedure.

Table V lists the mean values, autoscale factors, and weight vectors for the Set I discriminant. To predict whether an unknown has an activity of less than 100 min, the descriptors from Table IV are generated and

Table IV—Descriptors Selected for Cluster Sets I–III

Set I		Set II		Set III	
Atom and Bond Descriptors		Atom and Bond Descriptors		Atom and Bond Descriptors	
Number of oxygen atoms Number of double bonds		Number of oxygen atoms Number of double bonds		Number of oxygen atoms	
Substructural Descriptors	Environment Descriptors[a]	Substructural Descriptors	Environment Descriptors[a]	Substructural Descriptors	Environment Descriptors[a]
CH_3CH_2	CH_3 (General B) HC (General A) >C= (General C) >C= (Cyclic A) HC= (General A)	CH_3CH_2 HC HC=	CH_3 (General C) HC (General B) >C= (Cyclic A) >C= (Cyclic C)	CH_3 CH_2CH_2 $CH(CH_3)CH_2$	HC (General A) HC = (General A) >C= (General C) >C= (Cyclic C)
Molecular Connectivity[b]		Molecular Connectivity[b]		Molecular Connectivity[b]	
MC2 Average 93.8% predictive ability		MC4 Average 92.9% predictive ability		MC4 Average 93.7% predictive ability	

[a] See footnote *b* of Table II. [b] See footnote *c* of Table II.

Table V—Weight Vector and Normalizing Factors for Cluster Set I

Descriptor	Mean Value	Mult/Sigma	Weight Vector
5	3.907	431.4560	−0.2197
7	3.527	220.8830	−0.4915
10	18.376	15.6835	0.0441
15	8.457	13.7709	−0.2994
27	5.527	16.4919	0.4415
32	112.648	11.4809	0.2787
33	45.994	104.7780	0.2545
36	1.333	158.7940	−0.1682
45	63.752	15.3582	0.5009
$n + 1$	250		0.0453

these values are scaled by subtracting the mean value for that descriptor, multiplying the result by the normalizing factor, and truncating the results to integers. The result is a nine-component vector. By using a value of 250 as the 10th component, the dot product between this vector and the weight vector is calculated. If the sign of the dot product is positive, the activity is less than 100 min.

The calculation for the barbiturate having R_1 = ethyl and R_2 = *sec*-pentyl is given as an example. This compound is not part of the original data set and, therefore, constitutes an unknown. The duration is reported to be 180 min (32). Calculation of the descriptors in Set I yields the vector X = (3, 3, 19, 17, 0, 106, 47, 2, 71). Normalizing this vector yields X_n = (−41, −116, 9, 117, −91, −76, 105, 105, −42). Adding the extra component and calculating the dot products yield −30.6. Since the sign of the dot product is negative, the duration is estimated as being greater than 100 min. Discriminant functions can be used to predict activities and unknowns, as done here, using a simple calculation performed with a desk calculator (if the descriptors can be hand calculated).

DISCUSSION

The fact that discriminants could be developed successfully for a data set as diverse and heterogeneous as this one indicates that information concerning the duration of depressant effect is contained in the structures of these compounds. While it is possible to develop discriminant functions that show chance correlations (31), the experiments performed indicate that such correlations were not responsible for the behavior of the discriminants.

The reliability of relations found using nonparametric discriminant analysis is a function of the discrimination ability of the classifier. Discrimination ability is a measure of the classifier's ability to find a separating discriminant function. In each set studied, the learning machine could find a discriminant that separated short duration members from long duration members. If chance correlations were responsible for this separation ability, the features selected for the members in the training set would not support a separating discriminant for members of the training and prediction set. For each different training set in Table III, the descriptors chosen were similar. Additional experiments showed that these descriptors would support a discriminant that separated all members of the training and prediction sets. Thus, the structural features intrinsic to the development of a discriminant for the training set are also intrinsic to the relations for the prediction set.

The predictive ability of a discriminant is dependent on how the discriminant was developed. The linear learning machine does not necessarily provide a discriminant that yields the best predictive ability. While a training set may be linearly separable, there is an infinite number of separating discriminants. Thus, even though it is possible to use the descriptors selected by the training set to develop a discriminant for the prediction set, that fact does not imply that such a function will perform well. The predictive ability is a gauge of how well a discriminant will classify the data not used in developing that function. When developed from several different training sets, a discriminant developing chance correlations would demonstrate low or variable predictive ability. The studies described in Table III show that decreasing the number of members used to develop the discriminant function does not substantially degrade its predictive performance.

The discriminants developed by the learning machine are quite general. Not only can they distinguish between different structures, but they also can describe differences in duration between congeners. Structures 17, 20, and 24 constitute a congeneric series of increasing alkyl chain length. The fact that discriminants could be developed for all 61 possible divisions of the data set indicates that these compounds can be distinguished. Similarly, a branched series such as Compounds 16, 19, and 27 is accounted for. The duration of structural isomers such as Compounds 24, 25, and 27 and between Compounds 53 and 54 is also described.

In this light, it is interesting to investigate some of the six compounds that could not be accounted for using these descriptors. Compound 29 is a member of the series of Compounds 14, 15, 17, 20, and 24 and its duration time might be expected to be less than that for Compound 24. Its duration, however, deviates from the order implied by these compounds since it is unexpectedly large. Most likely, the differences in its activity can be attributed to changes in lipophilic properties due to the sizable side chain. Similarly, Compound 38 belongs to the series of Compounds 16, 19, and 27. Its duration also deviates from that implied by the other members in the series. Similar arguments can be made for Compounds 5 and 44, which do not fit the pattern followed by the remainder of the data set.

The ability to identify quickly those compounds differing from the larger body of data is an advantage of this approach to structure–activity studies. Once these differences are identified, they can be used to gain further information concerning the action of these compounds.

The structural parameters used in these studies appear consistent with the observed properties of the barbiturates. The lack of any dominant structural feature indicates a lack of specificity for the receptor site with which the compounds interact. The descriptors chosen through feature selection indicate that properties of chain length and the extent of branching are the major influences on barbiturate duration. The amount of shielding of the 5-position may be responsible for many of the lipophilic properties (33). Descriptors 32, 33, and 35 were included in the final sets of features for the three thresholds. These environment descriptors would extend to the secondary position of R_1 and R_2 and could conceivably account for this shielding. Similarly, the molecular connectivity descriptors provide information on the degree of branching which, in turn, can be related to lipophilic properties.

While it was not the intent to develop a discriminant for all barbiturates, evidence of the utility of the discriminants developed was provided by prediction of the activity of the compound having R_1 = ethyl and R_2 = *sec*-pentyl. The duration time of this compound was correctly predicted to lie between 100 and 200 min by using the discriminant given in Table V in combination with those from Sets II and III. Clearly, once a discriminant has been developed and the descriptors generated, the actual prediction process is quite straightforward.

A question naturally arises concerning the possibility of using the parameters from pattern recognition analysis to produce structures of a specific activity. A direct path to this goal is not possible. Table VI can

Table VI—Statistics for Final Set of Descriptors Selected for Cluster Set I

	Mean		Standard Deviation			
Descriptor	Compounds below Threshold[a]	Compounds above Threshold[a]	Compounds below Threshold[a]	Compounds above Threshold[a]	Highest Value	Lowest Value
5	3.04	3.18	0.24	0.38	4	3
7	3.52	3.54	0.56	0.61	5	3
10	19.91	16.19	9.00	7.55	41	0
15	8.56	8.34	9.26	9.30	36	0
27	6.50	4.15	9.20	5.40	29	0
32	114.96	109.35	13.20	8.78	141	102
33	46.16	45.77	1.17	1.13	49	43
36	1.37	1.28	0.82	0.78	3	0
45	66.24	60.21	9.70	8.00	81	42

[a] Fifty-six compounds in long duration cluster, and 90 in low duration cluster.

be used to demonstrate this statement. The table gives pertinent statistics for each feature used to define the duration of the barbiturates with respect to Set I. Listed are the standard deviation of the descriptors and the numerical average of the descriptor values. The highest and lowest values give information concerning the range of descriptor values.

Although the average values for the two classes differ from each other, the standard deviation is larger than this difference. Therefore, the individual descriptors, while providing structural information, are not the sole indicators of activity. With atom, bond, and substructural descriptors, the average value can be related directly to the structural composition of the molecule. However, average values for the environment and molecular connectivity descriptors are difficult or impossible to interpret because they are related to the structure in a complex manner.

Average values indicate only the relative presence of a particular descriptor and cannot be construed as indicating the amount necessary for activity. A case in point is the number of oxygen atoms in the molecule. This number ranges between three and four. The barbiturate ring accounts for three of these atoms. A value greater than three accounts for the number in the side chains. The fact that, on the average, the class of longer acting molecules contains slightly more oxygen atoms does not imply that adding oxygen guarantees an increase in the activity. The number, placement, and chemical environment of an oxygen govern its effectiveness, not its mere presence. If the activity of a molecule is to be described by use of structural parameters, each must be viewed as a single contribution to, rather than the single indication of, that action.

Since structurally derived descriptors reflect the composition of the structure, they are interdependent. Changes in composition generally affect the value of several structural parameters simultaneously. Most notably affected are the environment and molecular connectivity descriptors. However, substructural content is also affected by slight alterations in the structure. Such alterations affect the placement of the molecule in the space formed by its descriptors and, therefore, affect the results of classification. This effect occurs because the biological activity expressed as a function of the molecular structure is a vector representation of the descriptors describing that structure.

The discriminant developed from pattern recognition analysis can be thought of as a transform, which maps a structure vector onto one of the two cluster regions. The reverse mapping cannot be accomplished directly. Alternatively, structural descriptors can be viewed as indicating the electronic, steric, and lipophilic properties of a molecule. No one descriptor is an effective gauge of all of these properties. Each is a component in their description. Knowing these properties does not allow the direct construction of active molecules.

Although the parameters used in the pattern recognition analysis cannot be used directly to construct active molecules, they can be used to predict the effectiveness of hypothetical structures. This prediction offers a pragmatic aid in the synthesis problem; *i.e.*, given a choice of molecules that appear to be equally likely candidates for synthesis, how does one optimize the chances of synthesizing the most active. If a data base exists that details past successes and failures, then a plausible solution is to use the data base to develop rules that estimate the activity of a candidate structure. Since pattern recognition develops rules that define "similarity," application of the methods as described in the preceding sections will aid in the synthesis decision.

Another manner in which these techniques could prove useful is for large-scale prescreening. The derivation of structural parameters is rapid enough to allow several thousand prospective structures to be described and tested using a discriminant developed from a set of compounds known to be active. Those compounds that the discriminant notes as being the highest acting can then be considered for further testing. Prediction results from the sets studied indicate that such classifiers can perform with a high degree of reliability.

The ultimate purpose in the study of effects of structural alterations on biological action is to produce new, more effective compounds. Useful tools are those that produce information pertinent to this problem. Historically, the chemist has used a structural diagram of the molecule as a gauge for altering the structure. The number and diversity of active compounds attest to the success of this approach. Since the effectiveness of this approach is dependent on the judgment of the chemist, use of mathematical techniques to augment these judgments may well increase the effectiveness of this procedure.

REFERENCES

(1) W. J. Dunn, *Ann. Rep. Med. Chem.*, **8**, 313 (1973).

(2) G. Redl, R. D. Cramer Tert., and C. E. Berkoff, *Chem. Soc. Rev.*, **3**, 273 (1974).

(3) C. Hansch, S. H. Unger, and A. B. Forsythe, *J. Med. Chem.*, **16**, 1217 (1973).

(4) K. L. Ting, R. C. T. Lee, G. W. A. Milne, M. Shapiro, and A. M. Guarino, *Science*, **180**, 417 (1973).

(5) B. R. Kowalski and C. F. Bender, *J. Am. Chem. Soc.*, **96**, 916 (1974).

(6) K. C. Chu, R. J. Feldman, M. B. Shapiro, G. F. Hazard, Jr., and R. I. Geran, *J. Med. Chem.*, **18**, 539 (1975).

(7) A. J. Stuper and P. C. Jurs, *J. Am. Chem. Soc.*, **97**, 182 (1975).

(8) F. Darvas, *J. Med. Chem.*, **17**, 799 (1974).

(9) K. C. Chu, D. E. Goldender, and A. B. Rosonblit, *Comp. Biomed. Res.*, **6**, 411 (1973).

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

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

(12) C. L. Perrin, *Science*, **183**, 551 (1974).

(13) J. T. Clerc, P. Naegeli, and J. Seibl, *Chimia*, **27**, 639 (1973).

(14) S. H. Unger, *Cancer Chem. Rep., Part 2*, **4**, 47 (1974).

(15) W. E. Brugger, A. J. Stuper, and P. C. Jurs, *J. Chem. Inf. Comp. Sci.*, **16**, 105 (1976).

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

(17) J. T. Tou and R. C. Gonzalez, "Pattern Recognition Principles," Addison-Wesley, Reading, Mass., 1974.

(18) R. O. Duda and P. E. Hart, "Pattern Classification and Scene Analysis," Wiley–Interscience, New York, N.Y., 1973.

(19) L. Kanal, *IEEE Trans. Inf. Theory*, **IT-20**, 697 (1974).

(20) P. A. Lachenbruch and R. M. Micke, *Technometrics*, **10**, 1 (1968).

(21) A. J. Stuper and P. C. Jurs, *J. Chem. Inf. Comp. Sci.*, **16**, 99 (1976).

(22) F. F. Blicke and R. H. Cox, "Medical Chemistry," vol. IV, Wiley–Interscience, New York, N.Y., 1959.

(23) M. Randic, *J. Am. Chem. Soc.*, **97**, 6609 (1975).

(24) L. B. Kier, W. J. Murray, and L. H. Hall, *J. Med. Chem.*, **18**, 1272 (1975).

(25) L. B. Kier and L. H. Hall, "Molecular Connectivity in Chemistry and Drug Research," Academic, New York, N.Y., 1976.

(26) L. B. Kier, L. H. Hall, W. J. Murray, and M. Randic, *J. Pharm. Sci.*, **64**, 1971 (1975).

(27) L. H. Hall, L. B. Kier, and W. J. Murray, *ibid.*, **64**, 1975 (1975).

(28) W. J. Murray, L. H. Hall, and L. B. Kier, *ibid.*, **64**, 1979 (1975).

(29) W. J. Murray, L. B. Kier, and L. H. Hall, *J. Med. Chem.*, **19**, 573 (1976).

(30) G. S. Zander, A. J. Stuper, and P. C. Jurs, *Anal. Chem.*, **47**, 1085 (1975).

(31) A. J. Stuper and P. C. Jurs, *J. Chem. Inf. Comp. Sci.*, **16**, 238 (1976).

(32) E. E. Swanson and W. E. Fry, *J. Am. Pharm. Assoc., Sci. Ed.*, **29**, 509 (1940).

(33) C. Hansch and S. M. Anderson, *J. Med. Chem.*, **10**, 745 (1967).