Classification of Psychotropic Drugs as Sedatives or Tranquilizers Using Pattern Recognition Techniques

Andrew J. Stuper and Peter C. Jurs*

Contribution from the Department of Chemistry, The Pennsylvania State University, University Park, Pennsylvania 16802. *Received June 12, 1974*

Abstract: Adaptive binary pattern classifiers have been applied to the problem of classifying psychotropic drugs as tranquilizers or sedatives based on their structures. A set of 219 tranquilizers and sedatives representing a wide variety of compound types has been coded using 69 descriptors of three types: (a) numeric and binary fragment descriptors, (b) binary substructure descriptors, and (c) topological descriptors. Only an ordinary two-dimensional structural diagram of the drug molecule is used as input to the system. When the 69 structural descriptors are used to code the set of drugs, the two classes are linearly separable and all molecules can be classified correctly. A feature selection procedure is employed to reduce the number of descriptors necessary for classification by approximately one-half with no loss of separability of the classes. When the classifiers are used to categorize unknowns of the same general molecular types, predictive abilities of approximately 90% are obtained.

The elucidation of structure-activity relationships of drugs has long been an area of concern. Various methods of correlating structural parameters with observed activity have been developed. One such method is the Hansch approach which attempts to relate a change in the level of biological activity with changes in the physical and chemical properties of a series of drug molecules.[1,2] This is accomplished by fitting standard or relative biological activities A_i to an equation of the form

$$\log A_i = -k_1 \log^2 P_i + k_2 \log P_i + k_3 \qquad (1)$$

where P_i is an octanol-water partition coefficient which models the partition coefficient of the drug between the nonpolar biophases and the aqueous phase of the biological system, and the k's are constants determined by regression analysis.

Recently, several techniques have been applied to studies of structure-activity correlations.[3-6] The studies relating mass spectra of various sedatives and tranquilizers to their activities[4] and of structure-activity relations of anticancer drugs[5] have drawn some criticism in the literature.[7-9] However, the techniques of pattern recognition should offer a viable alternative approach to studies of structure-activity correlations when used cautiously with large data sets and appropriate numbers and types of descriptors.

The present study is concerned with the implementation of adaptive binary pattern classifiers in order to distinguish between drug molecules which exhibit activity as sedatives and tranquilizers. The classifier bases its decision entirely on information available from a standard two-dimensional structural representation of the molecule. In the present implementation no geometrical descriptors or physical property descriptors other than molecular weight are employed, although their use is not precluded by the techniques being used. The results of classifications are then further used to deduce which of the given parameters are most effective in the determination of a given activity.

Binary Pattern Classifiers. Implementation of adaptive binary pattern classifiers involves representing a drug structure as a vector in n dimensional space, $\mathbf{X} \equiv (x_1, x_2, x_3, \ldots, x_n)$, where each x_j, $j = 1, 2, \ldots, (n - 1)$, represents some parameter related to the structure of each molecule. In order to ensure a common origin, all pattern vectors are given an extra component, x_n, the value of which is optimized during the initial stages of training.

These pattern vectors can be thought of as points in n dimensional space (hyperspace). If the descriptor list is sufficiently informative, compounds of similar activity should be found to cluster in the same relative region in space. If these clusters can be distinguished from one another by a linear decision surface or hyperplane, they are said to be linearly separable. In practice, discrimination between classes is made by calculating the dot product of a data vector with a vector (called the weight vector, $\mathbf{W}$) normal to the surface of the plane, $s = \mathbf{W} \cdot \mathbf{X}_i$. All vectors on one side of the plane will have a positive dot product; all those on the opposite side a negative dot product. The classification rule is then as follows: If $s > 0$, then classify $\mathbf{X}_i$ in category one; if $s \leq 0$, then classify $\mathbf{X}_i$ in category two. If a nonzero threshold, or deadzone, Z, is employed, the classification rule is as follows. If $s > Z$, then classify $\mathbf{X}_i$ in category one; if $s < -Z$, then classify $\mathbf{X}_i$ in category two; if $-Z < s < Z$, then do not classify $\mathbf{X}_i$.

Given a set of points in the n dimensional space, a useful decision surface is found by an operation called training. To train a decision surface, a set of data points of known class called the training set is used. They are presented sequentially to the binary pattern classifier being developed. Whenever any member of the training set is misclassified, the weight vector is modified as follows

$$\mathbf{W}' = \mathbf{W} + c\mathbf{X}_i \qquad (2)$$

in which

$$c = -2(\mathbf{W} \cdot \mathbf{X}_i)/(\mathbf{X}_i \cdot \mathbf{X}_i) \qquad (3)$$

and $\mathbf{X}_i$ is the pattern which was just misclassified, $\mathbf{W}$ is the old weight vector, and $\mathbf{W}'$ is the new improved weight vector. This corresponds to moving the decision surface so that the misclassified point is on the opposite side of the decision plane, thereby causing the point to be correctly classified. The procedure is continued until all points are correctly classified. A decision surface can also be trained in a manner such that the points not only lie on the correct side of the plane but must be a given distance on either side. This corresponds to giving the decision plane thickness and requiring that pattern points not only be on the correct side but also outside the volume occupied by the decision surface. Adaptive binary pattern classifiers have been described in detail elsewhere.[10-13] The adaptive binary pattern classifiers used in this work are coded in FORTRAN IV and executed on the Penn State Computation Center IBM 370/168 and the Department of Chemistry MODCOMP II/25.

The Data Set. The set of drugs used in this study consisted of 219 compounds selected from a standard reference.[14] The set contains the 140 tranquilizers and 79 sedatives list-

Table I. List of Compounds in Data Set

TRANQUILIZERS		SEDATIVES	
1 A124	2 ACEPROMAZINE	1 PROFENAMINE	2 PROMETHAZINE
3 ACEPROMETAZINE	4 ACETOPHENAZINE	3 CLOXYPENDYL	4 FENOHARMAN
5 BUTAPERAZINE	6 BUTYRYLPROMAZINE	5 CANNABIOEROL	6 D-58S1
7 CARPHENAZINE	8 CB 1519	7 LORAZEPAM	8 ALLOBARBITAL
9 CB 1658	10 CHLORIMIPIPHENINE	9 ALPHENAL	10 AMOBARBITAL
11 CHLORPROETHAZINE	12 CHLORPROMAZINE	11 APROBARBITAL	12 BARBITAL
13 CHLORPROMAZINE (1)	14 CIBA 17040	13 BUTALBITOL	14 BUTETHAL
15 CPD 12	16 CYAMEPROMAZINE	15 BUTALLYLONAL	16 CYCLOBARBITAL
17 CYCLOPHENAZINE	18 DICHLORPROMAZINE	17 CYCLOPAL	18 FEBARBAMATE
19 DIXYRAZINE	20 ETHYLISOBUTRAZINE	19 HEPTABARBITAL	20 HEXETHAL
21 FLUOROPHENOTHIAZINE	22 FLUPHENAZINE (2)	21 HEXOBARBITAL	22 MEPHOBARBITAL
23 FLUPHENAZINE	24 FLUPHENAZINE (3)	23 METHABARBITAL	24 METHITURAL
25 HEPLTYLPROMAZINE	26 HOMOPHENAZINE	25 METHOHEXITAL	26 NEALBARBITONE
27 KS-33	28 MD 5501	27 PENTOBARBITAL	28 PHENOBARBITAL
29 MEPAZINE	30 MESORIDAZINE	29 PROBARBITAL	30 SECOBARBITAL
31 METHIOMEPRAZINE	32 METHOPHENAZINE	31 TALBUTAL	32 THIAMYLAL
33 METHOTRIMEPRAZIN	34 METHOXYPROMAZINE	33 THIOPENTAL	34 VASALGIN
35 OXAFLUMAZINE	36 P 824	35 NSD 2023	36 ANILERIDINE
37 P 1030	38 PERAZINE	37 CATAPRESAN	38 CHI 21
39 PERIMETAZINE	40 PERPHENAZINE	39 CHI 34	40 CHI 38
41 PERPHENAZINE (1)	42 PHENAZIN	41 CHI 42	42 CLOMETHIAZOLE
43 PHENAZINE	44 PIPAMAZINE	43 DICHLORMETHYMONE	44 ES 708
45 PIPERACTAZINE	46 PIPERIDOCHLOR- (4)	45 ETHINAZONE	46 GLUTETHIMIDE
47 PROCHLORPERAZINE	48 PROMETAZINE	47 HOMOCHLORCYCLIZINE	48 K-2004
49 PROMAZINE	50 PROPIOMAZINE	49 LB 50160	50 MECLOQUALONE
51 PROPIOPROMAZINE	52 RIDAZINE	51 METHAQUALONE	52 METHYPRYLONE
53 RP 3300	54 R.P. 4627	53 OXYPENAYL	54 TETRIDIN
55 R.P. 6696	56 R.P. 9153	55 THALIDOMIDE	56 ETHOMOXANE
57 SA 124	58 SAF 5657	57 PARALDEHYDE	58 WB 4123
59 SKF 6333	60 T 412	59 TRICETAMIDE	60 CAITAMINE
61 SPICLOMAZINE	62 THIETHYLPERAZINE	61 RD 6020	62 CD 6030
63 THIOPROPAZATE	64 THIOPROPERIZINE	63 CHLORAL HYDRATE	64 DISPRANOL
65 THIORIDAZIDE	66 TPN 12	65 ETHINAMATE	66 MEBUTAMATE
67 TRIFLUOPERAZINE	68 TRIFLUOPERAZINE	67 MEPROBAMATE	68 NISUBAMATE
69 TRIFLUPROMAZINE	70 TRIFLUTRIMEPRAZINE	69 ETHCHLORVYNOL	70 METHYLPENTYNOL
71 TRIMEPRAZINE	72 VALEROYL-PERAZINE	71 PETRICHLORAL	72 ACETYLCARBOROMAL
73 WIN 13,645-5	74 CHLORPROHEPTADIENE	73 AEC	74 CARBROMAL
75 CLOMACRAN	76 CLOPENTHIXOL	75 BROMISOVAIUM	76 ECTYLUREA
77 CLOTHIAPINE	78 CLOTHIXAMIDE	77 IPC	78 VALNOCTAMIDE
79 CYANOTHEPIN	80 DESMETHYL-DOXEPINE	79 CHLORETHATE	
81 DOXEPIN	82 FLUPENTHIXOL		
83 G 22150	84 ID 22		
85 LOXAPINE	86 TRIFLUTHEPIN		
87 TRIMEPRAMINE	88 XANTHIOL		
89 BISHOMORESERPINE	90 RESERPEDINE		
91 METHYL-18-KETORE	92 RAUJEMIDINE		
93 RAUNESCINE	94 RENOXIDINE		
95 RESCINNAMINE	96 SU5171		
97 8842	98 SU10704		
99 RAUBASINE	100 BENANSERIN		
101 BENZINDOPYRINE	102 DIM		
103 3-IAAR	104 IN 399		
105 MILIPERTINE	106 OXYPERTINE		
107 PI 11	108 SOLPYERTINE		
109 DIMECHROM	110 BROMAZEPAM		
111 CHLORAZEPATE	112 CHLORDIAZEPOXIDE		
113 CLOAZEPAM	114 CLOXAZOLAZEPAM		
115 CT 5104	116 CYPRAZEPAM		
117 DIAZEPAM	118 ISOQUINAZEPON		
119 LORAZEPAM	120 MEDAZEPAM		
121 NITRAZEPAM	122 NITRAZEPATE		
123 OXAZEPAM	124 OXAZOLAM		
125 PRAZEPAM	126 RO5-2180		
127 RO-53027	128 SULAZEPAM		
129 TEMAZEPAM	130 TETRAZEPAM		
131 ACEPERONE	132 AHR 1900		
133 FR-33	134 DIPHENCHLOXAZINE		
135 MISAFLUR	136 PROTHIPENDYL		
137 TRIOXAZINE	138 CAPTODIAME		
139 PHENYLTOLOXAMINE	140 CINTRLAMIDE		

(1) SULFOXIDE (2) DECANOATE

(3) ENANTHATE (4) -PROMAZINE

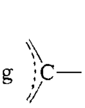
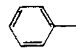
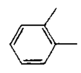
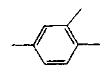
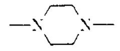
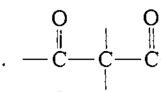
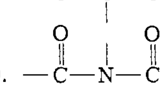
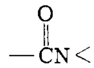
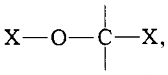
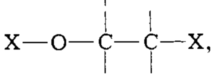
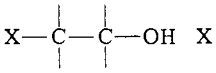
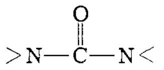
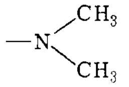
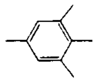
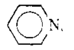
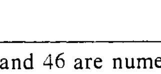
ed in Table I. A number of different parent ring types are represented, including phenothiazines, indoles, benzodiazepines, barbiturates, heterocyclic butyrophenones, nitrogen heterocycles, nonnitrogen heterocycles, and diphenylmethane derivatives. All the compound types represented and the number of each type are given in Table II.

Many medicinal chemists disagree on the precise classification of a large segment of the psychotropic agents and therefore the classifications of sedative and tranquilizer are not exacting. Many drugs show activities in both of these classes as well as in others, *e.g.*, hypotensive, muscle relaxant, etc. Generally tranquilizers are classified as being either major or minor tranquilizers, while many sedatives show hypnotic action. The method used in this paper to classify the compounds' major action as sedative or tranquilizer is based upon the information in the reference used (see ref 14). In this reference compounds were classified as either major tranquilizers (TMa), minor tranquilizers (TMi), tranquilizers (T), sedatives (Sed), hypnotics (Hyp), or sedative-hypnotics (Hyp-Sed). The classification rules

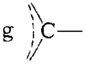
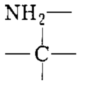
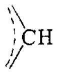
Table II. Compound Classes Represented in Data Set

Compound type	No. of tranquilizers	No. of sedatives
Phenothiazines	73	2
Phenothiazine analogs and isomers	15	1
Indoles		
Reserprine and derivatives	10	0
Haramine and derivatives	1	1
Others	9	0
Cannabis derivatives	0	1
Other heterocycles		
Chromone derivatives	1	0
Benzodiazepines	21	2
Barbiturates	0	27
Heterocyclic butyrophenones	3	1
Other N heterocycles	4	20
Benzodioxone derivatives	0	1
Non-N heterocycles	0	2
Aromatic compounds		
Diphenylmethane derivatives	2	0
Benzoic acid derivatives	0	1
Other	1	3
Aliphatic compounds		
Glycols	0	2
Carbamates	0	4
Carbinols ·	0	3
Amides and hydrazines	0	7
Others	0	1
	140	79

Table III. Descriptor List[a]

1. Molecular weight
2. Number of nonring carbon
3. Number of nonring oxygen
4. Number of nonring nitrogen
5. Number of nonring sulfur
6. Number of fluorine
7. Number of chlorine
8. Number of oxygen
9. Number of nitrogen
10. Number of sulfur
11. Number of carbon
12. Number of C=C
13. Number of C—C
14. Number of phenyl bonds
15. Total weighted bond length
16. —OC(=O)—
17. —C(=O)—
18. Aromatic ring 
19. >N—
20. 
21. 
22. 
23. 
24. Ring >S
25. Ring >N
26. Ring $>NCH_2$—
27. Ring >O
28. 
29. 
30. 
31.  , X ≠ H
32.  , X ≠ H
33.  X ≠ H
34. —OH
35. >C=C<
36. 
37. 
38. 
39. Ring $>N—CH_3$
40. —C—C—C—[b]
41. —C—C—C—
42. —C—C—[c]
43. —C—C—
44.  any substitution
45. Longest chain of nonaromatic carbon
46. Terminal $X—CH_3$ where X ≠ a carbon chain
47.  X ≠ H; Y = H or C

Topological Descriptors[d]

48. Aromatic ring 
50. NH_2—
52. 
54. >N—
56. >CH—
58. CH_3—
60. $—CH_2—$
62. —NH—
64. —O—
66. Aromatic ring 
68. —C(=O)—

[a] Descriptor numbers 1-15, 45, and 46 are numeric descriptors; the others through 47 are binary. [b] Value is 1 if CH_3CC is present, 2 if -CCC- is present, and 3 if both are present. [c] Value is 1 is CH_3C is present, 2 if -CC- is present, and 3 if both are present. [d] The topological descriptors were developed in pairs—bond and weighted—for the fragments shown.

were generally applied as follows: (1) if the compound was TMa, TMi, or T, then classify as tranquilizer; (2) if the compound was Hyp, Sed, or Hyp-Sed, then classify as sedative; (3) if the compound was a combination of activities such as (T Sed), (T Hyp), (TMi Sed), classify as a tranquilizer; (4) if in any of the multiple classifications given there was a preponderance of one class over the other, the compound was classified as belonging to the major class.

In no case was the classification changed in order that a more favorable result with respect to recognition rate be effected. The ability of the pattern recognition approach to deal with such a heterogeneous data set may be one of the strengths of the technique.

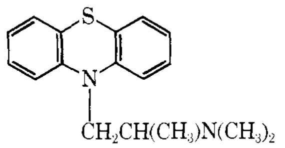
Descriptor Development. The success of the application of binary pattern classifiers to structure-activity correlations will depend upon the method used in describing the molecular structures. In this study three types of descriptors were employed: binary and numeric fragment descriptors, binary substructure descriptors, and topological descriptors. A list of the 69 descriptors is given in Table III. Each descriptor is contained in a minimum of 10% of the drug structures, and in no case does any one descriptor contain enough information to successfully classify the compounds.

Binary and numeric fragment descriptors have been used previously in studies of the interpretation of mass spectra.[15,16] A binary descriptor is simply an indication of the presence or absence of a certain structural parameter. Numeric descriptors indicate the number of times a descriptor appears in a molecule. Substructure descriptors, which are binary, tell whether or not a particular, explicitly described substructure is embedded in the molecular structure. Additional information concerning these types of descriptors can be found in ref 13.

While the nature of most of the descriptors is evident, some of them require further explanation. Descriptor 15, total weighted bond length, is calculated by summing four for each single bond, three for each phenyl bond, two for each double bond, and one for each triple bond in the molecule and dividing by two. (Note that these are in the same sequence as the lengths of these bond types.) Descriptor 18 is a binary descriptor which indicates whether an aromatic ring is branched. Descriptors 20, 21, 22, and 38 are binary descriptors which indicate whether or not these explicitly described substructures are present. Descriptor 48 differs from 18 in that 48 is a topological descriptor. Descriptor 41 differs from 40 and 43 from 42 in that 41 and 43 are binary substructure descriptors and the others are coded as shown and are numeric descriptors.

Along with information provided by fragment and substructure descriptors, information as to the molecular environment of various atoms within the molecule is of interest.

For instance, carbonyl carbons within aliphatic chains should be differentiated from carbonyl type carbons involved in such groups as acids, amides, or esters. A measure of this type of environment is provided by topological descriptors. Computerized generation of these parameters is facilitated by representing the structures of the molecules in the form of connection tables whereby each atom in the molecule is sequentially numbered and a matrix is then developed using the following criteria: (1) rows of the matrix indicate the numbered atoms; (2) columns indicate the possible atoms connected to this atom; (3) whenever an atom of row I is bonded to an atom in column J, the type of bond is entered into the (I,J)th position of the matrix (*e.g.*, single = 1, double = 2, triple = 3, phenyl = 4); (4) the main diagonal holds a numerical representation of the type of atom in the *N*th position of the structure (*e.g.*, C = 1, O = 2, N = 3, S = 4, Cl = 5, F = 6, Br = 7). Each topological descriptor's value is calculated by finding the atom center of interest and determining from the connection table the number of atoms connected to its first nearest neighbors and nearest neighbors of the first nearest neighbors. Thus, the topological descriptor holds the sum total of all the connectivities of the atom of interest. The contributions to the sum can be weighted by a constant dependent on bond type (*e.g.*, single = 1, double = 2, etc.). The former are called bond environment descriptors (BED) and the latter weighted environment descriptors (WED). As an example, the values for some of the BED descriptors for the sedative promethazine (a) are as follows: (48) 72; (54) 29; and (58) 22.



a

Thus, the raw data set consists of 219 drug structures each coded with 69 descriptors. Preprocessing of the raw data prior to training consisted of normalizing, autoscaling, and variance weighting. Normalizing consisted of multiplying each component of the data set by a factor such that the average value of all nonzero components was equal to 20. Then each descriptor was truncated to an integer value. This process yielded a normalized data set called the NDATA set. Secondly, the normalized data were subjected to autoscaling and variance weighting,[12] thus giving a normalized, autoscaled, variance weighted data set called the NAVDATA set. After autoscaling the normalized data were multiplied by a factor of 20 and truncated. The normalized and autoscaled data were variance weighted and multiplied by a scaling factor of 500 before truncation. A value of 20 was used for x_n because it provided fast training and high predictive ability.

The correlation coefficients $\rho(x_k,x_l)$ between each pair of descriptors was calculated for the $(69^2 - 69)/2 = 2346$ pairs. Of these, 24 (1%) had ρ values greater than 0.90, of which 11 were correlations between adjacent BED and WED topological descriptors. A total of 756 (32%) had ρ values in the range $-0.1 < \rho < 0.1$. The average ρ value for all 2346 pairs was 0.07.

Results

Training consisted of choosing 20 random sets of 209 compounds each from the total data set of 219 compounds using a random selection routine. This resulted in 20 sets of 209 knowns and ten unknowns each. These sets were then used to train two binary pattern classifiers: one to make the classification tranquilizers *vs.* nontranquilizers, the other to make the classification sedatives *vs.* nonsedatives. Having two BPC's allows the inclusion in the data set of molecules which are neither sedatives nor tranquilizers or that have been reported as showing both activities. After training of an individual BPC on 209 compounds, the ten unknowns were predicted. The overall predictive ability was taken to be the average percentage of all 200 unknowns which were correctly classified after all 20 trainings.

Since the BPC develops a decision surface in order to separate the classes, the predictive ability of the classifier is dependent to a large extent upon the number of members of the two classes in the training set which are nearest to the opposite class. A measure of the ability of the classifier to predict a compound having a rather small perturbation in properties from others in the data set is most accurately measured by sequentially removing one member from the data set, training, and predicting using the removed member as an unknown. This is known as the leave out one procedure. The rationalization for leaving ten out was to provide a measure closely related to that of the leave one out procedure, while reducing the amount of computer time necessary to obtain such a measurement. The average percentage predictive ability is then an approximate measurement of the success with which a BPC developed from all 219 compounds would have in correctly classifying an unknown not contained in the original data set. This ability would depend upon the extent to which the data set is representative of that compound. As in any learning process the smaller the perturbation from previous experience the more likely the chance for a successful classification.

The entire procedure was repeated a number of times using different threshold values. Table IV summarizes the results of these studies for the tranquilizer *vs.* nontranquilizer pattern classifier. It is seen that increasing the threshold causes an increase in predictive ability. The NDATA generally yielded better prediction than the NAVDATA; however, the NAVDATA normally required fewer feedbacks during training. In every case reported in Table IV training was performed with the threshold value, *Z*, shown and prediction was performed in two ways—with a zero threshold and with the same threshold as used during training. The two predictive ability results are reported separately.

Generally it would not be thought that all of the descriptors were of equal value in the training of the classifiers. In order to determine which of the 69 descriptors were of the greatest importance, weight-sign feature selection[17] was employed.

Each of the 20 randomly selected training sets was used for training using a threshold of 1.75, and the remaining ten drugs were predicted. The average of all 20 predictions were taken as a measure of the predictive ability prior to feature selection. Then the entire data set was used for training two weight vectors using one of the initializations listed in Table V and with the sequence of the compounds randomly scrambled. Any descriptor whose weight vector component sign did not agree for the two weight vectors

Table IV. Results of Training and Prediction

NDATA			NAVDATA		
Threshold	% prediction, $Z = 0$	% prediction. $Z > 0$	Threshold	% prediction. $Z = 0$	% prediction. $Z > 0$
0	89.5		0	86.0	
0.5	90.5	92.0	0.5	87.5	90.0
			1.0	89.0	90.0
			1.5	87.8	90.5
			2.0	88.3	90.0

Table V. List of Weight Vector Initializations

1. $w_j = 0, j = 1, 2, \ldots, (n - 1); w_n = 1$
2. $w_j = 0, j = 1, 2, \ldots, (n - 1); w_n = -1$
3. $w_j = 1, j = 1, 2, \ldots, (n - 1); w_n = 1$
4. $w_j = 1, j = 1, 2, \ldots, (n - 1); w_n = -1$
5. $w_l = 1; w_j = 0, j = 2, 3, \ldots, n$
6. $w_l = -1; w_j = 0, j = 1, 2, \ldots, n$
7. $w_j = -1, j = 1, 2, \ldots, (n - 1); w_n = 1$

Table VI. Results for Feature Selection[a]

	69 descriptors			After feature selection			
	% prediction	No. not predicted	Av feedbacks	% prediction	No. not predicted	Av feedbacks	Descriptors remaining
1.	85.08	5	386	88.94	2	340	40
2.	89.44	2	491	92.00	1	331	40
3.	86.00	1	266	90.00	0	274	38
4.	86.00	0	257	87.00	0	304	44
5.	85.89	1	300	91.44	1	187	35
6.	88.50	0	289	90.00	0	262	40
7.	86.00	0	280	87.50	0	223	34

[a] Training and prediction were done with $Z = 1.75$.

was discarded. The process was then repeated until no further features could be eliminated. Then the predictive ability was tested using the same procedure as used before feature selection. Results obtained for the seven independent implementations of the weight-sign feature selection procedure for the tranquilizer classification are shown in Table VI. The columns labeled "number not predicted" give the total number of unknowns for all 20 trainings which were not classified because their dot products fell inside the dead zone, *i.e.*, $-Z < s < Z$. It is seen that in every case the predictive ability was higher after the unnecessary descriptors had been discarded. Also note that the total number of compounds not predicted was never greater for the feature selected data than the nonfeature selected data and was, in fact, generally smaller for the former.

Since none of the original 69 descriptors alone has the ability to correctly classify all the data, some combinations of features must be used to attain separation.

The results of the feature selection procedure can be used to group the descriptors as to relative importance in performing the classification. Table VII lists the descriptors as a function of the fraction of the time they were retained during the seven trainings. Eleven descriptors (16%) were retained every time, while only four descriptors (6%) were eliminated in all the runs. The remaining 54 descriptors (78%) were intermediate in importance.

Generally, the exact sequence in which descriptors are eliminated and, therefore, the identity of the remaining features depend on how the weight-sign feature selection procedure is implemented. For each of the seven weight vector initializations, a different set of features was eliminated, and different average predictive abilities were obtained, as shown in Table VI. In all cases the data remained linearly separable. This behavior is possible because there exists more than one unique set of descriptors which afford linear separability. Further feature selection could be done by eliminating descriptors retained infrequently and redoing the entire weight-sign procedure on the new data set.

Table VII. Overall Feature Retention

Fraction of times retained	Descriptors retained
7/7	9, 15. 22. 26. 37, 39, 42, 45. 48. 49, 61
6/7	1, 5, 6, 10, 24, 28, 31. 35, 52, 57, 58, 60
5/7	7, 11, 13, 20. 32, 47. 54, 59
4/7	2, 12, 14, 18, 21, 25, 34
3/7	23, 29, 31, 33, 44. 46, 51. 53, 56
2/7	3, 8, 16, 27, 30. 38, 50, 55, 65
1/7	4, 19, 34, 40. 41, 43, 63, 66, 67
0/7	17, 62, 68, 69

Discussion

The data set used here represents a fairly diverse set of compounds, not all of which have their mode of action in the same area of the biosystem. It is rather unlikely that the exact mechanism of these compounds' actions is all intrinsically related, although all are CNS agents. However, many other factors come into play in the relation of structure to function. Solubility and those parameters that affect solubility, geometric consideration, electronic parameters, etc., also affect the observed activity. Given the sum total of all this, similarities may exist which in general distinguish one class of compounds from the other.

The development of relevant descriptors and an efficient means of feature selection is then the crux of this problem. As shown in Table VI any single feature selection was able to reduce the features required for separation by at least 40% while maintaining or increasing the predictive ability as well as further reducing the number of compounds which were not classified.

The possibility of those compounds which were not put into either class, being those compounds whose activity is in question, was not investigated. The present data set is not sufficiently feature selected for such a study to be meaningful since those descriptors that are not necessary for separation, while not influencing the ultimate ability to separate the classes, do influence the outcome of the prediction procedure. This can be evidenced from the fact that the number of compounds not predicted decreased as the number of unnecessary descriptors decreased. Future studies will investigate this question. The ultimate aim of this set of experiments was to demonstrate the potential usefulness of pattern recognition in this and other areas of activity correlations.

Preliminary studies showed that descriptors such as the number of nitrogens, total bond length, and various fragment and topological descriptors ranked high in their importance for discrimination as applied to this data set. This does not necessarily indicate that the properties of sedation and tranquilization are related by a cause and effect relationship to these parameters, but only that these parameters have a high degree of usefulness in the mathematical discrimination of these properties as they pertain to this data set. Also, generalizations based upon the preliminary feature selection attempted would be unwarranted as it is incomplete. The present data set does not contain any compounds which could be considered as inactive for both classes, although families were included that showed activities in both classes, *e.g.*, the benzodiazepine derivatives. The problem of what types of inactive compounds to include so as to create a meaningful data set is not trivial. At present only those compounds which are analogs of the present compounds or are CNS agents are under consideration for inclusion in the data set. Besides expansion, further improvement of the data set could be accomplished by addition of other types of topological or environment descriptors, addition of parameters obtained from computerized modeling of the structure, and more sophisticated methods of feature selection. Also, studies dealing with various homologous series within the data set are being investigated.

In conclusion, it is felt that these results indicate that further studies should be made in this area and that pattern recognition techniques can be of use in the classification of drugs as to their pharmacological activity.

Acknowledgment. The financial support of the National Science Foundation is gratefully acknowledged.

References and Notes

(1) W. J. Dunn, *Annu. Rep. Med. Chem.*, **8,** 313 (1973).
(2) *Advan. Chem. Ser.*, **No. 114,** 1 (1972).
(3) C. Hansch, S. H. Unger, and A. B. Forsythe, *J. Med. Chem.*, **16,** 1217 (1973).
(4) K-L. H. 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. Amer. Chem. Soc.*, **96,** 916 (1974).
(6) K. C. Chu, R. J. Feldman, M. Shapiro, G. F. Hagard, Jr., C. L. Chang, and R. Geran, Abstracts, 167th National Meeting of the American Chemical Society, Los Angeles, Calif., 1974.
(7) C. L. Perrin, *Science*, **183,** 551 (1974).
(8) J. T. Clerc, P. Naegeli, and J. Seibl, *Chimia*, **27,** 639 (1973).
(9) S. N. Unger, *Cancer Chem. Rep.*, in press.
(10) N. J. Nilsson, "Learning Machines," McGraw-Hill, New York, N.Y., 1965.
(11) H. C. Andrews, "Mathematical Techniques in Pattern Recognition," Wiley-Interscience, New York, N.Y., 1972.
(12) B. R. Kowalski and C. F. Bender, *J. Amer. Chem. Soc.*, **94,** 5632 (1972).
(13) P. C. Jurs in "Computer Representation and Manipulation of Chemical Information," W. T. Wipke, *et al.*, Ed., Wiley-Interscience, New York, N.Y., 1974.
(14) E. Usdin and D. H. Efron, "Psychotropic Drugs and Related Compounds," 2nd ed, DHEW Pub. No. (HSM) 72-9074, 1972.
(15) J. Schechter and P. C. Jurs, *Appl. Spectrosc.*, **27,** 30 (1973).
(16) J. Schechter and P. C. Jurs, *Appl. Spectrosc.*, **27,** 225 (1973).
(17) P. C. Jurs, *Anal. Chem.*, **42,** 1633 (1970).

Matrix Reactions of Alkali Metal Atoms with Tetrahalomethanes. Evidence for a Novel Carbenoid

Douglas A. Hatzenbühler, Lester Andrews,* and Francis A. Carey

Contribution from the Department of Chemistry, University of Virginia, Charlottesville, Virginia 22901. *Received May 16, 1974*

Abstract: Reaction of alkali metal atoms with carbon tetrachloride at high dilution in argon deposited at 15°K produced, in addition to trichloromethyl radical, dichlorocarbene, and the carbenoid Cl_3CM, a new carbenoid species identified as a dichlorocarbene–metal halide complex. Infrared examination of the matrix-isolated complex indicated a nonrandom orientation of the two components. Analysis of the effect of the metal atom, the carbon isotope, and the halide ion dependence of the spectrum allowed the determination that the new carbenoid involves interaction of the metal ion and the lone pair of the carbene carbon, *i.e.*, $X^- \cdots M^+ \cdots CCl_2$.

The application of α-elimination routes to the generation of carbenes[1] in solution remains a favored method in spite of ambiguities encountered in interpreting the results of trapping experiments. These ambiguities arise from similarities in reactivity between free divalent carbon species (*carbenes*) and, for example, the α-haloorganometallics generated by metalation of an alkyl halide.[2] Species of this latter type have been included in a class of reactive intermediates termed *carbenoids*[3] which include not only the structurally well-defined α-haloorganometallic reagents just mentioned but also less well-defined carbene–alkali halide complexes of various (usually unstated) degrees of aggregation as well as complexes between α-haloorganometallics and alkali halides. Considerable progress has been made in understanding a carbene–carbenoid relationship in the case of the dichlorocarbene–trichloromethyllithium equilibrium (eq 1).

$$Cl_3CLi \rightleftharpoons CCl_2 + LiCl \tag{1}$$

The position of equilibrium lies heavily to the side of trichloromethyllithium at −100°.[4] Reaction with olefins, however, to afford 1,1-dichlorocyclopropanes appears to involve only CCl_2 and not Cl_3CLi since the pattern of reactivity *vs.* olefin structure is identical with that observed using free CCl_2 generated in the gas phase.[5] While it may be convenient to envision a continuum of complexes between free dichlorocarbene plus lithium chloride at one extreme and trichloromethyllithium at the other, it is much more instructive to probe for those arrangements which correspond to energy minima.

Infrared spectroscopy in association with matrix isolation techniques has proved to be valuable in studies of structure and bonding of the dihalocarbenes and trihalomethyllithiums generated by the reaction of tetrahalomethanes with alkali metal atoms in solid argon (eq 2–4).[6]

$$Li + CCl_4 \longrightarrow LiCl + CCl_3 \tag{2}$$

$$2Li + CCl_4 \longrightarrow 2LiCl + CCl_2 \tag{3}$$

$$2Li + CCl_4 \longrightarrow LiCl + LiCCl_3 \tag{4}$$

Matrix-isolated CCl_2 generated according to eq 3 is readily identified by comparing vibrational frequencies with CCl_2 generated by photochemical routes[7] and by thermolysis of phenyltrichloromethylmercury.[8] Association of lithium chloride molecules with CCl_2 or with $LiCCl_3$ in the same matrix site is also possible but has so far not been systematically studied. It is of chemical interest to examine the infrared spectra of dichlorocarbene–alkali halide "complexes" to determine: (a) the extent of interaction of the two species when trapped together; and (b) whether the perturbation of CCl_2 vibrations by alkali halides corresponds to a unique and determinable orientation of the two species.

Experimental Section

Matrix reactions of CCl_4, $^{13}CCl_4$, CCl_3Br, and CCl_2Br_2 with Li, Na, K, or Cs atoms were performed using experimental techniques described previously.[9] Matrix samples (Ar/CCl_4 = 100:1 to 400:1) were codeposited with alkali metal vapor beams on a CsI substrate maintained at 15°K. Infrared spectra were recorded during and after sample deposition on a Beckman IR-12 spectrophotometer. Wave-number accuracy is ± 1 cm^{-1}.

Results and Discussion

Pertinent portions of the infrared spectra obtained on reaction of CCl_4 with ^6Li, Na, K, and Cs atoms in solid

Andrews, et al. / Reactions of Alkali Metal Atoms with Tetrahalomethanes