Journal of Medicinal Chemistry

Volume 18, Number 6 June 1975

Pattern Recognition and Structure–Activity Relationship Studies. Computer-Assisted Prediction of Antitumor Activity in Structurally Diverse Drugs in an Experimental Mouse Brain Tumor System

K. C. Chu,*[,1a] R. J. Feldmann,[1a] M. B. Shapiro,[1a] G. F. Hazard, Jr.,[1b] and R. I. Geran[1b]

Division of Computer Research and Technology and National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20014. *Received November 25, 1974*

This paper reports the application of pattern recognition and substructural analysis to the problem of predicting the antineoplastic activity of 24 test compounds in an experimental mouse brain tumor system based on 138 structurally diverse compounds tested in this tumor system. The molecules were represented by three types of substructural fragments, the augmented atom, the heteropath, and the ring fragments. Of the two pattern recognition methods used to predict the activity of the test compounds the nearest neighbor method predicted 83% correctly while the learning machine method predicted 92% correctly. The test structures and the important substructural fragments used in this study are given and the implications of these results are discussed.

Predicting the activity of a compound has been a primary goal of structure–activity relationship (SAR) studies for many years.[2] When the SAR problem involves correlating congeners such as finding the most active compound in a series and quantitative biological data are available then quantitative structure–activity relationship (QSAR) methods, such as the Hansch analysis and the Free–Wilson approach, may be useful.[3] On the other hand, if the SAR problem deals with structurally diverse compounds then substructural analysis may be appropriate.[4] And if quantitative biological data are unavailable then other techniques may be appropriate. Recently, pattern recognition has been applied to predicting pharmacological activity.[4a,5-7] Although some of these studies[5,7] have drawn criticism[8-10] with regard to the data and its representation, the pattern recognition techniques may offer a useful complement to QSAR methods. In an effort to help delineate the role of pattern recognition in SAR studies, this study attempts to predict the antineoplastic activity of test compounds in an experimental mouse brain tumor using pattern recognition based on structurally diverse compounds tested in this tumor system.[11]

In general, this problem involved (1) defining and assigning biological activity to a set of drugs (called the training set) which was used to establish the criteria for activity, (2) creating mathematical representations of the molecules, (3) selecting and applying the pattern recognition methods, (4) predicting the activity of a set of test drugs (called the test set), and (5) analyzing the results.

Criteria for Biological Activity. The structures of the drugs and their biological activity used in this paper were taken from a study on the effects of drugs on a murine ependymoblastoma in mice.[11] In that study, a transplantable solid mouse brain tumor was implanted into the brain of mice. Then groups of six mice were treated with drugs with a control group of 25 nontreated mice. The parameter for measuring the activity of the drugs was the median survival time of the mice. That is, a comparison of the test median survival time to the control median survival time gave the degree of increased life span (T/C). A drug was considered active if in two separate experiments at the same dosage tests gave a 25% increase in the life span of the treated mice, that is $T/C > 125\%$.

Of the 177 compounds in the ependymoblastoma study, 27 were incompletely tested and 12 had undefined structures, leaving 138 compounds. The training set (the set of drugs whose activity was known) consisted of 32 active drugs and 106 nonactive compounds while the test set (the set of drugs whose activity is to be predicted) consisted of all drugs (17 inactives and 7 actives) whose activity was completely determined after the reported ependymoblastoma study. In general, the compounds used in this study showed some type of antineoplastic activity toward other experimental tumor systems, generally L1210, the leukemia tumor system.

Mathematical Representation of Molecules. In order to apply pattern recognition to SAR problems, a molecule is represented as a point in n-dimensional space (referred to as a feature space). Each dimension or component represents a property such as the partition coefficient or a substructure of the molecules under study, where the type of properties selected to represent a molecule depends on the particular SAR problem. In our study, the antineoplastic agents span a wide range of structures, such as metal complexes, antimetabolites, alkylating agents, and alkaloids. This diversity of structure could be represented in several ways. One way is to use a global physicochemical parameter such as the partition coefficient, log P;[12] however, these parameters were not available for all the compounds in the study. Furthermore, this type of physical data may be difficult to obtain for a large diverse set of reactive molecules. A logical alternative was to use the structure of the molecule.

Table I. Augmented Atom Fragments for Three Simple Molecules

	Fragments	$CH_3-C(=O)-H$ D	$CH_3-C(=O)-CH_3$ K	$CH_3-C(=O)-OH$ A
1	O=C	1	1	1
2	H_3C-C	1	2	1
3	$C-C(=O)-H$	1	0	0
4	$C-C(=O)-C$	0	1	0
5	$C-O-H$	0	0	1
6	$C-C(=O)-O$	0	0	1

Thus, in this study a molecule is represented as a point in n-dimensional space where each dimension represents a substructural unit present in the molecules under study and the value for each dimension for any molecule is the number of occurrences of that unit in the molecule. For instance, if the first dimension represented amines, then the value of this dimension for a molecule is the number of amine groups in the molecule.

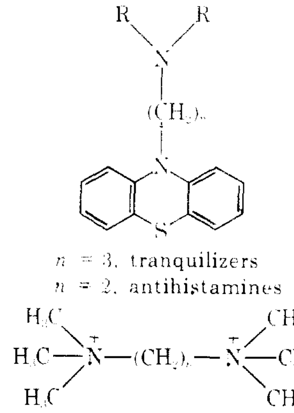
There are two important factors in determining the types of substructural features chosen to represent a molecule. First, the substructural fragments should be chemically meaningful, such as functional groups and rings. Second, the fragments should allow the use of computer structure files, such as the Chemical Abstract Service (CAS) files of over 2.7 million registered compounds. With these factors in mind, the three types of substructural units used in this study were (1) the augmented atom, (2) "heteropath", and (3) ring fragments.

The augmented atom or the atom-centered fragment is the basic fragment used in substructural retrieval systems such as the CAS system. This fragment, which is created for every nonhydrogen atom in a molecule, is defined as an atom and its adjacent atoms and bonds.[13] Simple examples of the decomposition of molecules into their augmented atom fragments and the creation of a vector are given in Table I. In this table, the rows represent the number of occurrences of a feature in the molecules and the columns represent a 6-dimensional vector representation of the drugs for the augmented atom fragments.

The augmented atom fragment has several important structural characteristics. First, it is an exact representation of many functional groups, such as fragment 3 which defines an aldehyde and fragment 4 which defines a ketone, and an excellent approximation to others, such as fragment 6 which represents the carboxy group. In other cases, this fragment can represent subfunctional groups such as the carbonyl moiety, fragment 1, which is common to the aldehyde, ketone, and acid. In addition, the CAS bond types not only specify whether the bond is aromatic, tautomeric, single, double, or triple but also specify whether the bond is in a ring nucleus.

The second substructural unit, the "heteropath" fragment, was created because the pharmacological activity of a set of derivatives can change with the number of carbon units between two heteroatoms. For example, in Chart I the activity of phenothiazine derivatives is antihistaminic if $n = 2$ and tranquilizing if $n = 3$.[14] In another case, the activity of bis-quaternary ammonium compounds varies from ganglionic if $n = 10$ to ganglionic-blocking if $n = 6$ (see Chart I).[14] In order to express these differences explicitly the heteropath fragment, which is defined as the path from one heteroatom to another, was generated. An immediate advantage of this fragment was that some large pharmacodynamic groups were explicitly represented as heteropath fragments.

Chart I. Compounds with Pharmacological Activities Dependent on the Number of Methylene Groups between Two Heteroatoms

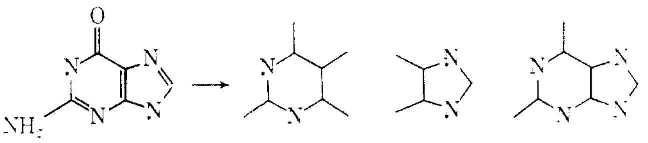


n = 3, tranquilizers
n = 2, antihistamines

n = 10, ganglionic agents
n = 6, ganglionic-blocking agents

Neither the heteropath fragment nor the augmented atom fragment contain information regarding the rings in a molecule, e.g., ring size, type of heteroatoms in the rings, and the relationship of several rings to one another. Thus, the third fragment involved the ring nuclei. All combinations of fused ring systems with their heteroatoms were identified. In the case of six-membered rings, the substitution pattern was also included. One purpose of this fragment was to discover important imbedded substructural rings. For example, the purine structure in Scheme I would be represented by three ring fragments including the pyrimidine nucleus.

Scheme I. Ring Fragments for Guanine



All the substructural units described above were generated automatically from connection tables derived from the Chemical Abstract Service Registry II system connection tables by a substructural retrieval system developed by Feldmann.[15]

Pattern Recognition Techniques. After the substructural fragments were created for the molecules in the training set these fragments were used to create points in n-dimensional space. Then the drugs in the test set were represented as points in the n-dimensional space created by the training set fragments. Hence, pattern recognition techniques were applied to the training set points allowing the classification of the test set.

A standard pattern recognition technique is the "nearest neighbor" method.[16] It involves the computation of Euclidean distances between points (drugs) in the n-dimensional space. The formula for the distance (squared) between two such drugs, **T**(J) a test set drug and **P**(K) a training set drug, is

$$D(\mathrm{K,J}) = \sum_{i=1}^{n} (T_i(\mathrm{J}) - P_i(\mathrm{K}))^2$$

where $T_i(\mathrm{J})$ and $P_i(\mathrm{K})$ represent the ith component of vectors $\mathbf{T}(\mathrm{J})$ and $\mathbf{P}(\mathrm{K})$. This distance is calculated for a test sample and each compound in the training set. The method assumes that the closer two points are in space the more they are alike pharmacologically. Thus the activity of the test sample is determined by the activity of the compound in the training set with the smallest Euclidean distance measurement to the test sample. In Figure 1, a two-dimensional representation of drugs is given where test samples X and Y are closest to nonactive compounds and were classified as such.

Another powerful pattern recognition technique, the "learning machine",[16,17] is an error-correcting procedure which attempts to create a linear mathematical function called a linear decision surface (a plane in two dimensions or hyperplane in higher dimensional space) which can separate the active drugs of the training set from the nonactives. In order to allow the decision surface to pass through the origin an additional feature usually given the value 1 was added to all the samples in the training set. A drug is thus represented as a vector (point), $\mathbf{P}'(\mathrm{K})$, in $n + 1$ dimensional space. To determine the linear decision surface, the following equation was evaluated for each point

$$\mathbf{W}\cdot\mathbf{P}'(\mathrm{K}) = \sum_{i=1}^{n+1} W_iP_i'(\mathrm{K}) = S$$

where $\mathbf{W}$ (the weight vector) is a vector normal to the linear decision surface and defines this surface, and $\mathbf{P}'(\mathrm{K})$ is the vector describing the kth drug in the training set and S is their dot product. After the initial values of the weight vector, $\mathbf{W}$, were set, the dot product of the weight vector with a drug vector $\mathbf{P}'(\mathrm{K})$ was evaluated. If the sign of S was positive, then $\mathbf{P}'(\mathrm{K})$ was classified as active; otherwise $\mathbf{P}'(\mathrm{K})$ was nonactive. If the classification according to the sign of S agreed with the category of $\mathbf{P}'(\mathrm{K})$, then $\mathbf{P}'(\mathrm{K})$ was classified correctly. If $\mathbf{P}'(\mathrm{K})$ was classified incorrectly, then the plane defined by $\mathbf{W}$ must be moved to classify this point correctly. A standard correction factor is

$$\mathbf{W}' = \mathbf{W} + (-2S/\mathbf{P}'(\mathrm{K})\cdot\mathbf{P}'(\mathrm{K}))\mathbf{P}'(\mathrm{K})$$

This correction factor moves the plane an equal distance on the other side of the point, thus classifying it correctly. This process was continued until all the patterns were classified correctly or until a present number of attempts had been performed. This gave a final hyperplane [$\mathbf{W}$ (final)] which was used to predict the activity of a test sample, $\mathbf{T}'(\mathrm{J})$:

$$\mathbf{W}(\mathrm{final})\cdot\mathbf{T}'(\mathrm{J}) = S'$$

The sign of S' determined the classification of $\mathbf{T}'(\mathrm{J})$, plus for active and minus for nonactive. An example of a linear decision surface for a two-dimensional problem is given in Figure 1. In this case sample X was classified as active and Y as nonactive.

Pattern Spaces. Coding the compounds into the three types of substructural fragments resulted in 421 distinct substructural features, that is, all the substructures with two or more occurrences (represented as F421): 161 augmented atom fragments, 129 heteropath fragments, and 131 ring fragments. Since a sample to feature ratio of three or greater is desirable for a learning machine decision surface,[18] the initial feature space was reduced. In general, the method employed to reduce the features allowed the learning machine to determine which features would be retained. The initial features were reduced by selecting those features which did not change the sign of their weight vector components when +1 then −1 were used as initial weight values for the learning machine. This procedure has been called the weight–sign change feature selection technique.[19] The application of this procedure reduced 421 features to 127 (F127) then from 127 to 70 (F70) at which point no sign changes occurred. Then the removal of insignificant heteropath fragments with paths of seven carbons or more and the application of the sign changing procedure reduced the features from 70 to 51 (F51).

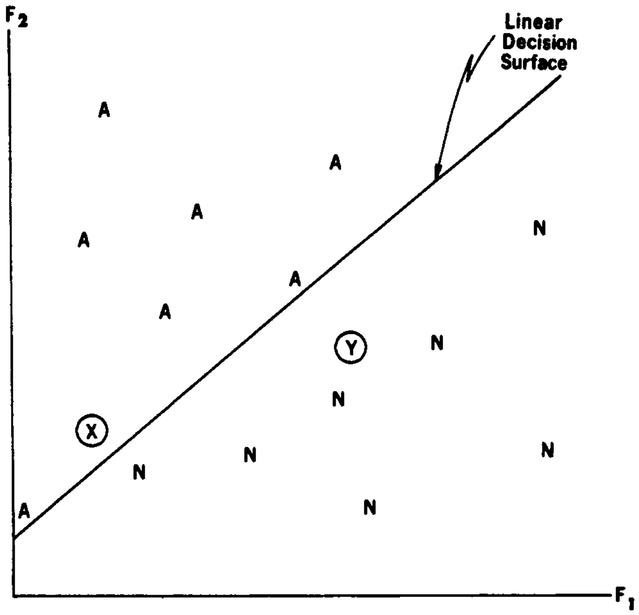


Figure 1. Two-dimensional plot of points with two unknowns X and Y, where F_1 and F_2 are the number of occurrences of features F_1 and F_2.

Because we were dealing with antineoplastic drugs it was highly desirable to assign high priorities to all drugs which possess potentially active structural units. And since the learning machine indicates which features contribute to activity and inactivity by the sign of their weight component, all nonactive features which were not required for 100% recognition were removed. This process reduced 51 features to 37 (F37). Since the weight vector (F37 hyperplane) represented a sample to feature ratio of 3.7, it was used to predict the activity of the test samples. These features and their individual weight components are given in Tables II–IV.

In the case of the nearest neighbor method there is no theoretical sample to feature ratio requirement. As a consequence, the five pattern spaces described above, F421, F127, F70, F51, and F37, were used to assign priorities to the test set. In general, the higher dimensional spaces, such as F421 and F127, reflect the overall nature of the molecule since most structural features are included. Thus, these feature spaces are used to find congeners. However, slight structural changes which may effect activity can be obscured when a large number of features are being considered. Thus, the lower dimensional spaces F70, F51, and F37 accentuate specific substructural characteristics while disregarding large portions of the molecule. However, these specific features may be overemphasized.

Results

A common way of predicting if an unknown compound will be active is to compare its structure to compounds known to be active.[20] If the compounds are similar, that is, they are congeners or the unknown has important sub-

Table II. 18 Augmented Atom Fragments and Their Learning Machine Weights for F37

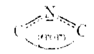
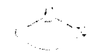
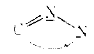
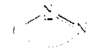
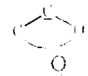
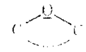
Feature no.	AA fragment	LM weight	Feature no.	AA fragment	LM weight
1	$C{=}\underline{C}{-}C$	0.57	10	$H_2\underline{N}{-}C$	−0.05
2	$H_3\underline{C}{-}C$	−0.75	11		−1.11
3		0.14	12		−1.52
4	$H_3\underline{C}{-}N$	0.53	13		1.39
5	$N{-}\overset{\overset{O}{\parallel}}{\underline{C}}{-}N$	1.76	14	$C{-}\overset{\overset{N}{\vert}}{\underline{N}}{-}C$	0.12
6		1.26	15	$N{-}\underline{N}{=}O$	1.72
7	$C{-}\overset{\overset{O}{\parallel}}{\underline{C}}{-}\dot{C}$	−0.01	16		−0.59
8	$\underline{Cl}{-}C$	0.29	17	$\underline{O}{=}C$	−0.16
9	$\underline{Cl}{-}Pt$	0.71	18	$\underline{O}{\cdots}C$	−0.74

Table III. Nine Heteropath Fragments and Their Learning Machine Weights for F37

	HP fragment			
	Heteroatoms			
Feature no.	1st	2nd	No. of carbons	LM weight
19	N	N	0	−0.38
20	N	N	2	0.45
21	N	N	6	0.35
22	O	N	6	0.39
23	O	O	2	0.22
24	O	O	4	0.54
25	O	O	6	−0.95
26	S	N	3	0.61
27	S	O	0	−1.56

structural units, then it can be given a high priority for biological testing. In our study, the pattern recognition method established separate criteria for assigning a high priority to the test samples for determining the similarity of compounds. However, the underlining premise of these criteria was that since 100% correct prediction may not be possible, in dealing with antitumor drugs, it was more desirable to predict that a compound be a false-positive than a false-negative. Thus, if a test sample contained new substructural fragments, that is, fragments not found in the training set fragments such as a new ring nuclei, then the compound could be assigned to be tested regardless of the pattern recognition predictions since these fragments may indicate that there is insufficient information in the training set to accurately predict activity.

For the nearest neighbor method, a test sample was assigned a high priority if any of the five pattern spaces predicted the compound to be active. While in the learning machine method, a test sample is assigned a high priority if the F37 hyperplane predicts the sample is active. As stated earlier, the substructural fragments and their weights for F37 are reported in Tables II–IV. A positive weight indicates that the feature is important for activity, while a negative value indicates a feature is important for inactivity. The larger the magnitude of the weights the greater their importance. The prediction results are reported in Table V, and the structures of some test samples and their pattern recognition results are reported in Tables VI–IX.

To aid in the evaluation of the prediction results, the probability of obtaining as good or better results was determined under a simple model. Suppose the 7 active and 17 inactive test compounds used in this study are presented to an investigator who knows only the training set probabilities of 0.232 (32/138) for the actives and 0.768 (106/138) for the inactives. If the investigator uses these probabilities to independently classify each unknown at random, then the probability of correctly classifying n or more of the actives and m or more of the inactives is

$$\sum_{X=n}^{7}\sum_{Y=m}^{17}\binom{7}{X}(0.232)^X(0.768)^{7-X}\binom{17}{Y}(0.768)^Y(0.232)^{17-Y}$$

where n is the number of correctly predicted actives and m is the number of correctly predicted inactives. For the learning machine results, $n = 7$ and $m = 15$ because this method correctly predicted 7 actives and 15 of 17 inactives. The probability of the investigator getting as good or better results is less than 0.00001. For the nearest neighbor method with $n = 6$ and $m = 14$, the probability is less than 0.0005. Clearly, the prediction results could not be duplicated by a random classification using only the training set probabilities.

Discussion

The prediction rates suggest that pattern recognition with substructural analysis can predict the antineoplastic activity of the test compounds. This result is due to several factors. First, the training set is representative of the different types of active compounds. Conversely, the test set does not introduce a new type of active compound. Second, similar structures had, in general, similar activities. For these data, the relative number of false-negatives to false-positives also suggests that the active substructural features in F37 are necessary but not sufficient conditions for activity. Conversely, the absence of these units seems to indicate inactivity.

As stated earlier, a basic premise of the pattern recognition methods is that similar structures have similar activity. When substructurally similar compounds, such as the aromatic nitrogen mustard, 3088, or the terephthalanilide, 60339, and their respective nearest neighbors give different activities, then other parameters, inadequately described by substructural features, must be important, such as partition coefficients and electronic effects.

A key pattern recognition limitation is that the methods are only as good as the scope of the data in the training set since the pattern recognition methods cannot predict accu-

Table IV. Ten Ring Fragments and Their Learning Machine Weights for F37

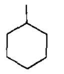
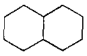
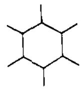
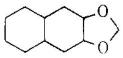
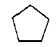
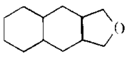
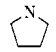
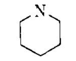
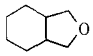
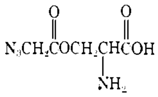
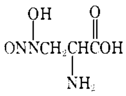
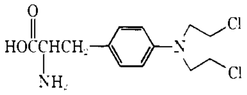
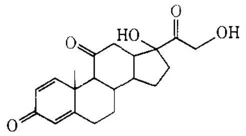
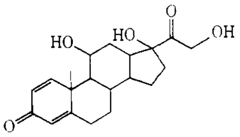
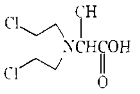
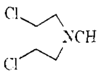
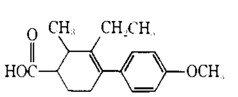
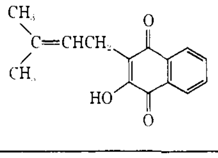
Feature no.	RN fragments	LM weight	Feature no.	RN fragments	LM weight
28		0.58	34		0.22
29		-0.26	35		0.69
30		-2.15	36		0.53
31		-2.18	37		1.90
32		0.69	38	Constant	-1.15
33		1.22			

Table V. Results of the Pattern Recognition Prediction of Ependymoblastoma Activity

	No. of samples sucessfully predicted as nonactive	No. of false-negatives	No. of samples successfully predicted as active	No. of false-postives	Percentage of test samples predicted correctly
NN	14	1	6	3	83
LM	15	0	7	2	92
Ideal	17	0	7	0	100

Table VI. Test Samples Correctly Predicted as Nonactive

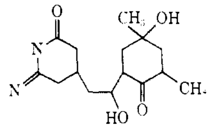
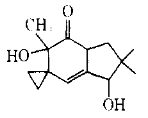
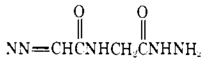
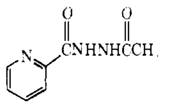
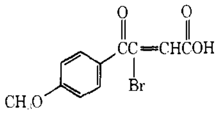
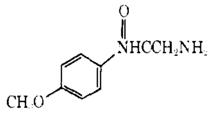
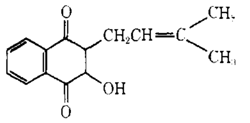
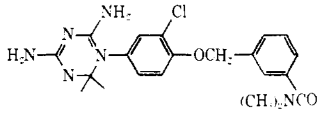
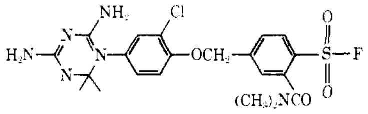
NSC no.	Structure	Nearest neighbor[a]	Dist.[b]	Prediction NN[c]	Prediction LM[d]	EM activity[e]
742			34.0	0-5	NA	Nonactive
8806	 L	DL	0.0	0-5	NA	Nonactive
10023			4.0	0-5	NA	Nonactive
17663			13.0	0-5	NA	Nonactive
19962			99.0	0-5	NA	Nonactive

[a] Nearest neighbor to test sample for pattern space F421. [b] Euclidean distance measurement. [c] The number of times the nearest neighbor method predicted a compound active out of the five pattern spaces, e.g., 0-5 means none out of five. [d] Learning machine prediction for activity: NA for nonactive and ACT for active. [e] Epe..dymoblastoma activity.

rately beyond the scope of the training set data. For instance, the misclassification of cyclic ether, 77037, containing several isoquinoline units could be attributed to the fact that isoquinoline units were present only in active compounds in the training set. Therefore, given the information available at the time of the evaluation, the computer predictions were logical. Of course, the training set can be updated to include these new results and new decision surfaces can be identified.

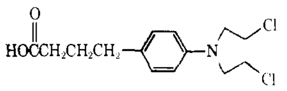
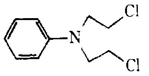
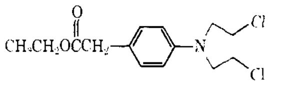
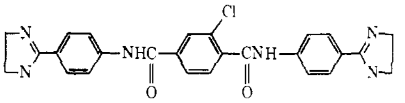
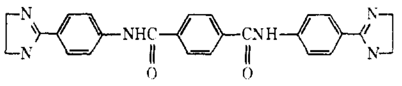
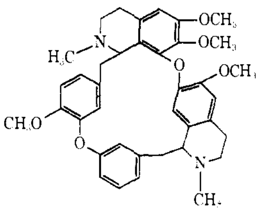
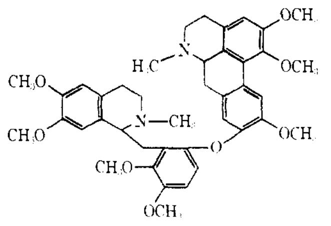
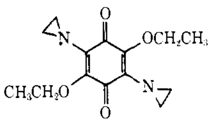
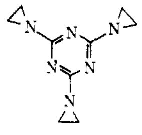
The two pattern recognition methods used in this study were designed to mimic two common ways a medicinal chemist recognizes the potential of a compound.[20] One way

Table VII. Test Samples Correctly Predicted as Nonactive

NSC no.	Structure	Nearest neighbor[a]	Dist.[b]	Prediction NN[c]	Prediction LM[d]	EM activity[e]
39147			110.0	0-5	NA	Nonactive
58404			20.0	0-5	NA	Nonactive
104801			58.0	0-5	NA	Nonactive
			58.0	0-5	NA	Nonactive
139105			88.0	0-5	NA	Nonactive

[a–e]See footnotes, Table VI.

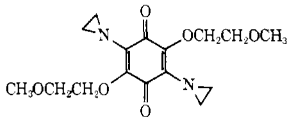
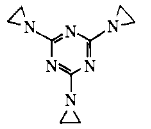
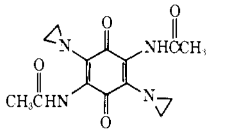
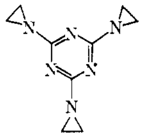
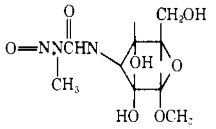
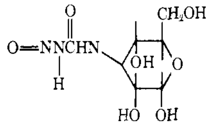
Table VIII. Test Samples Misclassified

NSC no.	Structure	Active neighbor[a]	Neighbor activity[b]	Prediction NN[c]	Prediction LM[d]	EM activity[e]
3088			150+	2-5	NA	Nonactive
			150+			
60339			125–150	5-5	ACT	Nonactive
77037			150+	0-5	ACT	Nonactive
18270			150+	0-5	ACT	Active

[a]Nearest neighbor to test sample for pattern space F421 except for NSC.77037 and 18270. For these samples, the active neighbor was the active compound in the training set with the highest correlation coefficient with the test samples. [b] T/C for active neighbor. [c]The number of times the nearest neighbor method predicted a compound active out of the five pattern spaces, e.g., 0-5 means none out of five. [d]Learning machine prediction for activity: NA for nonactive and ACT for active. [e]Ependymoblastoma activity.

is to recognize that the overall structure of the new compound resembles some active drugs. This type of procedure was mimicked by the nearest neighbor method when large numbers of features were used. In addition, the Euclidean distance measurement gave a qualitative measure of the similarity of compounds since the smaller the distance the greater the similarity. Another way to find potentially active compounds is to recognize important "active" substructural units imbedded in a molecule. The learning machine with the weight–sign change technique mimicked this

Table IX. Test Samples Correctly Predicted as Active

NSC no.	Structure	Active neighbor[a]	Neighbor activity[b]	Prediction NN[c]	Prediction LM[d]	EM activity[e]
17262			150+	2-5	ACT	Active
18269			150+	4-5	ACT	Active
160466			125–150	5-5	ACT	Active

[a] The active neighbor was the active compound in the training set with the highest correlation coefficient with the test samples. [b] T/C for active neighbor. [c] The number of times the nearest neighbor method predicted a compound active out of the five pattern spaces, e.g., 0-5 means none out of five. [d] Learning machine prediction for activity: NA for nonactive and ACT for active. [e] Ependymoblastoma activity.

process by determining which features are important for activity. Then the presence or absence of these features determined the activity of the compound.

These results suggest some interesting applications. One involves using this methodology to prescreen new compounds for biological testing. The pattern recognition methods could rank new compounds by the prediction of their activity and the predicted activities could be assigned higher testing priorities. As other pharmacological activities are added, a compound can be "prescreened" by the computer to direct it to the most promising biological testing. Thus, this methodology could optimize the use of previously collected chemical and biological data in directing current testing. In addition, the predictive capability of this methodology has been implemented on an interactive substructural retrieval system[21] and work is presently underway to apply this methodology to recognizing multiple pharmacological activities for use at the Food and Drug Administration.

A second possible application is in the area of drug design.[22] Presently, QSAR methods are used to find the most active compound in a series, that is, to optimize a "lead" drug. However, these methods cannot generate "new" lead compounds. Hopefully, pattern recognition with substructural analysis can aid in this area of drug development since substructural analysis allows the examination of structurally diverse drugs while pattern recognition methods determine the substructural units important for discriminating between pharmacological classes. For instance, this process may allow the development of "new" lead compounds composed of new combinations of important substructural features.

Conclusions

This study reported the successful use of substructural analysis and the nearest neighbor and learning machine methods to predict the antineoplastic activity of structurally diverse compounds in an experimental mouse brain tumor system. In the substructural analysis, the molecules were represented by three types of substructural units, the augmented atom, the heteropath, and the ring fragments while the pattern recognition techniques were designed to mimic the ways a medicinal chemist might determine potentially interesting compounds given the structures and activities of a series of known drugs.

Acknowledgment. The authors are grateful to J. Mosimann for statistical consultation.

References and Notes

(1) (a) Division of Computer Research and Technology; (b) National Cancer Institute.
(2) P. Ehrlich and S. Hata, "Die Experimentelle Chemotherapie der Spirillosen", Springer, Berlin, 1910.
(3) C. Hansch in "Drug Research", Vol. 1, E. J. Ariens, Ed., Academic Press, New York, N.Y., 1971, pp 271–342.
(4) (a) K. C. Chu, *Anal. Chem.*, **46**, 1181 (1974); (b) R. D. Cramer III, G. Redl, and C. E. Berkoff, *J. Med. Chem.*, 17, 533 (1974).
(5) K. L. Ting, R. C. T. Lee, G. W. A. Milne, M. B. Shapiro, and A. M. Guarino, *Science*, **180**, 417 (1973).
(6) S. A. Hiller, V. E. Golender, A. B. Rosenblit, L. A. Rastrigin, and A. B. Glaz, *Comput. Biomed. Res.*, **6**, 411 (1973).
(7) B. R. Kowalski and C. F. Bender, *J. Am. Chem. Soc.*, **96**, 916 (1974).
(8) C. L. Perrin, *Science*, **183**, 551 (1974).
(9) J. T. Clerc, P. Naegeli, and J. Seibl, *Chimia*, **27**, 639 (1973).
(10) S. H. Unger, *Cancer Chemother. Rep., Part 2*, 4, 45 (1974).
(11) R. Geran, G. F. Congleton, L. E. Dudeck, B. J. Abbott, and J. L. Gargus, *Cancer Chemother. Rep., Part 2*, 4, 53 (1974).
(12) A. Leo, C. Hansch, and D. Elkins, *Chem. Rev.*, **71**, 525 (1971).
(13) G. W. Adamson, M. F. Lynch, and W. G. Town, *J. Chem. Soc. C*, 3702 (1971).
(14) W. C. Cutting, "Handbook of Pharmacology", 5th ed, Appleton-Century-Crofts, New York, N.Y., 1972.
(15) R. J. Feldmann in "Computer Representation and Manipulation of Chemical Information", W. T. Wipke, S. R. Heller, R. J. Feldmann, and E. Hyde, Ed., Wiley-Interscience, New York, N.Y., 1974, pp 55–81.
(16) N. J. Nilsson, "Learning Machines", McGraw-Hill, New York, N.Y., 1965.
(17) T. L. Isenhour and P. C. Jurs, *Anal. Chem.*, **43**, 20A (1971).
(18) D. H. Foley, *IEEE Trans. Inf. Theory*, it-**18**, 618 (1972).
(19) T. L. Isenhour, B. R. Kowalski, and P. C. Jurs, *CRC Crit. Rev. Anal. Chem.*, **4**, 1 (1974).
(20) A. Burger, "Medicinal Chemistry", Part I, Wiley-Interscience, New York, N.Y., 1970, pp 25–245.
(21) K. C. Chu, R. J. Feldmann, and M. Spann, unpublished results.
(22) G. Redl, R. D. Cramer III, and C. E. Berkoff, *Chem. Soc. Rev.*, **4**, 273 (1974).