group, as substitution removes the mechanism from limiting SN2 characteristics. An investigation of the heat capacities of activation associated with the hydrolysis of the above compounds in water[8,11] support the conclusion that by and large the transition state for a nitrate is characterized by a higher degree of ionic character compared to the corresponding chloride. A similar trend in the ratio of isotope effects found for α-phenylethyl chlorides[2] can be understood on the basis of the arguments advanced above.

[11] K. M. Koshy, R. E. Robertson, and W. M. J. Strachan, *Can. J. Chem.*, in press.

K. M. Koshy, R. E. Robertson*
Department of Chemistry, University of Calgary
Calgary, Alberta, Canada
Received September 13, 1973

The Application of Pattern Recognition to Screening Prospective Anticancer Drugs. Adenocarcinoma 755 Biological Activity Test

Sir:

Pattern recognition[1] has been introduced to the chemical literature as a general tool which can be used by the chemist to reduce masses of experimental data to relevant information. Perhaps more importantly, it provides connections between raw, multivariant data and sought-for information without making restrictive assumptions about the underlying statistics of the data. The general problem has been stated as follows. Given a collection of objects and a list of measurements made on each object, is it possible to find and/or predict a property of the objects that is not directly measurable but is known to be related to the measurements *via* some *unknown* relationship? The only assumption made is that similarities and dissimilarities among objects are reflected in at least some of the measurements.

The above stated problem is general indeed, and pattern recognition can and has been used to solve problems in several diverse areas of science and engineering. Of particular interest to chemistry is the somewhat less general problem of learning something about a collection of objects when the objects are chemical compounds and the measurements are physical and/or structural properties of the molecules. Several possibilities exist. A chemist may want to determine the cause of a manufacturing problem by using pattern recognition to detect the discriminatory property or combination of properties between acceptable and unacceptable products. Material problems such as this have been solved by pattern recognition,[2] or, more fundamentally, one might wish to draw a relationship between the structure of a molecule and its activity (reactivity, response, etc.) in some system.

In this communication, a novel example of the latter pattern recognition application is presented. Potential anticancer drugs are screened for their chemotherapeutic activity by applying pattern recognition to 20 selected structural properties (Table I) of 200 drugs previously tested by the National Cancer Institute for activity in the solid tumor Adenocarcinoma 755 (CA 755) screening system.[3] In this test, the drug is administered to small animals with solid tumors, and tumor growth is measured. If the tumor weight inhibition (TWI) is greater than 70%, then the drug is considered positive and reproducible as an antineoplastic agent. Toxic molecules were not included in this study. Of the 200 drugs in this study, 87 had values above 70% and were included in the "positive" category, and 113 had values less than 50% (39 were 0%) and were included in the "negative" category. Some of the drugs in the study are currently in clinical chemotherapeutic use.

Table I. Features Used for the CA 755 Study

Feature no.	Feature	Variance wt	Rank	Correlation
1	Number of oxygens/number of atoms	1.36	9	−
2	Number of phosphorus/number of atoms	0.88	19	−
3	Number of sulfurs/number of atoms	3.99	1	+
4	Number of halogens/number of atoms	1.48	6	+
5	Number of carbons/number of atoms	1.48	5	+
6	Number of C—S bonds/number of carbons	3.86	2	+
7	Number of C═C bonds/number of carbons	1.53	4	+
8	Number of C—N bonds/number of carbons	1.27	12	−
9	Number of C—O bonds/number of carbons	1.39	8	−
10	Number of C═O bonds/number of carbons	1.08	14	−
11	Number of N—H bonds/number of nitrogens	1.34	10	−
12	Number of O—H bonds/number of oxygens	1.29	11	−
13	Number of PO_4 groups	0.88	20	−
14	Number of S—H bonds	2.35	3	+
15	Purine derivative	1.00	16	+
16	Pyrimidine derivative	1.00	17	−
17	Number of oxygens in rings	1.09	13	−
18	Number of nitrogens in rings	1.05	15	+
19	Number of phenyl groups	1.45	7	+
20	Substitution at the primary nitrogen in purine or pyrimidine	0.97	18	−

In order to limit the number of drugs studied in early experiments, it was decided to study a particular class of drugs instead of selecting drugs at random. The purine and pyrimidine nucleoside derivatives form a class of drugs that have produced several drugs of clinical interest. From a summary[3] of compounds in this class, 50 structural properties were extracted from the 200 drugs mentioned above. Preliminary data analysis was performed in order to eliminate several of the structural features from the study. First, 14 features were eliminated because of their scarcity in the 200 structures. Then 16 more were eliminated because they contained little or no useful information relating to biological activity in CA 755 as determined by variance weighting.[1] The remaining 20 structural features used in the study are listed in Table I.

Each of the 20 features was autoscaled[1] for the first stage of preprocessing. Autoscaling weights all of the features equally by producing new variables with zero mean and unit standard deviation. Then, the variance

[1] B. R. Kowalski and C. F. Bender, *J. Amer. Chem. Soc.*, **94**, 5632 (1972).

[2] B. R. Kowalski, *Chem. Tech.*, in press.

[3] A. Goldin, H. B. Wood, Jr., and R. R. Engle, *Cancer Chemother. Rep.*, 1 (1), 1 (1968).

weight[1] was calculated for each feature in order to determine the feature's usefulness in discriminating the positive drugs from the negative drugs. The variance weights, also found in Table I, were applied to the autoscaled features in order to enhance the discriminatory ability of the features. In Table I, the variance weights have been normalized to the purine (feature 15) and pyrimidine (feature 16) features which are the same because only purine and pyrimidine nucleoside derivatives were present in the 200 drugs. The "rank" column in Table I is the order of importance for each feature. This information, along with the results of other pattern recognition methods, can be extremely useful, and its importance to the prospective synthesis of new drugs cannot be overemphasized. Not only can chemical compounds be screened for their activity in a biological system, as will be shown later, but important feedback information can be made to the synthetic chemists who are always looking for improved methods of "drug design." The structural features or chemical and physical properties most discriminating in the retrospective studies can be emphasized in prospective studies.

The number of sulfur atoms normalized to the total number of atoms (Table I) is seen to be the most discriminating feature in this study. Of the sulfur-containing drugs in the study, 85 drugs had at least one C–S bond and only 37 had an S–H bond. However, closer examination of Table I shows that, although sulfur bonded to hydrogen is important, sulfur bonded to carbon is actually more important. Other important features indicate that unsaturation and phenyl and halogen substitutions are also important. The sign under "correlation" in Table I gives the direction of importance for each feature. The sign for feature 6 (C–S bonds) is positive, indicating that the greater the number of C–S bonds in a molecule the greater the tendency for the drug to be in the "positive" category. Conversely, the negative sign for feature 10 indicates that components in the "positive" category tend to have few or no carbonyls. All of these results are, of course, meaningful only to purine and pyrimidine nucleoside derivatives.

The 20 scaled and weighted features serve to position each of the 200 drugs in a 20-dimensional plot containing 200 points. The object of the pattern recognition methods discussed in the remainder of this communication is to develop classification rules that can separate the drugs in the positive category from those in the negative category.

Figure 1 shows an eigenvector projection[4] of the 200 points in 20-space. The eigenvector projection is a linear projection to 2-space and is optimized in the sense of preservation of variance. In Figure 1, 62% of the total variance information in the 20 features is preserved. Although the plot is a bit congested and overlap is considerable, an extremely encouraging separation of "positive" drugs (marked "2" in the plot) from "negative" drugs (marked "1" in the plot) is evident, suggesting that the application of classification procedures is warranted.

Since the ratio, R, of the number of patterns (drugs) to features is high ($R = 10$) and the problem is a 2-category dichotomization, separations are realistic for reasons discussed in an earlier paper.[1] Now, since a nonparametric classification method must be tested to determine its classification accuracy and since the maximum number of drugs should be used to develop or train the classification rule, the leave-one-out[1] procedure was used here. The first drug is removed from the training set, and the classification rule is developed on the remaining 199 drugs. The lone drug is then classified and returned to the training set, and the second drug is removed and training and classification repeated. This process is repeated 200 times, and the final results are reported on the 200 individual classifications. This procedure gives the best estimate of classification performance but is somewhat expensive.

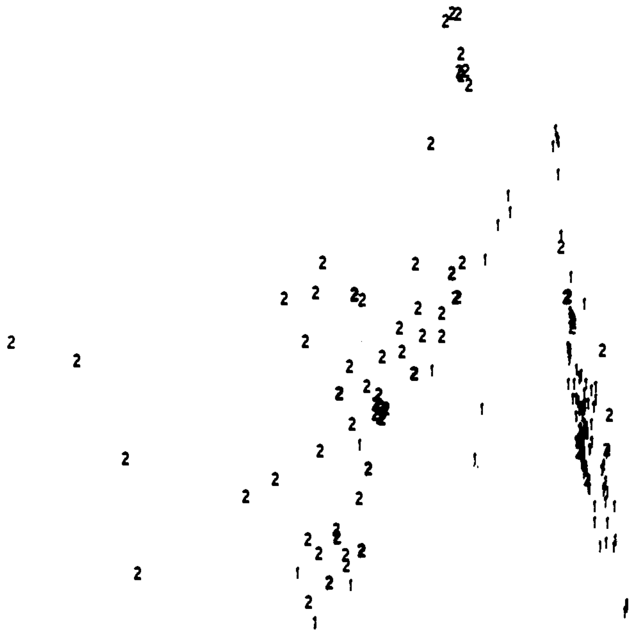


Figure 1. Training set for raw data.

Three classification procedures were applied to the autoscaled and weighted data: the k nearest neighbor rule[5] (kNN) with $k = 3$, a linear discriminant function trained by least squares,[6] and a linear discriminant function trained with an error correction procedure.[7] Table II shows the classification results using the three methods. The kNN method produced excellent results with slightly higher accuracy in detecting "positive" drugs. The linear discriminant function trained

Table II. Classification Results

Method	Per cent correct, positive	Per cent correct, negative	Per cent correct, total
k nearest neighbor	94.2	92.9	93.5
Least squares	89.7	90.3	90.0
Feedback learning	93.8[a]	88.5[a]	91.5[a]

[a] Data *not* linearly separable.

[4] B. R. Kowalski and C. F. Bender, *J. Amer. Chem. Soc.*, **95**, 686 (1973).

[5] B. R. Kowalski and C. F. Bender, *Anal. Chem.*, **44**, 1405 (1972).

[6] B. R. Kowalski, P. C. Jurs, T. L. Isenhour, and C. N. Reilley, *Anal. Chem.*, **41**, 695 (1969).

[7] N. J. Nilsson, "Learning Machines," McGraw-Hill, New York, N. Y., 1965.

by least-squares also performed quite well with about equal performance on the "positive" and "negative" drugs. The least-squares results are even more encouraging in light of the fact that the procedure used 1 order of magnitude less computer time. This may be more significant in the much larger studies currently in progress. The feedback learning results were also encouraging even though the two classes are not linearly separable.

In conclusion, the results of this first study show that the activity of a chemical compound can be predicted by computer analysis, using pattern recognition methodology, of structural features obtained from the molecule. It is not suggested that computer screening can or will eliminate the need for biological testing. Instead, computer screening can be used to provide priorities for drugs yet to be tested in over-taxed testing programs and can aid significantly in prospective studies by analyzing previously tested molecules and providing a more rational approach to "drug design." Pattern recognition "learns" about biological activity by processing data from retrospective studies and making connections to the structural features of the drugs. This advance can save the scientist much time and money.

Although several other screening applications are either in progress or planned for the near future, emphasis is currently on screening studies for anticancer activity. Computer screening applications in five other systems are in progress. These include the L1210 (leukemia) and Walker 256 (solid tumor) systems which have received most of the emphasis in the cancer chemotherapy program.

The Sarcoma 180, Lewis Lung, and KB tissue culture tests are also being studied. Results are equally encouraging and are forthcoming. Better structure coding methods are being considered to replace the more naive structural features used in this study. Computer extraction of molecular substructures from connection tables[8] and the Wiswesser line notation[9] are probably the most fruitful approaches being contemplated at this time. Also, the inclusion of chemical properties and pharmacological information will also be important. New data on several other classes of biologically active compounds are currently being amassed.

Classification performance will also improve when powerful preprocessing methods (*e.g.*, orthogonalization) within pattern recognition are applied to the input features. Also, unsupervised learning methods have yet to be applied. These and other improvements promise to aid significantly in screening applications and rational "drug design" for cancer chemotherapy as well as for any system where better and more active chemical compounds are needed.

Acknowledgments. The authors would like to express their gratitude to George Hess, M.D., University of California, Davis Medical School, for his help and encouragement and to Paul Murphy, Colgate University, for coding the structural features of the molecules used in this study. B. R. K. gratefully acknowledges the financial support of the National Science Foundation (GP-36578X) and the Office of Naval Research.

[8] M. F. Lynch, J. M. Harrison, W. G. Town, and J. E. Ash, "Computer Handling of Chemical Structure Information," American Elsevier, New York, N. Y., 1971.

[9] E. J. Smith, "W. J. Wiswesser's Line Formula Chemical Notation," McGraw-Hill, New York, N. Y., 1968.

[10] This work was performed under the auspices of the U. S. Atomic Energy Commission.

B. R. Kowalski*
Department of Chemistry, University of Washington
Seattle, Washington 98195

C. F. Bender[10]
University of California, Lawrence Livermore Laboratory
Livermore, California 94550

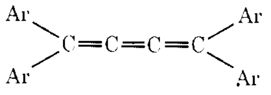
Received August 25, 1973

Photochemistry of Crystalline Cumulenes. Reassignment of the Structure of the Solid-State Photodimer of Tetraphenylbutatriene

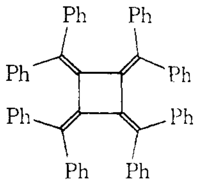
Sir:

The role of the crystal structures of substituted ethylenes in determining the stereochemistry of the cyclobutane photodimers obtained therefrom has been demonstrated.[1] A separation of 4.0 ± 0.2 Å between potentially reactive C=C double bonds is necessary for photodimerization. However, it has been observed[1,2] that neighboring C=C bonds, which fulfill the above requirement but which are far offset implying insufficient overlap of their π-electrons, do not dimerize. A system which allows systematic analysis of the importance of the alignment of the π-electrons is provided by the cumulenes in which the π-lobes of alternate C=C bonds lie in mutually perpendicular planes. We undertook photochemical and crystallographic studies of several substituted butatrienes.

The tetraarylbutatrienes (I and II) dimerize in the solid state.[3] The dimer of I was reported to be the radialene III.[4] On the assumption that the reaction



I, Ar = C_6H_5
II, Ar = p-$CH_3OC_6H_4$



III

I–III must involve pronounced overlap between the π-electrons in the ground state of the neighboring C=C bonds,[1,2] the monomer butatriene molecules must approach each other edge-on. Furthermore, steric factors permit the potentially reactive 4 Å approach between the central C=C bonds only if the molecules would be crossed in the unusual structure of Figure 1. However, in the course of our studies, we found that

[1] G. M. J. Schmidt, *Pure Appl. Chem.*, **27**, 647 (1971).

[2] (a) L. Leiserowitz and G. M. J. Schmidt, *Acta Crystallogr.*, **18**, 1058 (1965); (b) M. Lahav and G. M. J. Schmidt, *J. Chem. Soc. B*, 312 (1967); (c) N. J. Leonard, R. S. McCredie, M. W. Logue, and R. L. Cundall; *J. Amer. Chem. Soc.*, **95**, 2320 (1973); (d) J. K. Frank and I. C. Paul, *ibid.*, **95**, 2324 (1973).

[3] (a) K. Brand, *Ber.*, **54**, 1987 (1921); (b) K. Brand and F. Kercher, *ibid.*, **54**, 2007 (1921).

[4] (a) R. O. Uhler, H. Shechter, and G. V. D. Tiers, *J. Amer. Chem. Soc.*, **84**, 3397 (1962); (b) R. O. Uhler, *Diss. Abstr.*, **21**, 765 (1960). This work has been reviewed [M. P. Cava and M. J. Mitchell, "Cyclobutadiene and Related Compounds," Academic Press, New York, N. Y., 1967, Chapter 5; A. Schönberg, "Preparative Organic Photochemistry," Springer Verlag, New York, N. Y., 1968, p 89]. In terms of the correct structure IV the mechanism of its reactions with alkali metals should be revised: R. Nahon and A. R. Day, *J. Org. Chem.*, 30, 1973 (1965).