Relating Mutagenicity to Chemical Structure†

JOHN TINKER
Eastman Kodak, Rochester, New York 14650

Received July 31, 1980

A computer program correlating chemical structure with mutagenesis activity has been developed as a predictive test for hazard evaluation. Sets of substructural units are derived from complete structures, and the probability that a structure containing a given unit will have a designated category of activity is calculated. The program, validated by bacterial mutagenesis testing, is capable of identifying similar structures, showing why they are similar, and estimating the activity of a structure.

THE CHALLENGE

A reliable assessment of chemical hazard may require many tests, which may be needed for many chemical substances, including new compounds and compounds already in circulation, natural as well as synthetic (i.e., compounds in the environment as well as their metabolites and degradation products). Any of these may result in exposure, in the workplace or the environment, to industrial workers, consumers, or the general public.

There are many properties that are pertinent to evaluating hazard. These include lethality to experimental animals, capacity to irritate and sensitize, and prolonged effects such as mutagenicity, teratogenicity, and carcinogenicity. In assessing the environmental effects of a chemical we may need to know the response of invertebrates, fish, and possibly birds and plants.

For many of these effects, relatively few results are known in comparison to the number of chemicals that could be tested. For some of them, test procedures are not standardized, and for others there is often no general agreement on the testing protocols. Actual testing of a compound involves cost and time delays, in choosing both what should be tested and how it should be tested. The resources for decision, synthesis or isolation, and testing are limited. There is not the money, time, laboratory space, or personnel to test everything that could be tested. The selection of appropriate tests is therefore of great importance. Guidance as to the advisability of committing resources to a testing program is also needed. In consequence, there is a clear need for predictive tests for screening, prioritizing, and guidance in biological testing.

PREDICTIVE TESTS

If a procedure can be done relatively quickly and cheaply and yields results that can be regarded as reliable (i.e., correlate well with results obtained in more extensive testing), it would have two advantages. It could indicate what further testing would be appropriate, and it could serve as an early warning of possible hazards. A procedure that does not require a sample would be the quickest and cheapest. Traditionally, the procedure used is memory: someone becomes acquainted with the purpose and use of a test, memorizes the results for many chemical structures, and estimates the outcome for a new structure on the basis of similarity. Computer structure–activity correlation is a refinement of this process.

The basic assumption is that property depends on structure. A property is exhibited to varying degrees by different structures. It is natural to look for smaller parts of the structures, pieces that are identical or similar and account for the property or influence it. Many examples come to mind. Woodward's rules[1,2] correlate the ultraviolet absorption peak of α,β-unsaturated ketones. The position and identity of chromophores influence the hue of dyestuffs.[3-8] Chemical engineers can estimate many physical properties by adding contributions made by each atom and some functional groups. The general validity of this approach seems to be well accepted.

CORRELATION STRATEGY

After reviewing our requirements and surveying the available computer techniques, we chose to try qualitative correlation. To begin, activity data and structural features are needed.

1. Treatment of Activity Data. Activity data are expressed as levels of activity: low, medium, and high, for example. Estimates are made in terms of the same levels. The number of levels, or categories of activity intensity, can vary as desired from as few as two to as many as eight. For data expressed in quantitative terms, a category is defined as a range of values. For example, high acute oral toxicity is defined as LD_{50} smaller than 50 mg/kg. The definitions of activity categories can be changed from run to run.

2. Structural Features. It is possible to identify the important structural features during the statistical analysis. However, it is sufficient to choose an algorithm that will derive from the complete structure a set of structural features, provided that the important features are included.

3. Choosing the Algorithm. We must choose a set of substructure units to distinguish among groups of structures; a substructure present in all examples will not contribute to the distinction. A unit as small as a carbon atom is a poor choice because it appears in nearly every example. Nor can we choose a unit so large that it occurs only rarely: in this case similarity would be a rare phenomenon. A precedent for unit selection is given in the work of Hodes et al.[9] in relating structures to antitumor activity. These investigators used a unit of three connected nonhydrogen atoms, together with the terminal bonds, and referred to them as triplet ganglia (Figure 1). Any assortment of substructure units could be feasible, the selection being essentially pragmatic.

Were we to use such a set of units, equally weighted, in an ordinary statistical analysis, the results would be suspect because there are likely to be more adjustable parameters than data to determine them. The statisticians say that there are too few degrees of freedom: the dimensionality is too large. Operating on a large number of structures, the algorithm finds not only a large number of units for them all but, from one run to another, a variable number of units, because a new structure can contain a new string of atoms. That is, in continuing operation, the dimensionality in terms of the units is large and variable. The essential way of reducing the dimensionality is described by Hodes et al.[9] The weight, for each combination of substructural unit and category of activity, is calculated to give the probability that a structure containing the unit would have the designated category of activity. A

† Presented on April 23, 1980, as part of the Symposium on Development and Use of Reliable Data Bases for Quantitative Structure–Activity Relationships (QSARs) during the 14th Middle Atlantic Regional Meeting of the American Chemical Society, King of Prussia, Pa.

0095-2338/81/1621-0003$01.25/0

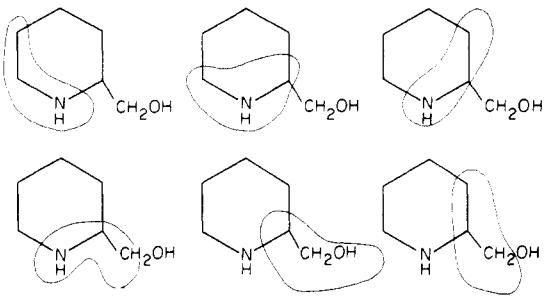

Figure 1. Triplet ganglia of Hodes et al., three connected nonhydrogen atoms, together with terminal bonds.

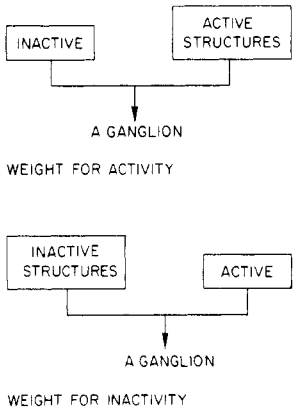

Figure 2. Weights for inactivity and activity.

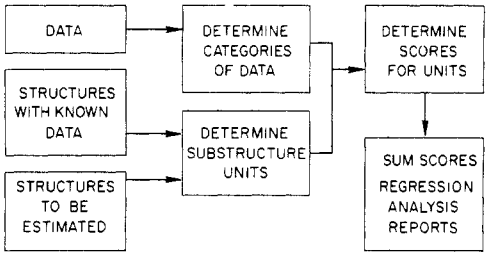

Figure 3. Organization of the program for substructure analysis and category of activity correlation.

weight is then given to each unit for each category of activity. The weight depends on the distribution of compounds in various categories. If a unit is present in two active compounds and six inactive ones, its weight is for inactivity; if another unit is present in ten active but in no inactive compounds, it receives a large weight for activity (Figure 2). The numeric value is a conditional probability. When these weights are summed for the units in a structure, they give a result which is the probability that the structure will possess the corresponding category of activity.

PROGRAM FOR SUBSTRUCTURE ANALYSIS AND CATEGORY OF ACTIVITY CORRELATION

Figure 3 shows the organization of the program. Given structures and activity data, the program calculates the degree to which the activity depends on a structure, averaged over all data. It calculates, for each structure, the category of activity and confidence rating. The degree of the dependence of activity on structure is reported as the confidence limit at which each category is statistically distinct from each of the other categories. For every structure submitted to the program, the activity category is calculated on the basis of the entire set of data whether or not its activity is known. In addition, the probability of the correctness of assigning a structure to other categories of activity is calculated on the basis of the sample size[10] in the categories. This value is an indication of the reliability of the activity prediction for each structure.

Table I. Fences Used to Categorize Data Obtained from Bacterial Mutagenesis Test

fence	unqualified data, revertants/nmol	qualified data, revertants/nmol
1	1.0 E−03[a]	0.0 E+00
2	1.0 E−01	4.0 E−01
3	1.0 E+01	4.0 E+01
4	1.0 E+03	8.0 E+00

[a] Low values imply low activity.

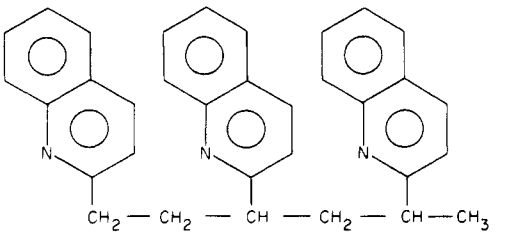

Figure 4. Vinyl polymer, represented as a trimer, with the terminal vinyl group unsaturated.

Any sort of data may be used. It may be qualitative or quantitative and expressed in different ways. For each run, the way each kind of expression is to be categorized must be explained.

An example of categorizing data is seen in the treatment of mutagenesis data from a thousand compounds selected from several hundred literature papers. Most of these are derived from tests of compounds using Ames' *Salmonella typhimurium* test. The results are expressed as revertants per nanomole of test compound. Data from several other tests have been included, for instance, tests based on *Escherichia coli, phage T4, Aspergillus* sp., *Saccharomyces* sp., and *Actinomyces* sp. Five categories, corresponding to inactive, weak, moderate, strong, and potent bacterial mutagens, are defined by the separation points or fences (Table I). From the numeric data from the *Salmonella* test these fences are selected as 10^{-3}, 10^{-1}, 10, and 1000 revertants/nmol. For qualitative data, the category in the literature is used as reported. We are using five activity classes for the bacterial mutagenesis data, although the program can handle as many as eight activity categories.

Structures are accepted by the program if expressed as *Chemical Abstracts* Connection Tables, Mechanical Chemical Codes ciphers,[11] or a mixture of both. Polymers are represented as trimers, with the terminal vinyl group of a vinyl polymer saturated, as shown in Figure 4. In this way all of the appropriate substructural units are included. The difference in molecular weight between the representation and the actual polymer is not important to this analysis, since the molecular weight does not appear explicitly in the calculations. Mixtures of substances must be treated as individual components.

The complete structures are used by the program to derive almost the same substructural units as in the Hodes et al. program. The calculation of weights for each substructural unit contributing to each category is done as in the strategy of Hodes et al. The data are then analyzed in the program by a stepwise discriminant analysis program.

In summary, the program is exactly like the one developed at the National Cancer Institute to estimate antitumor activity, except for the use of other data and the following program modifications.

MODIFICATIONS TO THE PROGRAM

The presently described program includes the following modifications:

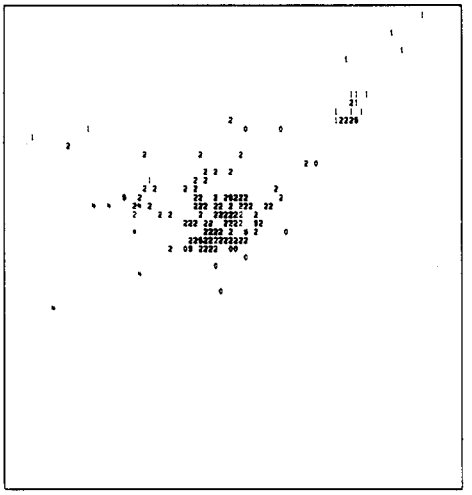

Figure 5. Diagram of distribution of activity into categories, showing weak activity (code 1) in the upper right and moderate activity (code 2) in the center.

(1) As many as eight activity categories may be specified.

(2) Categories may be defined separately for various activity codes and various types of data and may be combined for an analysis.

(3) Confidence limits for statistically significant differences between each pair of categories are calculated.

(4) For each substance, the input category, the calculated category, the distance in arbitrary units from the mean of each category, and the probability of classification in other significant categories are reported. The reports are sorted in several ways, so that a substance can be located easily if its registry number is known or its position is inconsistent with the rest of its category.

(5) Data can be traced through the program. For each structure, it is easy to find the substructures that contribute to its categorization and locate the references that contain the test results pertinent to that estimation. Thus, the resemblance of unknown to known structures, deduced by the program, is exactly laid out for consideration by the users of the program.

ADVANTAGES OF THE APPROACH

There are several advantages to the program.

(1) All of the activity data may be used. This includes mixed quantitative and qualitative data. They need not be perfectly consistent or statistically separated. Wrong data are tolerated and can be detected, as described later.

(2) The structures need not be closely related. A wide variety of structures can be used in a run, both with known results and results to be estimated.

(3) Several thousand structures can be used in a run and several hundred estimated at the same time, at a reasonable cost.

(4) Editorial effort is minimized. The program is tolerant of moderately inconsistent data and can make many estimates in a single run without recursive analysis.

(5) There is no need to discard data and no way to modify or alter the results by changing substructures or adjusting activity data.

VALIDATION OF THE PROGRAM WITH BACTERIAL MUTAGENESIS DATA

An actual run of the program was made by using bacterial mutagenesis test results on 1019 structures drawn from the literature and entered into the data base as described above. The distribution of activity into categories is shown in the diagrams of the decision space (Figures 5–10). These represent a two-dimensional view of the five-dimensional analysis space, viewed from the angle giving the best separation between

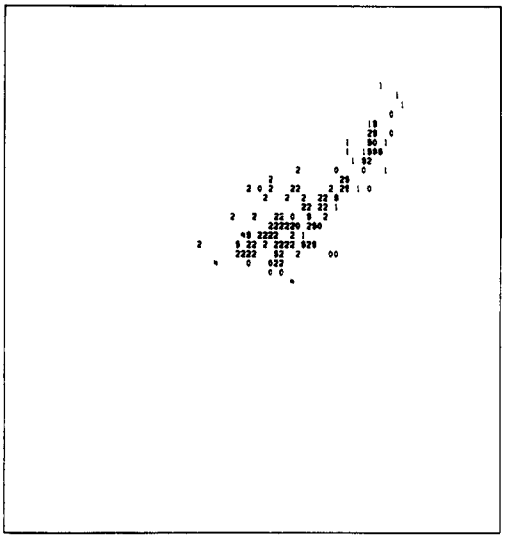

Figure 6. Diagram of distribution of activity into categories, showing weak activity (code 1) in the upper right and moderate activity (code 2) in the center. The overlapping of two or more activities is indicated by "$".

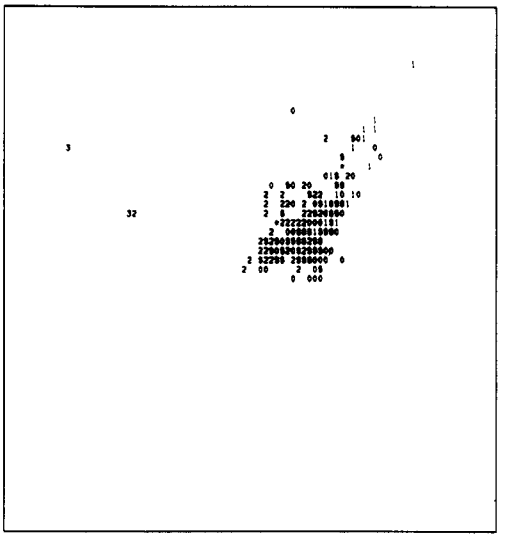

Figure 7. Diagram of distribution of activity into categories, showing moderate activity (code 2) in the center.

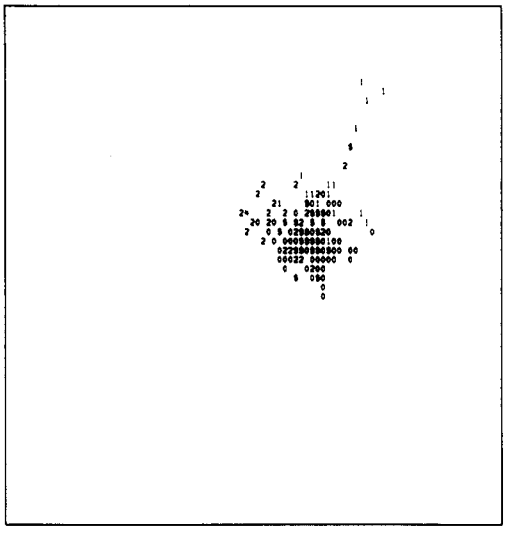

Figure 8. Diagram of distribution of activity into categories, showing weak activity (code 1) in the upper right.

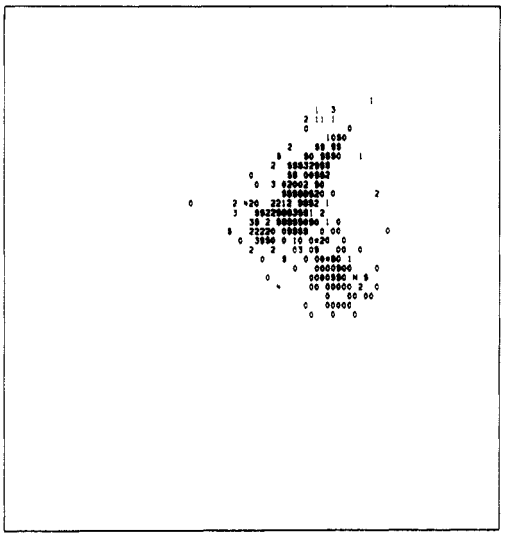

Figure 9. Diagram of distribution of activity into categories, showing inactivity (code 0) in the lower right.

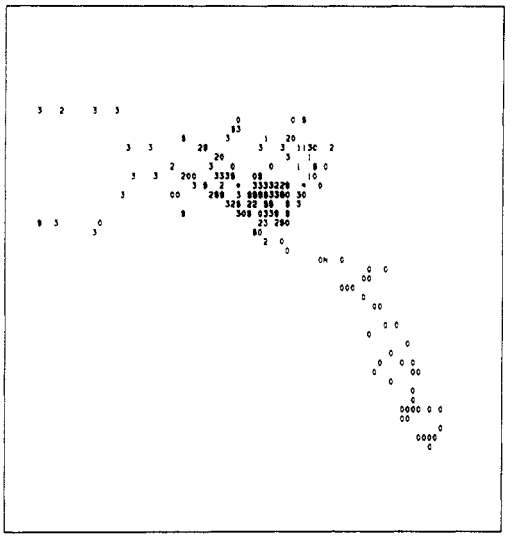

Figure 10. Diagram of distribution of activity into categories, showing strong activity (code 3) on the left and inactivity (code 0) on the right.

the means and divided into six parts along the third dimension. The separation of groups is apparent. The inactive category is in the lower right of the fifth and sixth slices (Figures 9 and 10), the weak to the upper right in slices 1 and 2 (Figures 5 and 6), the moderate in the center of the first three slices (Figure 5–7), the strong category on the left of the last slice (Figure 10), and the potent on the right in the first slice (Figure 5). Several points for the potent group lie off the edge to the lower right of slice 1 (Figure 5). The groups are not perfectly separated. Note the point coded moderate among many inactives on the lower right of slice 5 (Figure 9). It is the point for 7-methylguanine.

We need to test the distribution for internal consistency. Thus if knowns were reentered as unknowns and the predictions were the same as the actual values, the internal consistency would be perfect.

For an actual estimation run, the predicted categories for all of the known data tested in this manner were identical with the original categories in 66% of the cases and was within one category for 84% of them. This result gives us confidence that on the whole the data as categorized by the assigned fences are internally consistent to an acceptable degree.

The size of the data base has a strong influence on the optimum choice of substructure units and reliability of the predictions. When 700 structures are used, the triplet ganglia

Table II. Comparison of Structure–Activity Correlation and Bacterial Mutagenesis Testing of 34 Compounds

	number of compounds	
	results estimated at good confidence	results estimated at fair confidence
same category	22	4
one category higher	1	
one category lower	3	
two categories lower	4[a]	

[a] Some doubt regarding two tests.

Table III. Test Reclassification of Known Data

new classification (group)	original classification (group)				
	0[a]	1	2	3	4
0[a]	222	39	44	11	4
1	26	117	16	6	3
2	71	43	254	30	9
3	10	3	15	66	9
4	2	0	3	3	13
N[b]	17	9	17	1	0

[a] Represents low activity. [b] Represents unknown.

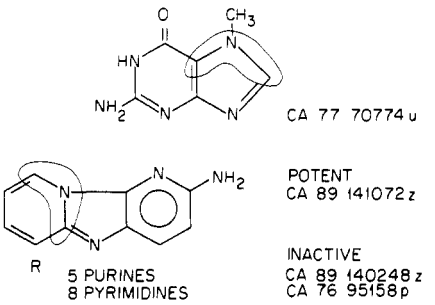

Figure 11. Substructural unit, referenced by *Chemical Abstracts* numbers and having a strong weight for potency. The triplet ganglion is outlined.

are slightly inferior to the triplet units without the terminal bonds, and the program estimated the correct category in only 42% of the cases. When 800 structures are used, the triplet ganglia are superior, and 59% of the estimates agreed with the published values.

Validation of the program was also tested by using compounds simultaneously submitted for bacterial mutagenesis testing and structure–activity correlation. Results are presently available for 34 compounds and are presented in Table II. Of these, 74% agree exactly with the predicted category, and 88% agree within one category. Of the estimates that are one category away, one was high and four were low. However, the program does not seem to have a bias, as supported by a test reclassification of known data (Table III). Thus 176 points were reclassified one or more categories lower, and 171 one or more higher.

IDENTIFYING CONTRIBUTING STRUCTURES

A particular value of the program is that structures contributing to an activity or inactivity finding in the analysis can be located and identified. Thus, 7-methylguanine was incorrectly entered as a moderate bacterial mutagen,[12] but the program correctly predicted that it should be inactive.[13,14] The structural analogy that the program locates in the case of 7-methylguanine can be followed.

There are several substructural units which make important contributions. The structures with known results and references are shown in Figures 11–13. The first unit (Figure 11) occurs in 15 structures. The two glutamic acid pyrolysis products, both potent mutagens, are responsible for the strong weight of this unit for potency.

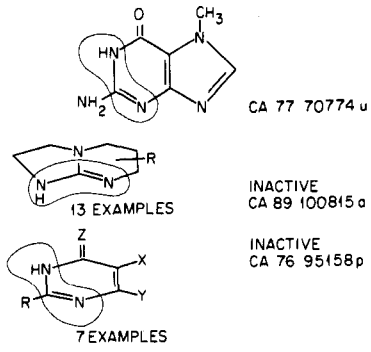

Figure 12. Substructural unit, referenced by *Chemical Abstracts* numbers and having a strong weight for inactivity. The triplet ganglia are outlined.

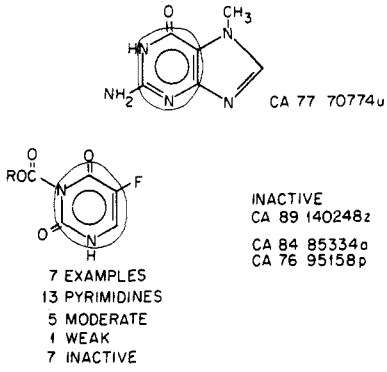

Figure 13. Ring subunit for a six-membered ring, referenced by *Chemical Abstracts* numbers and weighted for inactivity.

The next ganglion occurs in 24 structures, of which 20 are inactive. Thus its weight is strong for inactivity. Figure 12 shows the structures and references.

The third most important contributor is a ring subunit for a six-membered ring with four carbons, two nitrogens, and one ordinary double bond. For this ganglion, aromatic and tautomeric bonds are not counted. This unit occurs in 34 structures, 23 of which are inactive, so that its weight is for inactivity. The structures and references are shown in Figure 13. The contributions of the other subunits can be traced similarly.

Thus the program has revealed a large number of similar structures, most of them inactive. We would agree with its estimate that, on this basis, 7-methylguanine is inactive, contrary to the manner in which it was inadvertently entered.

SUMMARY

The statistical strategy of Hodes et al. has been embedded in a program that is capable of analyzing structures and data—which need not be selected or be consistent—and locating similar structures, showing exactly why they are similar, and estimating by analogy the activity of a structure.

REFERENCES AND NOTES

(1) Ostercamp, D. L. "Vinylogous Imides, II. Ultraviolet Spectra and the Application of Woodward's Rules", *J. Org. Chem.* **1970**, *35*, 1632–1641 (*Chem. Abstr. 73*, 3147w).
(2) Liljefors, T.; Allinger, N. L. "Conformational Analysis. 128. The Woodward-Fieser Rules and α,β-Unsaturated ketones", *J. Am. Chem. Soc.* **1978**, *100*, 1068–1073 (*Chem. Abstr. 89*, 59514m).
(3) Chu, K.-T.; Griffiths, J. "Color and Constitution of the Nitro- and Dinitro-*p*-phenylenediamines and their *N*-methyl Derivaties", *J. Chem. Soc., Perkin Trans. 1.* **1978**, 1194–1198 (*Chem. Abstr. 90*, 86166b).
(4) Klessinger, M. "The Constitution and Light Absorption of Organic Dyes", *Chem. Unserer Zeit* **1978**, *12*, 1–10 (Chem. Abstr. *88*, 171766a).
(5) Daehne, S. "Color and Constitution: One Hundred Years of Research", *Science (Washington, D. C.)* **1978**, *12*, 1–10 (*Chem. Abstr. 88*, 169182p).
(6) Griffiths, J. "Color and Constitution of Organic Molecules"; Academic Press: London, 1976 (*Chem. Abstr. 86*, 155129y).
(7) Hida, M. "Dyeing. Color and Chemical Constitution of Dyes", *Kagaku Kyoiku* **1976**, *24*, 70–77 (*Chem. Abstr. 86*, 6373w).
(8) Coates, E. "Color and Constitution", *J. Soc. Dyers Colour.* **1967**, *83*, 95–111 (*Chem. Abstr. 67*, 69101j).
(9) Hodes, L.; Hazard, G. F.; Geran, R. I.; Richman, S. "A Statistical-Heuristic Method for Automated Selection of Drugs for Screening", *J. Med. Chem.* **1977**, *20*, 469–475 (*Chem. Abstr. 86*, 114985f).
(10) Lachenbruch, P. A. "On Expected Probabilities of Misclassification in Discriminant Analysis, Necessary Sample Size, and a Relation with the Multiple Correlation Coefficient", *Biometrics* **1968**, 823–834.
(11) Lefkovitz, D.; Gennaro, A. R. "Utility Analysis for the MCC (Mechanical Chemical Code) Topological Screen System", *J. Chem. Doc.* **1970**, *10*, 86–94 (*Chem. Abstr. 73*, 10546d).
(12) Kononova, S. D.; Korolev, A. M.; Eremenko, L. T.; Gumanov, L. L. "Mutagenic Effect of Some Esters of Nitric Acid on Bacteriophage T4B", *Genetika.* **1972**, *8*, 101–108 (*Chem. Abstr. 77*, 70774u).
(13) Lakings, D. B.; Waalkes, T. P.; Borek, E.; Gehrke, C. W.; Mrochek, J. E.; Longmore, J.; Adamson, R. H. "Composition, Associated Tissure Methyltransferase Activity, and Catabolic End Products of Transfer RNA from Carcinogen-Induced Heptamona and Normal Monkey Livers", *Cancer Res.* **1977**, *37*, 285–292 (*Chem. Abstr. 86*, 53628t).
(14) Frei, J. V.; Swenson, D. H.; Warren, W.; Lawley, P. D. "Alkylation of Deoxyribonucleic Acid in Vivo in Various Organs of C57BL Mice by the Carcinogens *N*-methyl-*N*-nitrosourea, *N*-ethyl-*N*-nitrosourea and Ethyl Methanesulfonate in Relation to Induction of Thymic Lymphoma: Some Applications of High-Pressure Liquid Chromatography", *Biochem. J.* **1978**, *174*, 1031–1044 (*Chem. Abstr. 90*, 146748e).