An Assessment of Carcinogenicity of N-Nitroso Compounds by the SIMCA Method of Pattern Recognition†

W. J. DUNN III*

Department of Medicinal Chemistry, College of Pharmacy, University of Illinois at the Medical Center, Chicago, Illinois 60612

SVANTE WOLD

Research Group for Chemometrics, Umea University, S 901 87 Umea, Sweden

Received July 31, 1980

The ability to predict the toxic responses of potential environmental pollutants on the basis of their physicochemical properties has many advantages. Pattern recognition methods can be used to predict such pharmacological properties. In this report the SIMCA method of pattern recognition is used to predict the carcinogenicity of N-nitroso compounds, and the advantages of this method of pattern recognition in such applications are discussed.

I. INTRODUCTION

The evaluation of new and untested compounds for their potential to induce toxic effects in genetic material has become a major focus of research in both academic and industrial laboratories. Substances which induce genetic mutations can be a source of cancer and birth defects, for example. Therefore, it is important that the genetic toxicity of compounds which can enter the environment be known.

A number of model systems have been developed for the rapid assessment of the genetic toxicity of newly synthesized and untested compounds.[1] Most of these methods rely on the use of various microorganisms as model systems for mammalian metabolism and biochemical function. Other methods of assessment rely on exposing test animals, such as laboratory rats and mice, to potential environmental mutagens and carcinogens and observing the biological end points of these substances in the test sample.

Both approaches have their advantages and disadvantages. The microorganism-based methods have the advantage of being fast and relatively inexpensive. The whole animal method, while producing results which may better portray the effect of compounds on the human population, is time consuming and expensive. Both approaches require that an actual sample of the compound be in hand. It would be an advantage to be able to obtain an estimate of a compound's harmful effects strictly on theoretical grounds (i.e., without having to prepare the compound).

This problem can be stated as one of classification. One attempts to classify compounds as active or nonactive on the basis of their theoretical properties. A number of such classification methods, called pattern recognition methods, are available and have been applied to problems in structure–biological activity.[2-6] These applications include some recent efforts to assess the carcinogenicity of compounds.[7]

The SIMCA method of pattern recognition has been applied to a number of structure–chemical reactivity problems.[8] We have applied the method with considerable success to a number of problems of a structure–biological activity nature.[9] With regard to the problem of structure–carcinogenicity, the SIMCA method has been used in an attempt to assess the carcinogenic potential of 4-nitroquinoline 1-oxides,[10] polycyclic aromatic hydrocarbons,[11] and more recently N-nitroso compounds.[12] In these reports, not only was the method used to classify compounds as active or nonactive but information beyond classification was extracted from the data. This advantage of SIMCA to operate beyond classification when applied to structure–biological activity data will be discussed below. Following this, the application of SIMCA to structure–rat carcinogenicity data of N-nitroso compounds will be presented.

II. LEVELS OF CLASSIFICATION

We have recently formulated various levels of information which can be extracted from appropriate structure–(re)activity data.[13] Since this discussion deals specifically with structure–carcinogenicity data analysis, this topic will be presented in these terms.

Level I. Classification into one of a number of defined classes constitutes level I classification. An example is classifying a compound as a carcinogen of a specific mechanistic type when it is recognized that more than one mechanistic type is possible. Most methods of pattern recognition operate at this lowest level of classification. It should be noted that if a compound is not classified into this specified class, assignment as a noncarcinogen does not immediately follow.

Level II. If two or more classes of carcinogens can be identified and defined, classification of an unknown into one of the defined classes, with the possibility that the unknown may be a member of none of these classes, constitutes a level II classification. A compound being a nonmember in all classes is considered an outlier.

Level III. At level III, information beyond classification is extracted from the analysis. Once an unknown is classified, its carcinogenic potency can be estimated, or a prediction of its primary location of tumor induction can be obtained. In some cases, for example, prediction of a level of activity, a quantitative relationship between structure and activity is established.

Level IV. This level of classification requires that two separate analyses be carried out. In the data base for a class of compounds undergoing analysis (1) data describing the chemical profile of the compounds and (2) data describing the pharmacological profile of a compound are analyzed. Attempts are then made to quantitatively relate the two profiles. This may require simple regression methods or possibly canonical correlation methods, as the complexity of the analysis dictates. As will be seen, SIMCA can operate at all levels of classification as discussed. A number of classification results at level III have been reported, but to our knowledge no level IV examples have been reported in the literature.

† 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.

0095-2338/81/1621-0008$01.25/0

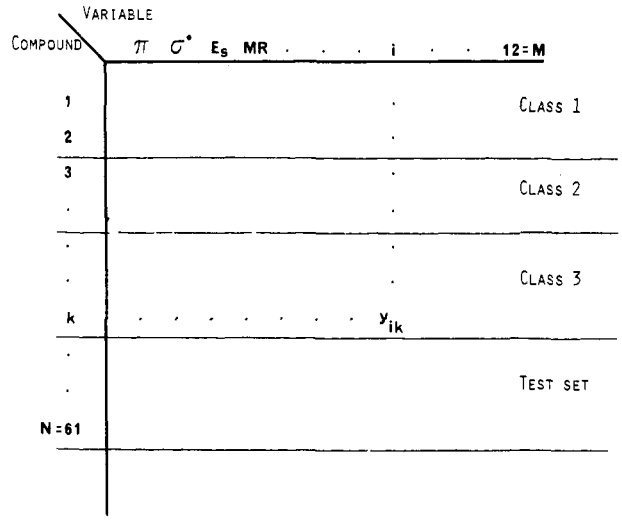


Figure 1. Data matrix for a three-class classification problem.

III. BASICS OF SIMCA PATTERN RECOGNITION

The objective of this study is to develop mathematical classification rules for *N*-nitroso compounds based on the structure–physical property relationships of such compounds of known carcinogenic potential. These rules are then used to predict the pharmacological response expected from similar untested compounds. The *N*-nitroso compounds of known carcinogenicity are called training or reference sets, while those compounds with unknown pharmacological properties are called the test set. The mathematical rules are referred to as similarity models. They are, in this study, derived from physicochemically based descriptors. These descriptors will henceforth be referred to as the "data".

In some cases simple classification rules can be derived based on, for example, the presence or absence of certain structural features. Therefore, in order for the results of such a classification study to be nontrivial, results beyond classification must be obtained. It is known from the work of Hansch[14] that within a class of structurally and pharmacologically similar substances, levels of activity can be a function of their physicochemical properties. Therefore, this approach to description offers a distinct advantage in that, in some cases, quantitative relationships between structure and activity can be derived and information at levels III and IV can be obtained.

For q reference sets of compounds described by M variables, a matrix such as that shown in Figure 1 is obtained. If the data in this matrix are represented in an M-dimensional space, each compound is defined as a single point, and, ideally, q clusters of points will result. This is shown in Figure 2 in three dimensions for $q = 2$. With SIMCA the mathematical regularity within the data for each class is described by eq 1, a principal component model. In this model

$$Y_{ik} = m_i + \sum_{\alpha=1}^{A} b_{ia} u_{ak} + e_{ik} \qquad (1)$$

Y is the observed value of variable i for compound k. The product terms $\alpha = 1, 2, 3 \ldots A$ are the component terms in the model, and associated with each component term is a variable specific term, b_i, and an object specific term, u_k. The difference in the observed Y and its model-predicted value is the residual, e_{ik}.

From the residuals, e_{ik}'s, for the objects of the reference sets, a residual standard deviation can be calculated for each class. Using this as a basis for a confidence interval, one encloses each class in a closed mathematical structure as shown in Figure 2. By fitting the data of unclassified compounds to these similarity models, an object can be classified as a member of one of the defined classes or as a member of neither.

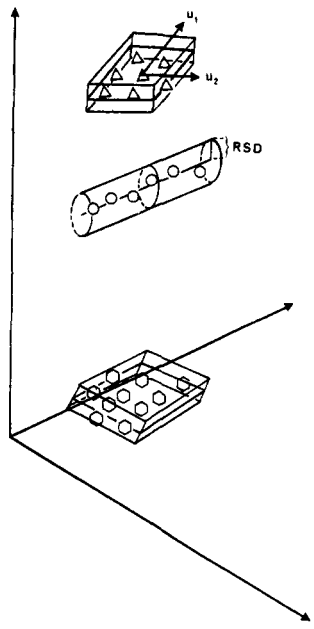


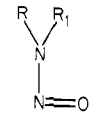
Figure 2. Geometric representation of a three-class problem.

The theoretical basis for the use of principal component models as measures of similarity has been published.[15] Two assumptions in this approach are (1) that the data for the objects are continuous and (2) that the class members are similar within each class.

Since our descriptors of the objects in this study are Hammett-type substituent constants and geometric measures of steric bulk of the substituents, the first assumption is easily satisfied. The second assumption, that the objects in each class are similar, implies two kinds of similarity, since this is a structure–biological activity study. The first type of similarity we call pharmacological similarity, which means that the objects in each class must elicit their response by a common mechanism. Chemical similarity requires that the objects in each class in some way be described as similar. Of the two types of similarity, pharmacological similarity presupposes chemical similarity, but pharmacological similarity does not imply chemical similarity.

IV. DATA

The reference sets for this SIMCA pattern recognition study were obtained from publications of Druckrey and co-workers,[16] Andrews et al.[17] and Kameswar Rao et al.[18] The carcinogen data of Druckrey are given in Table I and consist of the results of the evaluation of 45 compounds. These compounds are nitrosamines, *N*-nitrosoureas, and *N*-nitrosourethans. The data of Andrews et al. and Kameswar Rao et al. are also given in Table I and concern 28 nitrosamines. All compounds can be represented by the general structure



In both studies cited above, the nitroso compounds were given to rats as daily oral doses in drinking water. In a few cases where low water solubility dictated, the compounds were administered by gavage as solutions in oil. In addition to the evaluation of the carcinogenic potential for each compound, the location of tumor induction was determined and given for most of the carcinogenic compounds studied.

In the structural description of the compounds, each of the substituents R and R_1 is described by six constants, the Rekker

Table I. Biological data for *N*-nitroso compounds

Compound	Compound no.	Carcinogenicity: Druckrey et al. (1967)	Carcinogenicity: Andrews et al. (1978) and Kameswar Rao (1979)	Primary location of tumor induction	Class
N-Nitrosodiethylamine	01	+	+	Liver, esophagus, nasal turbinates	1
N-Nitrosodi-*N*-propylamine	02	+	+	Liver, esophagus, nasal turbinates	1
N-Nitrosodi-isopropylamine	03	+	+	Liver, nasal turbinates	1
N-Nitrosodi-*N*-butylamine	04	+	+	Liver, esophagus, bladder, nasal turbinates	1
N-Nitrosodi-isobutylamine	05		+		1
N-Nitrosodi-sec-butylamine	06		?		0
N-Nitrosodi-*N*-octylamine	07		-		0
N-Nitrosodiallylamine	08	-	-		0
N-Nitroso-bis(2-cyanoethyl)amine	09		?		0
N-Nitroso-bis(2-chloroethyl)amine	10		+	Liver, esophagus, forestomach	3
N-Nitroso-bis(2-chloropropyl)amine	11		?		0
N-Nitroso-bis(2-hydroxyethyl)amine	12	+	+	Liver	3
N-Nitroso-bis(2-hydroxypropyl)amine	13		+	Liver, nasal turbinates	3
N-Nitroso-bis(2-methoxymethyl)amine	14		+	Liver, esophagus, nasal turbinates	3
N-Nitroso-bis(2-ethyxyethyl)amine	15		+	Liver, esophagus, nasal turbinates	3
N-Nitroso-bis(2-oxopropyl)amine	16		+	Liver, esophagus, nasal turbinates	3
N-Nitroso-bis(2-trifluoromethylethyl)amine	17		-		0
N-Nitrosoethylmethylamine	18	+	+	Liver, esophagus, sarcoma	1
N-Nitrosomethylphenethylamine	19	+	+	Esophagus	1
N-Nitrosomethyl-*N*-propylamine	20		+		1
N-Nitrosomethylisopropylamine	21		+		1
N-Nitrosomethylneopentylamine	22		+	Esophagus	1
N-Nitrosoethyl-(2-hydroxyethyl)amine	23	+	+	Liver, esophagus	3
N-Nitrosodimethylamine	24		+	Liver	1
N-Nitrosomethylundecylamine	25		+	Liver, lung	1
N-Nitrosomethyldocecylamine	26		+	Bladder	1
N-Nitrosomethylcyclohexylamine	27	+	+	Esophagus	1
N-Nitrosomethylphenylamine	28	+	+	Esophagus	1
N-Nitrosomethylisopropylamine	29	+		Liver, esophagus, forestomach	1
N-Nitrosodi-*N*-pentylamine	30	+		Liver, lung	1
N-Nitrosodicyclohexylamine	31	-			0
N-Nitrosodiphenylamine	32	-			0
N-Nitrosodibenzylamine	33	-			0
N-Nitrosomethylvinylamine	34	+		Tongue, pharynx, esophagus	1
N-Nitrosomethylallylamine	35	+		Esophagus, nose	1
N-Nitrosomethylpentylamine	36	+		Esophagus	1
N-Nitrosomethyl-*N*-heptylamine	37	+		Lung	1
N-Nitrosomethylbenzylamine	38	+		Esophagus	1
N-Nitrosoethylvinylamine	39	+		Esophagus, forestomach	1
N-Nitrosoethyl-*N*-butylamine	40	+		Liver, esophagus	1
N-Nitrosomethyl-*tert*-butylamine	41	-			0
N-Nitroso-*n*-butyl-*n*-pentylamine	42	+		Liver	1
N-Nitroso-bis(2-hydroxyethyl)amine, diethyl ester	43	+		Liver	3
N-Nitrosobutyl-4-hydroxybutylamine	44	+		Bladder	1
N-Nitrosomethyl-2-chloroethylamine	45	+		Liver	3
N-Nitrosomethylcyanomethylamine	46	+		Liver	3
N-Nitroso-bis(cyanomethyl)amine	47	+		Liver, nasal turbinates	3
N-Nitrososarcosine	48	+		Esophagus	3
N-Nitrososarcosine, ethyl ester	49	+		Esophagus	3
N-Nitrosomethyl-1,1-dimethyl-3-oxobutylamine	50	+		Liver	3
N-Nitrosomethyl-(4-formylphenyl)amine	51	-			0
N-Nitrosoethyl-(4-pyridylmethyl)amine	52	+		Esophagus	1
N-Nitroso-*N*-methylacetamide	53	+		Forestomach	2
Ethyl-*N*-nitroso-*N*-methylcarbamate	54	+		Forestomach	2
Ethyl-*N*-nitroso-*N*-ethylcarbamate	55	+		Forestomach, nose, lung	2
N-Nitroso-*N*-methylurea	56	+		Forestomach	2
N-Nitroso-*N*,*N*'-dimethylurea	57	+			2
N-Nitroso-*N*,*N*',*N*'-trimethylurea	58	+			2
N-Nitroso-*N*-ethylurea	59	+			2
N-Nitroso-*N*-*N*-butylurea	60	+		Sarcoma	2
	61	+		Liver, nose	2

lipophilicity constant,[19] Taft's σ^* and E_s,[20] MR,[21] and Verloop's[22] steric constants L and B_4. In a few cases it was necessary to estimate some of the descriptors for some of the substituents, and these cases are noted. In order to be consistent throughout the study, the substituent with the smaller L constant is described first. Thus, the shorter group is always described first.

The Rekker lipophilicity constant for a substituent is somewhat analogous to the Hansch π constant in that it is a measure of the substituent's relative affinity for nonpolar biophases. The Verloop constants for a substituent are derived by obtaining the lowest energy conformation for the substituent and then calculating the length of the substituent (L) and the perpendicular width (B) at its widest point. The metric unit for these parameters is angstroms, and they are estimates of properties (e.g., area, volume) which are functions of these two parameters. The physiochemical descriptors have been published elsewhere.[9]

V. SIMCA METHODOLOGY

Because the details of a SIMCA analysis procedure have been published,[15] only a brief summary is presented here.

(1) Represent each object (compound) as an M-dimensional data vector. Then define the classes pertinent to the problem and select a representative training set for each class. (2) Normalize the data to zero mean and unit variance for each variable over the whole data set. This gives each variable equal initial weight in the analysis.

(3) Fit a separate similarity model to each class. The dimensionality of the models (A) is determined by cross validation. The cross validation may indicate that some classes lack structure (i.e., $A = 0$).

(4) Delete irrelevant variables, i.e., those variables not participating in the class models and not differentiating between classes. Also, delete obvious outliers among objects in the training sets.

(5) Fit new principal components models to the possibly reduced class matrices.

(6) Calculate the residual standard deviation for each class. These form the basis for the confidence intervals around the class together with the distribution of the u values in each class.

(7) Classify the objects in the training set (i.e., fit all class models to all training set data vectors).

(8) Validate the classification of the training sets (step 7) by deleting parts of the training set and calculating new

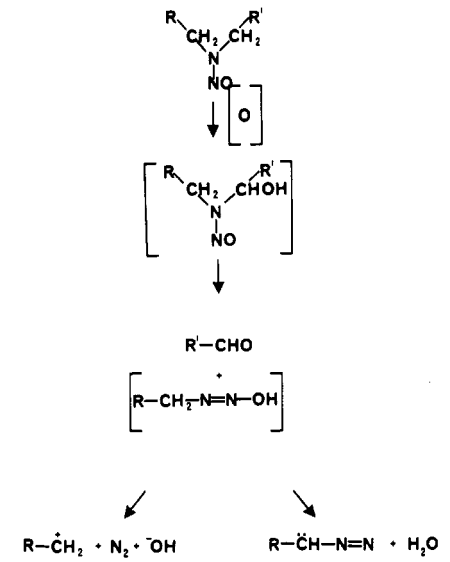


Figure 3. Metabolic activation of N-nitrosamines.

principal components class models from the reduced class matrices. The deleted objects are then fitted to the new principal components models to obtain their class assignment. This procedure is repeated until each object is deleted once and only once. The classification rate calculated from the assignment of the deleted objects gives a conservative estimate of the "correct" rate and also ensures stability in the data structure.

(9) Classify the objects in the test set by fitting each class model from step 7 to their data vectors.

(10) In postclassification analyses, searches for relations between the parameters u (the position of the objects in class) and other pharmacological properties can be made. Within each class, compounds of similar chemical and pharmacological properties may cluster. This can be detected by a graphical or regression analysis of the u's for each class.

VI. CHEMISTRY OF N-NITROSO COMPOUNDS

Prior to a presentation of the results of this classification study, a brief discussion of the chemistry of N-nitroso compounds is warranted.

The N-nitroso compounds in this study can be precursors to several reactive intermediates which may ultimately be responsible for their carcinogenicity.[16] Among these are (1) a carbocation intermediate or (2) a diazoalkyl intermediate.

N-Nitroso compounds can be activated in two different ways. These are shown in Figures 3 and 4. In Figure 3 the N-nitroso compound can be a substrate for metabolic α-hydroxylation, yielding an α-hydroxy intermediate which can decompose spontaneously to the appropriate aldehyde and the hydroxy form of the diazo cation. This can yield, if energetically favorable, an alkylating carbocation.

As indicated in Figure 4, if the N-nitroso compounds are N-nitrosoureas or -urethans, these intermediates are susceptible to hydrolysis. The result is the same hydroxyl form of the diazo cation, which yields the same alkylating carbocation. As the figures show, the same intermediate can be formed from chemically dissimilar compounds. The two chemically dissimilar carbocation precursors also represent pharmacologically dissimilar classes of potential carcinogens.

N-nitroso compounds can also yield highly reactive diazoalkyl intermediates; if carbocation formation is energetically unfavorable, water can be eliminated to yield the stabilized diazoalkyl intermediate. This pathway would be favored by R being a group which can stabilize the electron pair on the sp^3 carbon adjacent to the diazo group, by a field effect and/or

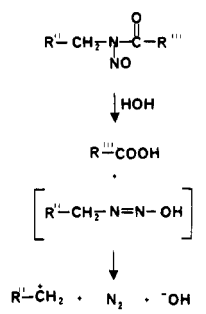


Figure 4. Hydrolytic activation of N-nitroso compounds.

a resonance effect. An example of such a group is R = CN or R = COCH. Therefore this scheme represents a third class of potentially carcinogenic N-nitroso compounds.

VII. CLASSIFICATION RESULTS

In the initial stages of the analysis of the data for the N-nitroso compounds in Table I, all of the active compounds were placed in the same reference set and the inactive and untested compounds in the test set. An attempt to derive a similarity model for the active compounds at this point failed. This class was then subdivided into the three classes given in Table I.

Since this subdivision is somewhat arbitrary, an explanation for it is as follows. Those compounds which required metabolic activation were separated from those which do not. The latter compounds, which include carbamates, ureas, etc., are the class 2 compounds in Table I. The former group was further subdivided into two classes corresponding to those N-nitrosoamines which may be expected to eventually yield a cation (class 1) and those which may be expected to yield a diazoalkane (class 3). Thus two classes of dialkylnitrosamines resulted, one group with electronically neutral and/or electron-donating substituents on the amine nitrogen (class 1) and one class of compounds which have at least one rather strongly electron-withdrawing substituent on the amine nitrogen (class 3). Any assumptions about routes of activation are not necessary at this point; other routes are equally conceivable. It is sufficient to say that this subdivision results in three classes of chemically dissimilar N-nitroso compounds. The training sets which resulted were as follows: class 1, which contained 27 compounds, class 2, which contained 9 compounds, and class 3, which contained 14 compounds.

The application of SIMCA to the three classes showed that there was considerable clustering. All 12 variables (see Table I) were significant for obtaining descriptions of the classes. The classification results are summarized as follows:

Class 1: A two-component similarity model resulted which classified 23 out of 27 correctly.

Class 2: A one-component similarity model resulted which classified 7 out of 9 correctly.

Class 3: A three-component similarity model resulted which classified all 14 compounds correctly.

The result of 44 out of 50, or 88%, of the active compounds correctly classified is highly significant. The compounds from class 1 which were incorrectly classified were N-nitrosodiisobutylamine (no. 5), N-nitrosomethylneopentylamine (no. 22), N-nitrosomethylphenylamine (no. 28), and N-nitrosoethyl(4-formylpyridyl)amine (no. 52). None are represented by the structure for this class. N-nitrosoethyl(4-formylpyridyl)amine is found to be in class 3. N-Nitroso-N-methylacetamide (no. 53) and ethyl N-nitroso-N-ethylcarbamate (no. 55) of the class 2 carcinogens are misclassified.

It might be noted that some of the carcinogens (e.g., N-nitrososarcosine) are equally well described by more than one of the similarity models. This indicates that there is some overlap between the classes, which might be anticipated in view

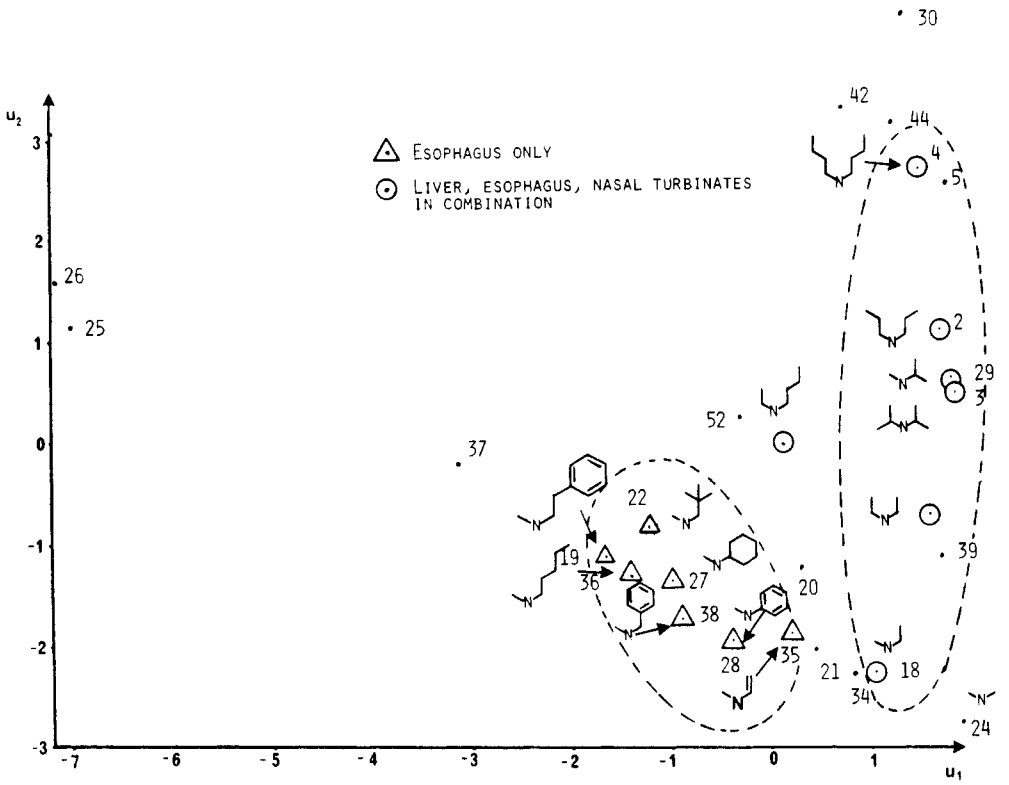


Figure 5. Clustering of pharmacologically similar class 1 *N*-nitrosamines.

of the nature of the subdivision of the training set.

Of the test set compounds, eight were reported to be inactive and three had, at the time this project was begun, incomplete test data. The completed test results indicate that the diisobutyl compound (no. 6) is inactive, the bis(2-chloropropyl) compound (no. 11) is a carcinogen whereas the bis(2-cyanoethyl) compounds (no. 9) are inactive.[23] SIMCA predicted these three as inactive, active, and active, respectively. The classification result of two out of three is encouraging. The diallyl compound is a false positive. The test set compound, *N*-nitrosodiphenylamine (no. 32), gives a very interesting classification result. During the preparation of this manuscript, Cardy et al.[24] reported that this substance, when tested at higher dosages than those tested by Druckrey,[16] proved to be carcinogenic. This result is not inconsistent with the SIMCA classification result, because being a member of none of the described classes does not preclude membership in an undescribed (undiscovered) class of carcinogens, which *N*-nitrosodiphenylamine apparently is.[24] All of the inactives, with the exception of *N*-nitrodiallylamine (no. 8), are classified as members of none of the three classes of carcinogens.

VIII. INTERPRETATION OF RESULTS

The evaluation of chemicals present in the human environment for their potential to produce harmful effects on the population is one of the most complex problems that face us today. Because of the nature of the problem, only an estimate of the harmful effects that a new or untested compound may have can be obtained. Therefore, the solution of the problem is a probabilistic one, and any such estimate, whether it be experimentally or empirically derived, must be in terms of a level of statistical significance.

Semiempirical classification results, such as those reported here, are only as reliable as the data and assumptions on which they are based. In this report two types of data are considered: (1) the physicochemical variables on which the calculations are made and (2) the biological assay results. The physicochemical variables are more quantitative, compared to the biological measurements, and we assume the limitations in the analysis to be the biological assessments.

Results based on the analysis of animal cancer tests are particularly interesting and controversial. Especially interesting is the question of whether there is a safe dose of exposure to a carcinogen. It is assumed in this analysis that compounds were tested at a concentration level that would produce a carcinogenic response, if carcinogenic. If these conditions are not met in animal testing, we agree with Ames[1] that the term "noncarcinogenic" becomes quantitatively meaningless.

Consequently, the interpretation of this SIMCA analysis becomes a probabilistic view of the carcinogenicity–noncarcinogenicity problem. In Figure 2 the carcinogens form well-defined clusters in descriptor space, and such clustering can be well approximated by similarity models. Within the first standard deviation of the model, compounds with the highest probability of being carcinogens are found, within the next standard deviation those compounds with a lower probability of being carcinogens are found, and at the extremes of the region of a class a definite classification as a carinogen becomes "fuzzy". A compound outside the carcinogen class means that it is not the same type of carcinogen as those inside the class or that it could be a "noncarcinogen". If this view of the problem of assessment of compounds as carcinogens or noncarcinogens is accepted, their classification as active or inactive becomes an oversimplification.

IX. POSTCLASSIFICATION ANALYSIS

In a postclassification analysis the objective is to attempt to relate some secondary property of the compounds in the training sets with their geometric position in the class structure. In most structure–activity studies, such as those resulting from the application of the method of Hansch, the secondary property is the level of activity of the compounds in a class of active compounds. In the case in which the compounds are carcinogenic, such a relationship between structure and level of carcinogenicity is not always possible due to the difficulties in quantitating this parameter.

For most of the compounds in the class 1 reference set, the primary location of tumor induction was reported from the results of an autopsy on each animal. If the parameters u_1 and u_2 from the SIMCA analysis for the class 1 *N*-nitroso compounds are plotted (Figure 5), it can be seen that definite clustering into one area takes place with compounds which induce tumors mainly in the esophagus, liver, and nasal turbinates. There is also a cluster of compounds which induce tumors only in the esophagus. If such phenomena are assumed to be a function of the physicochemical properties of the carcinogens, this result can be expected.

Those compounds inducing tumors only in the esophagus are found in the area described by $-2 < u_1 < 0$ and $-2 < u_2 < 0$ in Figure 5. If $u_1 = -1$ and $u_2 = -1$, for example, and class 1 m_i and b_i values are used for 12 variables, Y_{ik} values can be calculated for the compound with this physicochemical description. The nonscaled parameters calculated for these compounds are $f_R = 1.36$, $\sigma^*_R = -0.01$, $MR_R = 6.27$, $Es_R = 1.26$, $L_R = 3.11$, $B_{4R} = 2.13$, $f_{R_1} = 2.60$, $Es_{R_1} = 1.63$, $L_{R_1} = 6.86$, $B_{4R_1} = 4.55$, $\sigma^*_{R_1} = 0.01$, and $MR_{R_1} = 7.95$. The property which appears to distinguish this cluster of compounds is σ^*, with the esophagal carcinogens characterized by R and R_1 being electronically neutral or electron withdrawing (as modeled by σ^*).

X. DISCUSSION

Other attempts have been made to derive quantitative structure–activity relationships for carcinogenic *N*-nitroso compounds[25,26] by using regression methods. Recently a classification study dealing with such agents was reported.[27,28] This study employed the linear learning machine as the classifier of *N*-nitroso compounds as carcinogenic or noncarcinogenic. We doubt the validity of such a classification, since neither carcinogens nor noncarcinogens represent homogeneous classes of pharmacologically and chemically similar compounds.

The results here with the SIMCA method, however, show that significant classification results consistent with, and explicable in terms of, the chemistry of these substances can be obtained. The success is probably due to the fact that classification is based on a description of the chemical similarity of the carcinogenic *N*-nitroso compounds. This led to division of the compounds into three classes. This division, while somewhat arbitrary, is consistent with the current ideas about the mechanism of carcinogenicity of *N*-nitroso compounds.[16,18]

In this study, in which descriptions of only carcinogens are obtained, classification of an active compound as not being a member of any well-described class cannot be considered an incorrect result. Such an object may be a member of an as yet undiscovered class of carcinogens, such as *N*-nitrosodiphenylamine. A false positive result, however, must be considered an incorrect result.

N-nitrosodiallylamine is consistently classified as a false positive. This indicates that there is some chemical property not included in its description which results in its being a poor substrate for activation and therefore being inactive. Such outliers as these are a natural and valuable result of the analysis and should be considered in more detail.

Some general comments on the use of pattern recognition in the manner proposed and illustrated in this report are in order. The interaction of carcinogens with biological systems is very complex, not only in terms of the physicochemical nature of these interactions but also in terms of the number of different kinds of significant biological events which can occur. Even for carcinogens of the general structure given in section IV, it cannot be expected that they will all be carcinogens by a common mechanism.

The use of pattern recognition to estimate the potential of an unknown or untested compound to induce cancer in test animals, in the opinion of present investigators, shows promise.

REFERENCES AND NOTES

(1) Ames, B. N. *Science (Washington D.C.)* **1979**, *204*, 587.
(2) Ting, K. H.; Lee, R. C. T.; Milne, G. W. A.; Shapiro, H.; Gaurino, A. M. *Science (Washington, D.C.)* **1973**, *180*, 417.
(3) Kowalski, B. R.; Bender, C. F. *J. Am. Chem. Soc.* **1974**, *96*, 916.
(4) Stuper, A. J.; Jurs, P. C. *J. Am. Chem. Soc.* **1975**, *97*, 182.
(5) Chu, K. C.; Feldman, R. J.; Shapiro, M. B.; Hazard, G. F.; Geran, R. I. *J. Med. Chem.* **1975**, *18*, 539.
(6) Cammarata, A.; Menon, G. K. *J. Med. Chem.* **1976**, *19*, 739.
(7) Jurs, P. C.; Chou, J. T.; Yuan, M. In "Computer Assisted Drug Design", *ACS Symp. Ser.* **1979**, *No. 112*, Chapter 4.
(8) Wold, S.; Sjostrom, M. In "Chemometrics, Theory and Application", *ACS Symp. Ser.* **1977**, *No. 52*.
(9) Dunn, W. J.; Wold, S. *Bioorg. Chem.*, in press.
(10) Dunn, W. J.; Wold, S. *J. Med. Chem.* **1979**, *21*, 1001.
(11) Norden, B.; Edlund, U.; Wold, S. *Acta Chem. Scand., Ser. B* **1979**, *B32*, 1.
(12) Dunn, W. J.; Wold, S. *Bioorg. Chem.*, in press.
(13) Albano, C.; Dunn, W. J.; Edlund, U.; Johannson, E.; Norden, B.; Sjostrom, M.; Wold, S. *Anal. Chem. Acta* **1978**, *103*, 429.
(14) Hansch, C. *Acc. Chem. Res.* **1967**, *2*, 232.
(15) Wold, S. *Pattern Recognition* **1976**, *8*, 127.
(16) Druckrey, H.; Preussmann, R.; Ivankovic, S.; Schmall, D. Z. *Krebsforsch.* **1967**, *69*, 103.
(17) Andrews, A. W.; Libault, L. H.; Lijinsky, W. *Mutat. Res.* **1978**, *51*, 319.
(18) Kameswar Rao, T.; Young, J. A.; Lijinsky, W.; Epler, J. L. *Mutat. Res.* **1979**, *66*, 1.
(19) Rekker, R. "The Hydrophobic Fragment Constant"; Elsevier: Amsterdam, 1977.
(20) Taft, R. W. In "Steric Effects in Organic Chemistry"; Newman, M. S., Ed.; Wiley: New York, 1956.
(21) Dunn, W. J. *Eur. J. Med. Chem.* **1977**, *12*, 109.
(22) Verloop, A.; Hoogenstraaten, W.; Tipker, J. In "Drug Design"; Ariens, E. J., Ed.; Academic Press: New York, 1977; Vol. 7.
(23) Lijinsky, W., personal communication.
(24) Cardy, R. H.; Lijinsky, W.; Hildebrandt, P. K. *Exotoxicol. Environ. Safety* **1979**, *3*, 29.
(25) Wishnok, J. S.; Archer, M. C. *Br. J. Cancer* **1976**, *33*, 307.
(26) Wishnok, J. S.; Archer, M. C.; Adelman, a. S.; Rank, W. M. *Chem.-Biol. Interact.* **1978**, *20*, 42.
(27) Jurs, P. C.; Chou, J. T.; Yuan, M. J. *J. Med. Chem.* **1979**, *22*, 476.
(28) Jurs, P. C.; Chou, J. T. *J. Med. Chem.* **1979**, *22*, 792.