BIOORGANIC CHEMISTRY 10, 29–45 (1981)

The Carcinogenicity of *N*-Nitroso Compounds: A SIMCA Pattern Recognition Study

W. J. DUNN III* AND SVANTE WOLD†

Department of Medicinal Chemistry College of Pharmacy, University of Illinois at the Medical Center, 833 S. Wood, Chicago, Illinois 60612, and †Research Group for Chemometrics, The Institute of Chemistry, Umea University, S 901 87, Umea, Sweden

Received October 11, 1979

A number of methods are available for evaluating the potential of chemical agents to induce cancer in test animals. Recent use of SIMCA pattern recognition in structure–biological activity studies suggests that this method may be useful in making such evaluations. In this report this method is used to estimate the potential of a number of *N*-nitroso compounds to induce tumors in the rat. The data treatment and estimation results are consistent with the chemistry of these substances.

INTRODUCTION

N-Nitroso compounds, such as nitrosoamines, *N*-nitroso ureas, and *N*-nitroso urethans, are organic compounds with the potential for being carcinogenic (1, 2). Certain of these compounds are environmental carcinogens. Nitrosamines, for example, have been shown to form in the environment from a number of naturally occurring amines such as amino acids or from commercially available amines and nitrite ion (*3*). Exposure to these thus-formed nitrosamines could technically result in tumor induction in humans. It is therefore of interest to increase our understanding of the effects of these and other *N*-nitroso compounds when they interact with living systems.

In attempts to assess the carcinogenic potential (CP) of these agents, a number of laboratories have synthesized and tested a large number of these compounds in experimental animals. Compilations of such data have been published (*1, 4, 5*) and are a valuable source of data for systematic evaluation of the CP for these agents.

Until recently, such evaluations were based on the comparison of the chemical structures of unknown or untested *N*-nitroso compounds with those of known CP (confirmed carcinogens or noncarcinogens). The result is an intuitive estimate of the CP for the unknown or untested compounds. Classification so based leaves much to be desired, and more reliable and quantitative estimates of CP can be made with mathematical methods of data analysis, so-called methods of pattern recognition (PaRC). PaRC methods have recently been applied to the problem of predicting the carcinogenicity of organic compounds. Jurs and co-workers (*5*) have used the linear learning machine to assess the carcinogenicity of a large

0045-2068/81/010029-17$02.00/0

number of structurally and pharmacologically diverse carcinogens and noncarcinogens. Using a method of PaRC based on analogy and similarity, the SIMCA method, recent evaluations of CP of 4-nitroquinoline-1-oxides (*6a*) and polycyclic aromatic hydrocarbons (*6b*) have resulted from these laboratories.

The SIMCA method of PaRC offers a number of advantages in classification studies based on pharmacological structure–activity data and these have been discussed (*7*). In this report we apply this method of PaRC to the problem of assessing the CP of *N*-nitroso compounds.

THE 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 can then be used to predict the pharmacological response expected from similar untested compounds. The *N*-nitroso compounds of known carcinogenicity are known as the *training* or *reference* sets while those compounds with unknown pharmacological properties are known as the *test* set. The mathematical rules we will refer to as similarity models. They are, in this study, derived from physicochemically based descriptors. These descriptors will henceforth be referred to as the *data*. The outcome of this approach is that classification as a carcinogen or noncarcinogen will be based on the physicochemical description of the subject compounds.

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 and his co–workers (*8*) that within a class of structurally and pharmacologically similar substances, levels of activity can be a function of these 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 beyond classification can be obtained from the analysis (*6, 7*).

For q reference sets of *N*-nitroso compounds which can be described by M variables, a matrix such as that shown in Fig. 1 can be obtained. If the data from 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 Fig. 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 components model. In this model

$$y_{ik}{}^{q} = m_i{}^{q} + \sum_{a=1}^{A} b_{ia}{}^{q} u_{ak}{}^{q} + e_{ik}. \qquad (1)$$

y is the observed value of variable i for the kth object. m_i is the mean value of variable i. The product terms, $A = 1, 2, 3 \ldots$ are the component terms in the

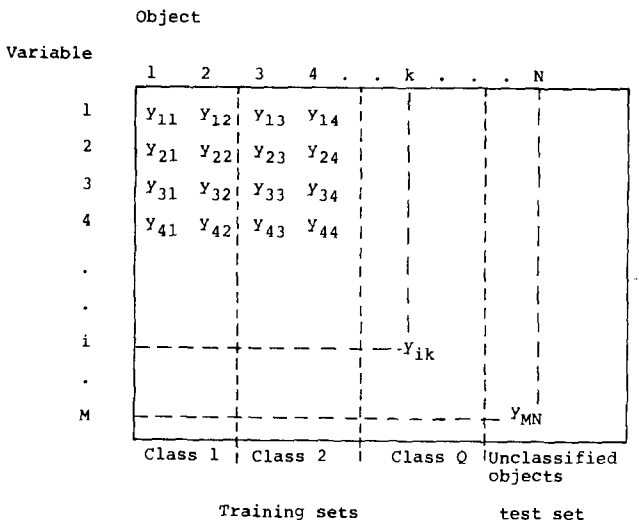


FIG. 1. Data matrix for a q-class classification problem.

model and associated with each component 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

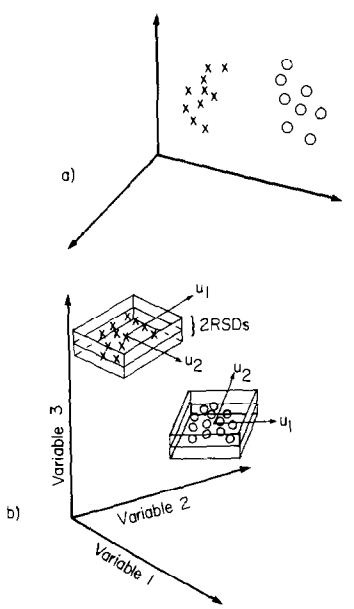
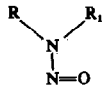


FIG. 2. (a) Data structure for two classes of compounds. (b) A SIMCA description of the similarity of the two classes.

TABLE 1

BIOLOGICAL AND PHYSICOCHEMICAL[a] DATA FOR *N*-NITROSO COMPOUNDS



Compound	Carcinogenicity Ref. (1)	Carcinogenicity Ref. (4)	Primary location of tumor induction[b]	Class	f_R	σ^*_R	MR_R	E_{sR}	L_R	B_{4R}	f_{R1}	σ^*_{R1}	MR_{R1}	E_{sR1}	L_{R1}	B_{4R1}	
01	+	+	Liver, esophagus, nasal turbinates	1	1.22	−0.10	10.30	−1.31	4.11	2.97	1.22	−0.10	10.30[b]	−1.31	4.11	2.97	*N*-Nitrosodiethylamine
02	+	+	Liver, esophagus, nasal turbinates	1	1.74	−0.12	14.96	−1.62	5.05	3.49	1.74	−0.12	14.96	−1.60	5.05	3.50	*N*-Nitrosodi-*n*-propylamine
03	+	+	Liver, nasal turbinates	1	1.74	−0.19	14.96	−1.71	4.11	3.16	1.74	−0.19	14.96	−1.71	4.11	3.16	*N*-Nitrosodi-isopropylamine
04	+	+	Liver, esophagus, bladder, nasal turbinates	1	2.26	−0.13	19.59	−1.63	6.17	4.42	2.26	−0.13	19.59	−1.63	6.17	4.42	*N*-Nitrosodi-*n*-butylamine
05		+	—	1	2.26	−0.13	19.59	−2.17	5.05	4.21	2.26	−0.13	19.59	−2.17	5.05	4.21	*N*-Nitrosodi-isobutylamine
06		?	—	0	2.26	−0.21	19.59	−2.37	5.05	3.49	2.26	−0.21	19.59	−2.37	5.05	3.49	*N*-Nitrosodi-*sec*-butylamine
07		−	—	0	4.34	−0.15	38.20	−1.57	10.27	6.85	4.34	−0.15	38.20	−1.57	10.27	6.85	*N*-Nitrosodi-*n*-octylamine
08	−	−	—	0	1.38	0.23	14.49	−1.60	5.11	3.78	1.38	0.23	14.49	−1.60	5.11	3.78	*N*-Nitrosodiallylamine
09		?	—	0	0.00	0.80	14.70	−2.23	6.28	3.05	0.00	0.80	14.70	−2.23	6.28	3.05	*N*-Nitroso-bis(2-cyanoethyl)amine
10		+	Liver, esophagus, forestomach	3	1.10	0.39	15.15	−2.14	5.57	3.25	1.10	0.39	15.15	−2.14	5.57	3.25	*N*-Nitroso-bis(2-chloroethyl)amine
11		?	—	0	1.61	0.35	19.80	−2.14	5.00	3.46	1.61	0.35	19.80	−2.14	5.00	3.46	*N*-Nitroso-bis(2-chloropropyl)amine
12	+	+	Liver	3	−0.43	0.21	11.84	−1.32	4.79	3.38	−0.42	0.21	11.84	−1.32	4.79	3.38	*N*-Nitroso-bis(2-hydroxyethyl)amine
13		+	Liver, nasal turbinates	3	0.09	0.16	16.44	−2.31	5.00	3.09	0.09	0.16	16.44	−2.31	5.00	3.09	*N*-Nitroso-bis(2-hydroxypropyl)amine
14		+	Liver, esophagus, nasal turbinates	3	0.14	0.24	16.68	−2.01	6.02	3.86	0.14	0.24	16.68	−2.01	6.02	3.81	*N*-Nitroso-bis(2-methoxymethyl)amine
15		+	Liver, esophagus, nasal turbinates	3	0.66	0.27	21.30	−2.21	7.13	4.74	0.66	0.27	21.30	−2.21	7.13	4.74	*N*-Nitroso-bis(2-ethyxyethyl)amine
16		+	Liver, esophagus, nasal turbinates	3	−0.42	0.60	15.06	−1.99	4.54	4.39	−0.42	0.60	15.06	−1.99	4.54	4.39	*N*-Nitroso-bis(2-oxopropyl)amine
17		−	—	0	1.79	0.32	14.32	−1.61	5.52	4.55	1.79	0.32	14.32	−1.61	5.52	4.55	*N*-Nitroso-bis(2-trifluoro-methylethyl)amine
18	+	+	Liver, esophagus, sarcoma	1	0.70	0.00	5.65	−1.24	3.00	2.04	1.22	−0.10	1 .30	−1.31	4.11	2.97	*N*-Nitrosoethylmethylamine
19	+	+	Esophagus	1	0.70	0.00	5.65	−1.24	3.00	2.04	2.88	0.08	34.60	−1.62	8.33	3.16	*N*-Nitrosomethylphenethylamine
20		−	—	1	0.70	0.00	5.65	−1.24	3.00	2.04	1.74	−0.12	14.96	−1.60	5.05	3.49	*N*-Nitrosomethyl-*n*-propylamine
21		+	—	1	0.70	0.00	5.65	−1.24	3.00	2.04	1.74	−0.19	14.96	−1.71	4.11	3.16	*N*-Nitrosomethylisopropylamine
22		+	Esophagus	1	0.70	0.00	5.65	−1.24	3.00	2.04	2.78	−0.17	24.24	−2.98	5.22	5.39	*N*-Nitrosomethylneopentylamine
23	+	+	Liver, esophagus	3	1.22	−0.10	10.30	−1.31	4.11	2.97	−0.43	0.21	11.84	−1.32	4.79	3.38	*N*-Nitrosoethyl-(2-hydroxyethyl)amine

24		+	Liver	1	0.70	0.00	5.65	−1.24	3.00	3.04	0.70	0.00	5.65	−1.24	3.00	2.04	*N*-Nitrosodimethylamine
25		+	Liver, lung	1	0.70	0.00	5.65	−1.24	3.00	2.04	5.90	−0.15	52.20	−1.54	13.27	9.30	*N*-Nitrosomethylundecylamine
26		+	Bladder	1	0.70	0.00	5.65	−1.24	3.00	2.04	6.42	−0.15	56.80	−1.54	14.38	10.27	*N*-Nitrosomethyldocecylamine
27	+	+	Esophagus	1	0.70	0.00	5.65	−1.24	3.00	2.04	2.94	−0.26	26.69	−2.03	6.17	3.49	*N*-Nitrosomethylcyclohexylamine
28	+	+	Esophagus	1	0.70	0.00	5.65	−1.24	3.00	2.04	1.84	0.60	25.36	−1.01	6.28	3.11	*N*-Nitrosomethylphenylamine
29	+		Liver, esophagus, forestomach	1	1.74	−0.19	14.96	−1.71	4.11	3.16	1.74	−0.19	14.96	−1.71	4.11	3.16	*N*-Nitrosomethylisopropylamine
30	+		Liver, lung	1	2.78	−0.16	24.24	−1.64	7.11	4.94	2.78	−0.16	24.24	−1.64	7.11	4.94	*N*-Nitrosodi-*n*-pentylamine
31	−		—	0	2.94	−0.26	26.69	−2.03	6.17	3.49	2.94	−0.26	26.69	−2.03	6.17	3.49	*N*-Nitrosodicyclohexylamine
32	−		—	0	1.84	0.60	25.36	−1.01	6.28	3.11	1.84	0.60	25.36	−1.01	6.28	3.11	*N*-Nitrosodiphenylamine
33	−		—	0	2.36	0.26	30.01	−1.62	6.02	3.63	2.36	0.26	30.01	−1.62	6.02	3.63	*N*-Nitrosodibenzylamine
34	+		Tongue, pharynx, esophagus	1	0.70	0.00	5.65	−1.24	3.00	2.04	0.86	0.40	10.99	−1.39	4.29	3.09	*N*-Nitrosomethylvinylamine
35	+		Esophagus, nose	1	0.70	0.00	5.65	−1.24	3.00	2.04	1.38	0.23	14.49	−1.60	5.11	3.78	*N*-Nitrosomethylallylamine
36	+		Esophagus	1	0.70	0.00	5.65	−1.24	3.00	2.04	2.78	−0.16	24.24	−1.64	7.11	4.94	*N*-Nitrosomethylpentylamine
37	+		Lung	1	0.70	0.00	5.65	−1.24	3.00	2.04	3.82	−0.17	33.55	−2.03	9.16	6.39	*N*-Nitrosomethyl-*n*-heptylamine
38	+		Esophagus	1	0.70	0.00	5.65	−1.24	3.00	2.04	2.36	0.26	30.01	−1.62	6.02	3.63	*N*-Nitrosomethylbenzylamine
39	+		Esophagus, forestomach	1	1.22	−0.10	10.30	−1.31	4.11	2.97	0.86	0.40	10.99	−1.39	4.29	3.09	*N*-Nitrosoethylvinylamine
40	+		Liver, esophagus	1	1.22	−0.10	10.30	−1.31	4.11	2.97	2.26	−0.13	19.59	−1.63	6.17	4.42	*N*-Nitrosoethyl-*n*-butylamine
41	−		—	0	0.70	0.00	5.65	−1.24	3.00	2.04	2.26	−0.30	19.62	−2.78	4.11	2.97	*N*-Nitrosomethyl-*tert*-butylamine
42	+		Liver	1	2.26	−0.13	19.59	−1.63	6.17	4.42	2.78	−0.16	24.24	−1.64	7.11	4.94	*N*-Nitroso-*n*-butyl-*n*-pentylamine
43	+		Liver	3	0.50	0.21[c]	21.13	−1.32[c]	6.02	5.39	0.50	0.21	21.13	−1.32	6.02	5.39	*N*-Nitroso-bis(2-hydroxyethyl)amine, diethyl ester
44	+		Bladder	1	2.26	0.13	19.59	−1.63	6.17	4.42	0.61	0.05[d]	21.13	−1.63[e]	7.01	5.24	*N*-Nitrosobutyl-4-hydroxybutylamine
45	+		Liver	3	0.70	0.00	5.65	−1.24	3.00	2.04	1.10	0.39	15.15	−2.14	5.57	3.25	*N*-Nitrosomethyl-2-chloroethylamine
46	+		Liver	3	0.70	0.00	5.65	−1.24	3.00	2.04	−0.52	1.30	10.11	−2.38	3.99	4.12	*N*-Nitrosomethylcyanomethylamine
47	+		Liver, nasal turbinates	3	−0.52	1.30	10.11	−2.38	3.99	4.12	−0.52	1.30	10.11	−2.38	3.99	4.12	*N*-Nitroso-bis(cyanomethyl)amine
48	+		Esophagus	3	0.70	0.00	5.65	−1.24	3.00	2.04	−0.14	1.05	11.88	−1.99[f]	5.02	3.59	*N*-Nitrososarcosine
49	+		Esophagus	3	0.70	0.00	5.65	−1.24	3.00	2.04	0.49	0.29	21.13	−1.99[f]	7.09	5.21	*N*-Nitrososarcosine, ethyl ester
50	+		Liver	3	0.70	0.00	5.65	−1.24	3.00	2.04	1.13	0.82	30.41	−2.78[g]	6.02[h]	2.97[i]	*N*-Nitrosomethyl-1,1-dimethyl-3-oxobutylamine
51	−		—	0	0.70	0.00	5.65	−1.24	3.00	2.04	1.35	0.60	25.36	−1.01	6.28	3.11	*N*-Nitrosomethyl-(4-formylphenyl)amine[j]
52	+		Esophagus	1	1.22	−0.10	10.30	−1.31	4.11	2.97	1.04	0.26	30.01	−1.62	8.33	3.16	*N*-Nitrosoethyl-(4-pyridylmethyl)amine[k]
53	+		Forestomach	2	0.70	0.00	5.65	−1.24	3.00	2.04	−0.94	1.65	11.18	−1.99[f]	4.06	2.93	*N*-Nitroso-*N*-methylacetamide
54	+		Forestomach	2	0.70	0.00	5.65	−1.24	3.00	2.04	−0.03	0.64	17.47	−1.99[f]	5.96	4.29	Ethyl-*N*-nitroso-*N*-methylcarbamate
55	+		Forestomach, nose, lung	2	1.22	−0.10	10.30	−1.31	4.11	2.97	−0.03	0.64	17.47	−1.99[f]	5.96	4.29	Ethyl-*N*-nitroso-*N*-ethylcarbamate
56	+		Forestomach	2	0.70	0.00	5.65	−1.24	3.00	2.04	−1.70	0.60	9.81	−1.99[f]	4.06	3.07	*N*-Nitroso-*N*-methylurea
57	+		—	2	0.70	0.00	5.65	−1.24	3.00	2.04	−1.47	0.60	14.57	−1.99[f]	5.00	3.07	*N*-Nitroso-*N,N'*-dimethylurea
58	+		—	2	0.60	0.00	5.65	−1.24	3.00	2.04	−1.52	0.60	19.58	−1.99[f]	4.61	3.83	*N*-Nitroso-*N,N',N'*-trimethylurea
59	+		—	2	−1.70	0.60	9.81	−1.99	4.06	3.07	1.22	−0.10	10.30	−1.31	4.11	2.97	*N*-Nitroso-*N*-ethylurea
60	+		Sarcoma	2	−1.70	0.60	9.81	−1.99	4.06	3.07	2.26	−0.13	19.59	−1.63	6.17	4.42	*N*-Nitroso-*N*-*n*-butylurea
61	+		Liver, nose	2	0.70	0.00	5.65	−1.24	3.00	2.04	−0.68	0.62	15.55	−1.99[f]	3.53	2.80	

[a] Pomona College Medicinal Chemistry Data Bank. [b] See Ref. (*1*). [c] Assumed equal to σ^* for CH_2COCH_3. [d] σ^* for CH_2COCH_3 attenuated by 0.25/CH_2. [e] Assumed equal to E_s for t-C_4H_9. [f] Assumed equal to E_s for CH_2COCH_3. [g] $L = L$ for CH_2COCH_3 + 1.11 Å for addition of CH_2. [h] $B_4 = B_4$ for t-C_4H_9. [i] Assumed equivalent to C_6H_5. [j] Assumed equivalent to $CH_2C_6H_5$.

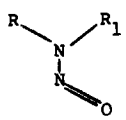
shown in Fig. 2. By fitting the data of unclassified compounds to these similarity models an object can be classified as a member of one of the defined classes or as a member of neither.

The theoretical basis for the use of principal components models as measures of similarity has been published (*9*). Two assumptions in this approach are that (i) the data for the objects be continuous and (ii) that the class members be similar.

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 be 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.

THE DATA

The reference sets for this SIMCA PaRC study were obtained from publications of Druckrey and co-workers (*1*) and from Lijinsky and co-workers (*4*). The carcinogen data of Druckrey are given in Table 1 and consist of the results of the evaluation of 45 compounds. These compounds are nitrosamines, *N*-nitroso ureas and *N*-nitroso urethans. The data of Lijinsky are also given in Table 1 and concern 28 nitrosamines. All compounds can be represented by the general structure (I) below.



I

19

In each of the studies cited above the nitroso compounds were given to rats as daily oral doses in drinking water. In a few cases where lack of water solubility dictated, the compounds were administered gavage as solutions in oil. In addition to the evaluation of the CP for each compound, the location of tumor induction was determined and given for most of the compounds studied. These data are also given in Table 1.

In the description of the compounds each of the substituents R and R_1 is described by six constants, the Rekker lipophilicity constant (*11*), Taft's σ^* and E_s (*12*), *MR* (*13*), and Verloop's steric constants (*14*) *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 analagous to the Hansch π constant in that it is a measure of the substituents' 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_4) at its widest point. These parameters are measured in angstroms, and these constants are estimates of properties (e.g., area, volume) which are functions of these two parameters.

THE SIMCA METHODOLOGY

Since the details of a SIMCA analysis have been published (*9*) 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., such 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 principal components class models from the reduced class matrices. The deleted objects are then fit 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 insures 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 (*7*) searches for relations between the param-

eters u_{ak} (the position of the objects in class) and other pharmacological properties can be made. Within each class described by a similarity model compounds of similar chemical and pharmacological properties may cluster. This can be detected by a graphical or regression analysis of the u_k's for each class.

THE 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 (*10*). Among these are (1) a carbocation intermediate or (2) a diazoalkyl intermediate.

The carbocation intermediate can be formed from *N*-nitroso compounds in two different ways. These are shown in Pathways 1 and 2 in Fig. 3. In Pathway 1 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 diazocation. This can yield, if *energetically favorable,* an alkylating carbocation.

As indicated in Pathway 2, if the *N*-nitroso compounds are *N*-nitroso ureas or urethans, these intermediates are susceptible to hydrolysis. The result is the same

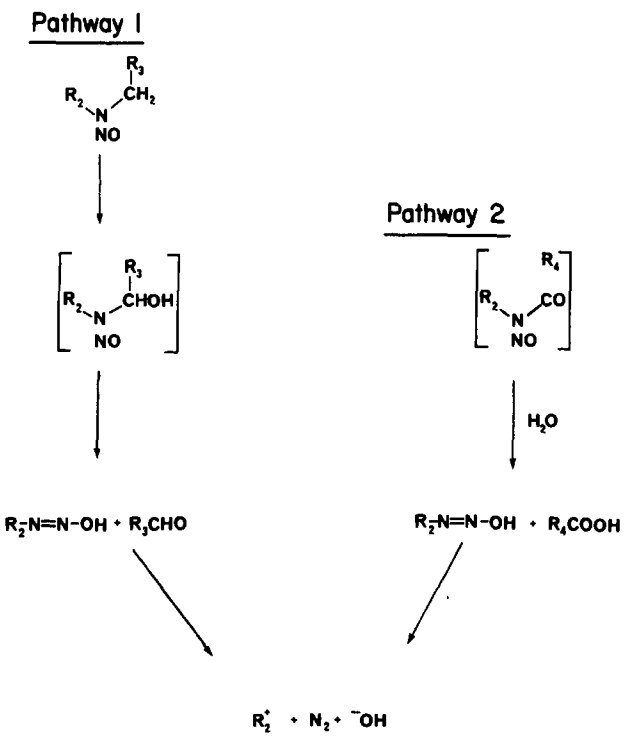


FIG. 3. Two pathways for formation of a carbocation.

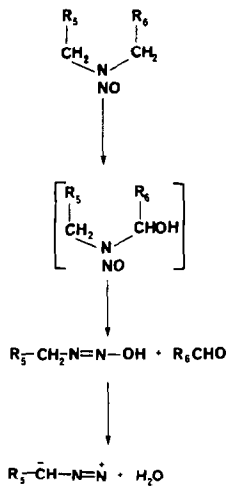


FIG. 4. Pathway for formation of diazoalkyl intermediates.

hydroxyl form of the diazocation which yields the same alkylating carbocation. As the Fig. 3 shows, 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 as shown in Fig. 4. α-Hydroxylation, as shown in Fig. 3, can lead to the hydroxyl form of the diazocation.

If carbocation formation is energetically unfavorable, water can be eliminated to yield the stabilized diazoalkyl intermediate. This pathway would be favored by R_5 being a group which can stabilize the electron pair of the sp^3 carbon adjacent to the diazo group, either by a field effect and/or a resonance effect. An example of such a group would be R_5=CN or R_5=$COCH_3$. Therefore this scheme represents a third class of potentially carcinogenic *N*-nitroso compounds.

CLASSIFICATION RESULTS

In the initial stages of the analysis of the data for the *N*-nitroso compounds in Table 1, 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 1.

Since this subdivision is somewhat arbitrary, an explanation for it is as follows. Those compounds which require metabolic activation were separated from those which do not. The latter compounds, which include carbamates, ureas, etc., are the class 2 compounds in Table 1. 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). The outcome of this subdivision was that 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 only 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 1) were significant for obtaining descriptions of the classes. The classification results are summarized below.

Class 1: A two-component similarity model resulted which classified 23/27 correctly

Class 2: A one-component similarity model resulted which classified 7/9 correctly

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

The result of 44/50 or 88% of the active compounds correctly classified is highly significant. The compounds from class 1 which were incorrectly classified were *N*-nitrosodiisobutylamine (**5**), *N*-nitrosomethylneopentylamine (**22**), *N*-nitrosomethylphenylamine (**28**), and *N*-nitrosoethyl-(4-formylpyridyl)amine (**52**). None are inside the structure for this class. *N*-Nitrosoethyl-(4-formylpyridyl)amine is found to be inside class 3. *N*-Nitroso-*N*-methylacetamide (**53**) and ethyl *N*-nitroso-*N*-ethylcarbamate (**55**) are missed of the class 2 carcinogens.

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 and this might be anticipated in view of the nature of the subdivision of the training set.

Of the test set compounds, eight are reported[1] to be inactive and three have, at this time, incomplete test data. All of the inactives, with the exception of *N*-nitrosodiallyamine (**8**), are classified to be members of none of the three classes of carcinogens. The diallyl compound is clearly a false positive. These results are given in Table 2.

INTERPRETATION OF RESULTS

The problem of the evaluation of chemicals present in the human environment for their potential to produce harmful effects on the population is one of the most

[1] During review of this work, it was reported by Lijinsky *et al.* (*18*) that *N*-nitrosodiphenylamine, when tested at significantly higher doses than by Druckrey *et al.* (*1*) proved to be carcinogenic. This result is not inconsistent with our classification since 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 (*18*).

TABLE 2

F STATISTICS[a] AND STANDARD DEVIATIONS (SD) FOR CLASSIFICATION OF TRAINING SETS

Compound	Class	u_1	u_2	u_3	SD[b]	*F* Class 1[c]	*F* Class 2[d]	*F* Class 3[e]
1	1	1.84	−0.69		0.30	0.52	5.00	1.80
2	1	1.74	1.12		0.11	0.07	7.70	1.50
3	1	1.89	0.60		0.33	0.64	7.30	2.50
4	1	1.51	2.98		0.16	0.14	15.00	2.00
5	1	1.75	2.83		0.65	2.40	15.00	3.00
18	1	1.13	−2.20		0.28	0.46	3.70	2.60
19	1	−1.48	−1.12		0.48	1.30	7.20	4.60
20	1	0.37	−1.18		0.17	0.17	3.00	2.00
21	1	0.67	−1.94		0.26	0.38	3.20	2.40
22	1	−1.14	−0.88		1.20	8.00	9.90	4.50
24	1	2.04	−2.65		0.27	0.43	4.70	3.60
25	1	−6.80	1.14		0.26	0.39	34.00	26.00
26	1	−6.98	1.57		0.30	0.53	43.00	33.00
27	1	−0.78	−1.27		0.47	1.30	5.70	3.20
28	1	−0.20	−1.84		0.69	2.80	5.20	3.40
29	1	1.89	0.60		0.33	0.64	7.30	2.50
30	1	1.35	4.51		0.28	0.45	25.00	3.50
34	1	1.11	−2.26		0.31	0.57	2.00	1.20
35	1	0.38	−1.86		0.21	0.25	1.80	1.00
36	1	−1.25	−1.06		0.16	0.15	5.70	3.50
37	1	−2.92	−0.23		0.31	0.55	12.00	7.40
38	1	−0.75	−1.44		0.32	0.58	4.00	2.50
39	1	1.82	−0.75		0.37	0.77	3.20	1.10
40	1	0.18	0.13		0.17	0.17	5.20	1.20
42	1	0.79	3.31		0.16	0.16	17.00	2.30
44	1	1.32	3.09		0.41	0.96	15.00	2.00
52	1	−1.17	0.13		0.59	2.00	5.60	1.70
53	2	2.04			0.64	8.60	2.50	2.10
54	2	0.86			0.35	2.50	0.78	0.20
55	2	0.89			0.62	2.60	2.40	0.47
56	2	1.22			0.32	3.20	0.63	1.00
57	2	1.15			0.22	3.10	0.30	0.64
58	2	1.14			0.27	3.60	0.47	0.70
59	2	−4.12			0.37	12.00	0.86	6.08
60	2	−4.24			0.31	12.00	0.58	6.50
61	2	1.06			0.28	2.30	0.50	0.65
10	3	1.43	0.14	1.12	0.49	5.90	6.50	1.10
12	3	−0.05	−1.30	−1.75	0.49	4.50	2.40	1.10
13	3	1.32	0.13	1.50	0.59	7.80	6.00	1.60
14	3	1.78	−0.72	0.65	0.23	5.90	5.60	0.24
15	3	3.13	−1.15	1.59	0.36	7.40	14.00	0.60
16	3	2.03	0.86	−0.85	0.28	12.00	4.80	0.36
23	3	−1.38	−1.90	−0.96	0.50	1.10	3.40	1.20
43	3	1.86	−2.84	−1.18	0.60	6.60	12.00	1.70
45	3	−2.53	0.01	0.27	0.33	1.60	1.30	0.53
46	3	−2.71	1.37	−0.29	0.56	8.10	1.50	1.50
47	3	2.90	4.26	−1.34	0.21	32.00	16.00	0.20
48	3	−2.63	0.59	−0.57	0.30	4.10	0.42	0.44
49	3	−2.39	−0.49	0.16	0.62	1.80	2.50	1.80
50	3	−2.75	1.05	1.65	0.57	7.80	4.80	1.60

[a] The *F* statistics for each compound are calculated based on its fit to the respective class model.
[b] Class 1, SD = 0.42; class 2, SD = 0.40; class 3, SD = 0.46.
[c] For assignment to class 1, $F(9,240)_{\alpha 0.05} < 1.92$ for training set: others $F(10,240)_{\alpha 0.05} < 1.87$.
[d] For assignment to class 2, $F(9,77)_{\alpha 0.05} < 1.99$: others $F(11,77)_{\alpha 0.05} < 1.97$.
[e] For assignment to class 3, $F(6,90)_{\alpha 0.05} < 2.21$: others $F(9,90)_{\alpha 0.05} < 1.99$.

complex that faces us today. Due to 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: (i) the physicochemical variables on which the calculation are made and (ii) the biological assay results. The physicochemical variables on which the claculations are made compared to the biological measurements are more quantitative, and the limitations in the analysis we assume 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, if carcinogenic, were tested at a concentration level that would produce a carcinogenic response. If these conditions are not met in animal testing according to Ames (*16*) the term "noncarcinogenic" becomes quantitatively meaningless.

Considering this, the interpretation of this SIMCA analysis becomes a realistic

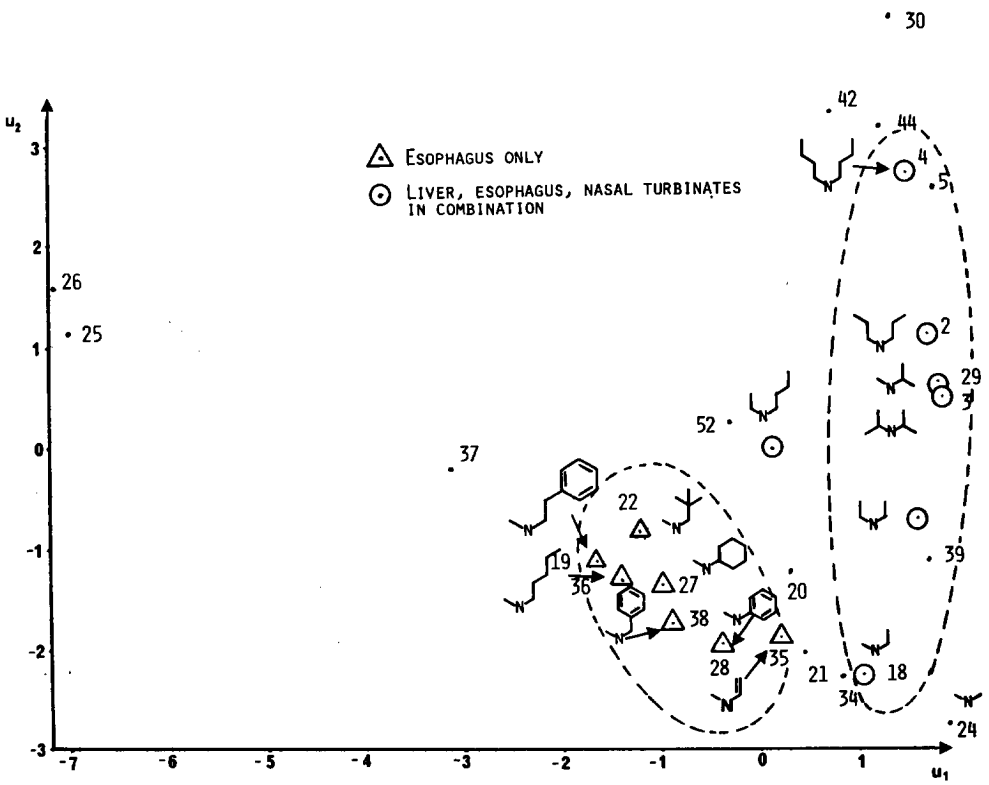


FIG. 5. u_1 plot for the class 1 carcinogenic *N*-nitroso compounds. Those compounds not indicated by △ or ⊙ induce tumors in other organs or do not have autopsy data.

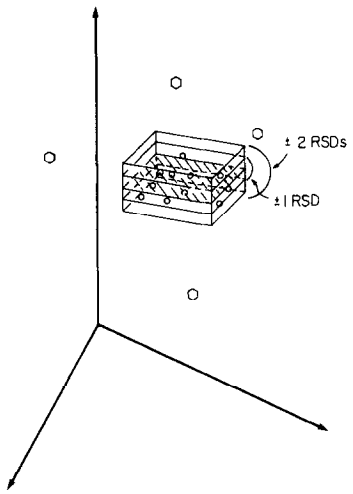


FIG. 6. SIMCA representation of members of a class of carcinogens (O) with nonmembers of that class (O).

view of the carcinogenicity–noncarcinogenicity problem. A graphical presentation is given in Fig. 6. In Fig. 6 the carcinogens form a well-defined cluster in descriptor space which can be well approximated by a similarity model. Within the first standard deviation of the hyperplane, 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 the class a definite classification as a carcinogen becomes "fuzzy." A compound outside of the carcinogen class means that it is not the same type of carcinogen as those inside of 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.

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 possible due to the inability to quantitate 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. By plotting (Fig. 5) the parameters, u_1 and u_2 from the SIMCA analysis for the class 1

TABLE 3

TEST SET CLASSIFICATION RESULTS

Compound	Class	SD (class)[a]			F (class)[a]		
		1	2	3	1	2	3
6	0	0.89	1.70	0.80	4.50	17.00	3.10
7	0	0.83	3.50	1.90	3.90	76.00	16.00
8	0	0.50	0.92	0.48	1.40	5.20	1.10
9	0	1.70	1.20	0.55	17.00	4.00	1.40
11	0	0.94	1.20	0.60	5.00	9.20	1.70
17	0	0.70	1.20	0.71	2.80	8.80	2.40
31	0	0.69	2.00	0.87	2.70	24.00	3.60
32	0	1.40	1.70	1.40	12.00	18.00	9.10
33	0	0.92	1.80	1.10	4.90	20.00	6.10
41	0	1.00	1.10	0.88	5.90	7.30	3.70
51	0	0.66	0.88	0.81	2.50	4.80	3.10

[a] See Table 2 footnotes.

N-nitroso compounds, it can be seen that there is definite clustering of those compounds which induce tumors mainly in the esophagus, liver, and nasal turbinates into one area of the class. 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 Fig. 5. If $u_1 = -1$ and $u_2 = -1$, for example, and using class 1 m_i and b_{ia} values for 12 variables from Table 4, y_{ik} values can be calculated for the compound with this physicochemical description. The nonscaled parameters calculated for this compound are $f_R = 1.36$, $\sigma_r^* = -0.01$, $MR_R = 6.27$, $E_{sR} = -1.26$, $L_R = 3.11$, $B_{4R} = 2.13$, $f_{R1} = 2.60$, $E_{sR1} = -1.63$, $LR_{R1} = 6.86$, and $B_{4R1} = 4.55$, $\sigma_{R1}^* = 0.01$, $MR_{R1} = 7.95$. The physical property

TABLE 4

m_i AND b_{ia} VALUES

	f_R	σ_R^*	MR_R	E_{sR}	L_R	B_{4R}	f_{R1}	σ_{R1}^*	MR_{R1}	E_{sR1}	L_{R1}	B_{4R1}
Class 1												
m_i	0.47	−0.44	0.01	0.22	−0.04	−0.06	0.63	−0.58	0.30	0.37	0.24	0.16
b_{i1}	0.17	−0.06	0.22	−0.13	0.20	0.20	−0.33	0.03	−0.49	0.05	−0.49	−0.47
b_{i2}	0.35	−0.10	0.47	−0.25	0.45	0.43	0.12	−0.09	0.18	−0.08	0.22	0.28
Class 2												
m_i	−0.69	0.26	−0.53	0.16	−0.48	−0.47	−0.92	0.74	−0.43	−0.20	−0.45	−0.35
b_{i1}	0.53	−0.43	−0.12	0.39	−0.14	−0.17	−0.30	0.40	−0.01	−0.24	−0.05	−0.04
Class 3												
m_i	−0.46	0.69	0.32	−0.54	0.39	0.42	−0.61	0.65	−0.31	−0.57	−0.18	−0.09
b_{i1}	−0.14	0.43	0.39	−0.51	0.41	0.45	−0.02	−0.10	−0.01	0.05	0.00	0.07
b_{i2}	−0.15	0.51	−0.27	−0.28	−0.30	−0.17	−0.05	0.45	−0.08	−0.46	−0.14	−0.07
b_{i3}	0.23	−0.33	0.11	−0.36	0.18	−0.21	0.23	−0.16	0.29	−0.62	0.22	−0.12

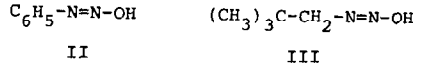
which appears to distinquish this cluster of compounds is σ^*, with the esophageal carcinogens characterized by R and R_1 being electronically neutral or electron withdrawing as modeled by σ^*.

DISCUSSION

Other attempts have been made to derive QSAR for carcinogenic *N*-nitroso compounds (*15*) using regression methods. Recently a classification study was reported dealing with such agents (*5b*). The latter study employed the linear learning machine as the classifier of *N*-nitroso compounds as carcinogenic vs noncarcinogenic.

The results here with the SIMCA method 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 subdivision of the compounds into three classes. This subdivision, while somewhat arbitrary, is consistent with the current ideas about the mechanism of carcinogenicity of *N*-nitroso compounds (*10, 17*). The observation of deutrium isotope effects *N*-nitrosodimethylamine-17-d_6 on the carcinogenic response of suggests that abstraction of an α-hydrogen is involved in the activation step leading to the ultimate carcinogen in this carcinogens. It is the nature of this intermediate(s) that must be resolved.

In this study, in which descriptions of only carcinogens are obtained, classification of an active compound as not being a member of any of the well-described classes cannot, technically, 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. It is instructive at this point to consider some of the substances which were incorrectly classified. The compounds *N*-nitrosomethylneopentylamine and *N*-nitrosomethylphenylamine are classified as being members of none of the three classes and both are carcinogens. If each is demethylated in the activation step, the species **II** and **III** below would be formed.



II III

Formation of a diazoalkyl-like intermediate would not be expected to be highly favorable from either of these intermediates.

Carbocation intermediates which may result from them would be extremely unstable and highly reactive. **II** and **III** may represent intermediates to new types of ultimate carcinogens from *N*-nitroso compounds.

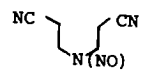
N-Nitrosodiallylamine is consistently classified as a false positive. This indicates that there is some chemical property not included in its description which may result in it being a poor substrate for activation and therefore being inactive. Such outliers as these are a natural and valuable result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

be considered in more detail. They may contain information which can account for their extraordinary classification results.

Since three compounds included in Table 1 have not been tested completely, this presents an opportunity to make predictions of their activities. The structures of these three compounds are given below and their classification results are in Table 3.

6

11



9

The di-*sec*-butyl compound is calculated to be inactive. This can also be deduced from the di-α-substitution which would make it a poor substrate for α-hydroxylation. Hence the calculations merely verify this although no parameters are included to directly identify this substitution pattern. *N*-Nitroso-bis(2-cyanoethyl) amine is placed in class 3 as *N*-nitroso-bis(2-chloropropyl) amine. Both, therefore, are expected to be carcinogens.

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 are 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 general structure, **I**, 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 these investigators, shows promise. However, such results should also be considered in their proper perspective. The present method, while theoretically based, gives an estimate of the carcinogenic potential just as the recently developed mutagenicity tests do, and this result should not be considered an end in itself but a part of the pharmacological profile of the compound. In this way the results can be used to make more sound judgments about further evaluation.

REFERENCES

1. H. Druckrey, R. Preussmann, S. Ivankovic, and D. Schmall, *Z. Krebsforsch.* **69**, 103 (1967).
2. P. N. Magee and J. M. Barnes, *Advan. Cancer Res.* **10**, 164 (1967).
3. W. Lijinsky and S. S. Epstein, *Nature (London)* **225**, 21 (1970).
4. (a) A. W. Andrews, L. H. Libault, and W. Lijinsky, *Mutat. Res.* **51**, 319 (1978). (b) T. Kameswar Rao, J. A. Young, W. Lijinsky, and J. L. Epler, *Mutat. Res.* **66**, 1 (1979).
5. P. C. Jurs, J. T. Chou, and M. Yaun, *J. Med. Chem.* **22**, 476 (1979). (b) J. T. Chou and P. C. Jurs, *J. Med. Chem.* **22**, 792 (1979).
6. (a) W. J. Dunn III and S. Wold, *J. Med. Chem.* **21**, 1001 (1978). (b) B. Norden, U. Edlund, and S. Wold, *Acta Chem. Scand. B* **32**. 1 (1979).

7. (a) W. J. DUNN III, S. WOLD, AND Y. C. MARTIN, *J. Med. Chem.* **21,** 922 (1978). (b) C. ALBANO, W. J. DUNN III, U. EDLUND, E. JOHANSSON, B. NORDEN, M. SJOSTROM, AND S. WOLD, *Anal. Chim. Acta* **103,** 429 (1978).
8. C. HANSCH, *Accts. Chem. Res.* **2,** 232 (1967).
9. S. WOLD, *Pattern Recog.* **8,** 127 (1976).
10. (a) W. LIJINSKY, H. W. TAYLOR, AND L. K. KEEFER, *J. Nat. Inst.* **57,** 1311 (1976). (b) K. K. PARK, J. S. Wishnok, AND M. C. ARCHER, *Chem.–Biol. Interact.* **18,** 349 (1977).
11. R. REKKER, "The Hydrophobic Fragment Constant." Elsevier, Amsterdam, 1977.
12. R. W. TAFT, "Steric Effects in Organic Chemistry" (M. S. Newman, Ed.). Wiley, New York, 1956.
13. W. J. DUNN III, *Eur. J. Med. Chem.* **12,** (1977).
14. A. VERLOOP, W. HOOGENSTRAATEN, AND J. TIPKER, "Drug Design" (E. J. Ariens, Ed.), Vol. VII. Academic Press, New York, 1977.
15. (a) J. S. WISHNOK AND M. C. ARCHER, *Brit. J. Cancer* **33,** 307 (1976). (b) J. S. WISHNOK, M. C. ARCHER, A. S. ADELMAN, AND W. M. RANK, *Chem.–Biol. Interact.* **20,** 43 (1978).
16. B. N. AMES, *Science* **204,** 587 (1979).
17. L. K. KEEFER, W. LIJINSKY, AND H. GARCIA, *J. Nat. Cancer Inst.* **51,** 299 (1973).
18. R. H. CARDY, W. LIJINSKY, AND P. K. HILDEBRANDT, *Exotoxicol. Environ. Safety* **3,** 29 (1979).