crystallization from aqueous ethanol (50%) afforded 650 mg (69%) of **7a**, mp 247–248 °C. Anal. ($C_{12}H_9N_3O_2$) C, H, N.

1-Methylpyrido[2,3-*b*]quinoxalin-2(1*H*)-one (8). A solution containing 381 mg (1.67 mmol) of **6a** and 134 mg (3.35 mmol) of sodium hydroxide in 40 mL of aqueous ethanol (50%) was refluxed 3 h. The solvent was concentrated under reduced pressure to 15 mL, diluted with water, and extracted with chloroform. The organic layer was dried (magnesium sulfate) and evaporated to a yellow solid. Recrystallization from methanol gave 160 mg (45%) of **8**, mp 243–245 °C. Anal. ($C_{12}H_9N_3O$) C, H, N.

Alternatively, **8** was prepared from **7a** by reduction with aqueous titanium trichloride as described for compound **6a**. The compound prepared in this fashion was identical in all respects with that obtained by base hydrolysis of **6a**.

[3-(Dimethylamino)quinoxalin-2-yl]acrylonitrile 5-Oxide (9b). A solution containing 500 mg (2.35 mmol) of either **2a** or **2b** and dimethylamine (23 mmol) in 125 mL of anhydrous methanol was stirred 16 h at 20 °C. The resulting suspension was filtered to give 337 mg (60%) of a red solid, mp 222–226 °C. The crude product was recrystallized from methanol to give 195 mg (35%) of **9b**,[10] mp 224–226 °C. The NMR spectrum (CDCl₃)

had absorptions at δ 3.10 [s, 6 H, N(CH₃)₂], 5.55 (d, 1 H, *J* = 12 Hz, CH=CHCN), 7.35–8.75 (m, 4 H, H-5, H-6, H-7, and H-8), 9.85 (d, 1 H, *J* = 12 Hz, C*H*=CHCN). The infrared spectrum had a nitrile absorption at 2222 cm⁻¹. Anal. ($C_{13}H_{12}N_4O$) C, H, N.

(3-*N*-Pyrrolidinequinoxalin-2-yl)acrylonitrile 5-Oxide (9c). This compound was prepared according to the method described for **10a**. The crude product was purified by trituration with ether/chloroform (10:1) to give **9c**[10] (60%), mp 215–216 °C. Anal. ($C_{15}H_{14}N_4O$) C, H, N.

(10) The olefinic coupling constant (*J* = 12 Hz) is consistent with cis stereochemistry. However, a definitive assignment cannot be made without comparison to the corresponding trans isomer, which could not be isolated from the complex mixture of products in the mother liquors.

Computer-Assisted Studies of Structure–Activity Relationships of *N*-Nitroso Compounds Using Pattern Recognition

Susan L. Rose and Peter C. Jurs*

*Department of Chemistry, The Pennsylvania State University, University Park, Pennsylvania 16802.
Received October 26, 1981*

Pattern-recognition techniques have been applied to the study of relationships between the molecular structure of nitrosamines and their carcinogenic potential. A set of 150 nitrosamines (112 carcinogenic and 38 noncarcinogenic) was used. Each compound was represented by a set of calculated molecular structure descriptors. Discriminants were found that could separate 146 of the compounds into the two activity classes based on a set of 22 descriptors. Internal consistency checking showed that the 22 descriptors used supported a meaningful discriminant. The results show that sufficient information is contained within the structures of *N*-nitroso compounds to allow classification into carcinogenic activity classes.

An abundance of research has been reported on the carcinogenic potential of *N*-nitroso compounds in recent years. Many compounds have been synthesized and tested for carcinogenic activity by repeated administration of small doses to animals. Most of these tests are long term and expensive. The tests have led to the following observations: (1) *N*-nitroso compounds are easily formed, (2) *N*-nitroso compounds have shown activity in a wide variety of species examined, (3) *N*-nitroso compounds are active even in small doses, (4) *N*-nitroso compounds are active through a variety of administrative routes, and finally, (5) *N*-nitroso compounds show organ-specific activity.[1-3]

Hecht and co-workers[4,5] have studied tobacco-specific nitrosamines. Lijinsky[6] has performed many studies of *N*-nitroso compounds. Some cross species testing has in-

dicated that species differences and dose rates are critical in assessing susceptibility but that all animal species are susceptible.[7] Some research has focused on nitrosamine formation by reactions of nitrites and amines[8] or through reactions of drugs with nitrites.[9,10]

Relationships between the molecular structure of *N*-nitroso compounds, their metabolism, and their carcinogenic potential have been studied extensively. Nitrosamines are thought to undergo metabolic activation to alkylating agents. The α carbon, adjacent to the *N*-nitroso group, has attracted the interest of investigators. The hydrogen atoms bonded to the α carbon have been studied using a variety of deuterated compounds to take advantage of the kinetic isotope effect. The results of these experiments have been confusing and contradictory and highlight organ specificity as a complicating factor.[11] The existence of competing metabolic pathways could explain the observations.[12,13] Studies focusing on metabolism of

(1) Lijinsky, W. In "New Concepts in Safety Evaluations"; Mahlman, M.; Shapirio, R.; Blumanbel, H., Ed.; Halsted Press: New York, 1979; Chapter 9.
(2) Lijinsky, W. In "N-Nitrosamines"; Anselme, J., Ed.; American Chemical Society: Washington, DC, 1979; Chapter 10.
(3) Gronon, X. *Chem. Biol. Interact.* 1980, *29*, 1–30.
(4) Hecht, S. S.; Chen, C. B.; Hoffmann, D. *Acc. Chem. Res.* 1979, *12*, 92–98.
(5) Hecht, S. S.; Chen, C. B.; Ohmori, T.; Hoffmann, D. *Cancer Res.* 1980, *40*, 298–302.
(6) Cardy, R. H.; Lijinsky, W.; Hildebrandt, P. K. *Ecotoxicol. Environ. Saf.* 1979, *3*, 29–35.

(7) Maduagwu, E. N.; Bassir, O. *Toxicol. Appl. Pharmacol.* 1980, *53*, 211–219.
(8) Walters, C. L.; Carr, F. P. A.; Dyke, C. S.; Saxby, M. J.; Smith, P. L. R. *Food Cosmet. Toxicol.* 1979, *17*, 473–479.
(9) Lijinsky, W,; Reuber, M. D. *Food Cosmet. Toxicol.* 1980, *18*, 85–87.
(10) Raisfeld, I. H.; Lin, C. *Biochem. Pharmacol.* 1979, *28*, 3451–3455.
(11) Lijinsky, W. unpublished work.
(12) Lijinsky, W.; Reuber, M. D. *Cancer Res.* 1980, *40*, 19–21.

tobacco-specific nitrosamines have appeared.[14-17] Some studies have been done with model compounds requiring no activation,[18] and other studies have investigated pancreatic carcinogenic nitrosamines.[19]

After about 30 years of testing and many concentrated efforts in the investigation of the mechanisms involved in nitrosamine carcinogenesis, no unifying theory is apparent. Much information has been uncovered, but it is not conclusive. However, the threat of these compounds is as grave a concern as ever, and animal testing does not seem to provide answers quickly or inexpensively enough.

In the last few years, interest has grown in short-term testing as a means for obtaining genetic toxicity data quickly and inexpensively. Since the molecular mechanisms involved in nitrosamine carcinogenicity are still not known, attention has been directed to a new approach involving the prediction of carcinogenicity of compounds through structure–activity relationships (SAR). These SAR methods seek correlations between the structures of tested compounds and the biological activities as measured in animal tests. In 1976, Wishnok and Archer[20] reported, with some degree of confidence, an estimate of carcinogenic activity for 51 nitrosamines through structure–activity relationships by correlating the number of carbon atoms with the carcinogenic activity. A year later, Singer, Taylor, and Lijinsky[21] linked liposolubility with nitrosamine carcinogenicity through quantitative SAR. Wishnok and co-workers[22] later reported a quantitative Hansch–Taft SAR for nitrosamine carcinogenicity which demonstrated that variation in carcinogenicity could be correlated with a number of molecular properties. Using 21 compounds, they determined water–hexane partition coefficients and electronic inductive effects of substituents at the α carbon to be the most important features, which suggested that transport properties were important in carcinogenicity. Wishnok and co-workers[23] have predicted organ specificity using physicochemical properties of N-nitrosodialkyl-amines. Partition coefficients, electronic factors, and a measure of steric hindrance gave a near perfect prediction of 19 compounds.

The previous structure–activity studies were limited in the number of compounds studied but did show that reasonable correlations could be drawn between the structure of compounds and their biological activity without a complete understanding of the underlying mechanisms involved. Chou and Jurs[24] expanded the approach to structure–activity relationships by applying computer-assisted mathematical and statistical methods to a large set of N-nitroso compounds. These methods

come under the broad heading of pattern-recognition techniques. The software used was developed by Jurs and his co-workers and is referred to as ADAPT.[25] Since physicochemical measurements were not available for such a large set of compounds, computer-generated parameters were used. A set of 15 molecular structure descriptors was found that could support a linear discriminant capable of separating 116 carcinogens from 28 noncarcinogens with only 5 compounds misclassified. The descriptors represented structural fragments, molecular branching, geometric parameters, environment parameters, and electronic properties. An internal consistency check was performed which resulted in an average correct classification rate of 91%. Some observations were made regarding the descriptors selected and current theories of N-nitroso compound carcinogenicity.[24]

A recent study of N-nitroso compounds has been presented by Dunn and Wold[26,27] using a pattern-recognition technique called SIMCA. This method performs principal components analysis for each class of compounds using physicochemical measurements as descriptors. Their study included 61 compounds divided into three classes for modeling. Only the carcinogenic compounds were modeled, and the N-nitroso compounds were classified by fitting them to the models. They reported an 88% correct classification of the carcinogens and stated that if a compound was not a member of one of the carcinogenic classes based on the models formed, the nonclassified compound could be a member of another carcinogenic class of compounds not yet modeled. Therefore, a nonclassified compound was not considered to be a noncarcinogen.

The nitrosamines lend themselves nicely to structure–activity studies using computer-assisted pattern recognition techniques. The available data set is large and diverse, not only in structural types but also in the complexity of its activity. Due to these characteristics, the compounds are difficult to analyze and understand through ordinary means. Small sets of compounds can be easily investigated by a chemist using nothing more than his own insight, but as the data sets grow in size and complexity, alternative approaches become attractive.

The investigation described here involved the study of a set of 150 nitrosamines (112 carcinogenic, 38 noncarcinogenic) and the development of discriminants to separate these compounds into the two activity classes. The compounds were represented only by calculated molecular structure descriptors.

Experimental Section

Methodology. The methods used in this research are implemented in an interactive computer software system known as ADAPT (Automated Data Analysis by Pattern recognition Techniques).[24] This set of computer software provides capabilities for performing structure–activity studies.

The chemical structures of the compounds comprising the data set are entered into computer disc files by sketching the structures on a graphics display terminal. Each structure is modeled using a force-field molecular mechanics model builder.

Descriptors, a means of encoding structural information in a numerical form, can be based on the entire molecule or on a part of it defined by a substructure. Substructures are chosen by the user to represent regions of interest within the molecule. The use of substructures allows the user to encode knowledge he has of his data set regarding hypotheses on metabolic activation,

(13) Lijinsky, W.; Reuber, M. D.; Saavedra, J. E.; Blackwell, B. N. *Carcinogenesis* 1980, *1*, 157–160.

(14) Hecht, S. S.; Chen, C. B.; Hoffman, D. *Cancer Res.* 1978, *38*, 215–218.

(15) Hecht, S. S.; Chen, C. B.; McCoy, D. G.; Hoffmann, D. *Cancer Lett.* 1979, *8*, 35–41.

(16) Hecht, S. S.; Chen, C. B.; Hoffmann, D. *J. Med. Chem.* 1980, *23*, 1175–1178.

(17) Hecht, S. S.; Young, R.; Chen, C. B. *Cancer Res.* 1980, *40*, 4144–4150.

(18) Saavedra, J. E. *J. Org. Chem.* 1979, *44*, 4511–4515.

(19) Gingell, R.; Brunk, G.; Nagel, D.; Pour, P. *Cancer Res.* 1979, *39*, 4579–4583.

(20) Wishnok, J. S.; Archer, M. C. *Br. J. Cancer* 1976, *33*, 307–311.

(21) Singer, G. M.; Taylor, H. W.; Lijinsky, W. *Chem Biol.-Interact.* 1977, *19*, 133–142.

(22) Wishnok, J. S.; Archer, M. C.; Edelman, A. S.; Rand, W. M. *Chem.-Biol. Interact.* 1978, *20*, 43–54.

(23) Edelman, A. S.; Kraft, P. L.; Rand, W. M.; Wishnok, J. S. *Chem.-Biol. Interact* 1980, *31*, 81–92.

(24) Chou, J. T.; Jurs, P. C. *J. Med. Chem.* 1979, *22*, 792–797.

(25) Stuper, A. J.; Brugger, W. E.; Jurs, P. C. "Computer Assisted Studies of Chemical Structure and Biological Function"; Wiley-Interscience: New York, 1979.

(26) Dunn III, W. J.; Wold, S. *J. Chem. Inf. Comput. Sci.* 1981, *21*, 8–13.

(27) Dunn III, W. J.; Wold, S. *Bioorg. Chem.* 1981, *10*, 29–45.

receptor interaction, etc. For example, the N–NO group in a nitrosamine is considered a highly reactive region either through α-carbon reactions or N-oxidation. Therefore, the N–NO group was used as a substructure in several descriptors. Four general classes of descriptors are available: physicochemical, topological, geometrical, and electronic.

Several types of descriptors can be generated by using substructures. A substructure count descriptor consists of the number of occurrences of the substructure in the structure being represented. Environment descriptors encode the surroundings of the substructure as imbedded within the target structure. The surroundings of the substructure can be described by a molecular connectivity computation, a path-tracing procedure, by electronic properties, or other methods.

A recently implemented descriptor routine encodes molecular symmetry. It was implemented because metabolic studies have suggested a difference in the potency between symmetrical and nonsymmetrical compounds. The routine generates a symmetry parameter by dividing the number of uniquely connected nonhydrogen atoms by the total number of nonhydrogen atoms. Three example values are as follows: dioctylnitrosamine, 11/19 = 0.579; N-nitrosomorpholine, 6/8 = 0.75; N-nitrosonornicotine, 13/13 = 1.00.

A variety of descriptors can be computer generated on a theoretical basis. The utility of the descriptors depends on the nature of their application and characteristics of the study. Care must be taken in the development and selection of descriptors to obtain the best results in a given study. Whalen-Pedersen and Jurs[28] reported that the probability of finding a linear discriminant that completely separates two classes by chance can be kept extremely low by ensuring that the number of compounds divided by the number of descriptors or features is greater than about three and that the number of patterns in the less populated class exceeds the total number of descriptors. The goal of the analysis is to thoroughly represent the compounds with a minimum number of chemically plausible descriptors. Therefore, the user must rely on insight and knowledge of the data set to aid in descriptor generation. This proper selection of descriptors is the key to any successful structure–activity study.

Many of the available descriptors encode related information. Therefore, multicollinearities among the descriptors can be a problem, giving rise to numerical instabilities in the analysis phase of the study. Mulitple linear regression analysis can be used to discover multicollinearities so they can be eliminated. This allows one to find a set of descriptors that minimizes the overall redundancy.

An important step in the overall study is the selection of sets of descriptors to analyze. This is usually an iterative process guided by the experience of the user. A method for automatically selecting the best subset of descriptors for use in discriminant analysis has been implemented. This involves the computation of all possible U statistics for all possible subsets of a given list of descriptors. The U statistic is the ratio of the determinants of the within class cross-products matrix and the total cross-product matrix. Because the computation time doubles with each additional descriptor in the list, it is only practical to use this routine to refine a proven set of descriptors with minimal additions.[29]

The variance method of feature selection[25] can also be used to identify meaningful descriptors by ranking them according to the relative variation of the discriminant function coefficients within a series of discriminants. The descriptors corresponding to larger relative variations can be identified and discarded in the order of their decreasing variances.

Data Set. In 1979, Chou and Jurs[24] studied a set of 149 nitrosamines and nitrosamides using the ADAPT software system. Their structure–activity study of N-nitroso compounds resulted in 94% of the compounds being correctly classified by discriminants based on 15 molecular structure descriptors. Since that N-nitroso compound study, test results for many additional nitrosamines have been reported, and sufficient data became

Chart I. Substructures Used for Development of Descriptors
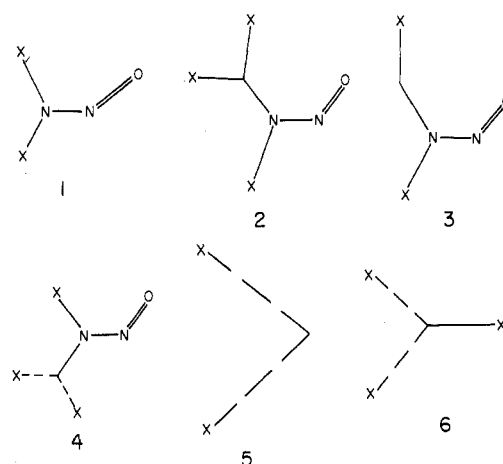
available to study a set of nitrosamines alone. Here, the 115 nitrosamines included in the Chou and Jurs data set were combined with the more recently tested nitrosamines to form a larger data set solely comprised of nitrosamines. The nitrosamides were left out of the present study because they are overwhelmingly carcinogens and can therefore bias the study. When more testing was completed, 4 of the 115 compounds were found to have activities different from those observed earlier, and these most recently reported activities were used in this work. Dr. William Lijinsky of the Frederick Cancer Research Center provided the test results for most of the compounds. The remainder were obtained from the literature.[4,30-39]

The nitrosamines data set used in this project contained the 150 compounds listed in Table I. The data were separated into two classes: class 1 consisted of 98 carcinogenic compounds and 14 weakly carcinogenic compounds, and class 2 contained 38 noncarcinogens. The set of compounds contains alkyl-substituted N-nitrosamines, piperidines, pyrrolidines, morpholines, piperazines, hydrazines, hyroxylamines, and other structural types. Thus, there is some degree of structural diversity within the data set. Discrepancies exist between the activities in Table I and recently published results for compounds **101**, **111**, **126**, **129**, **136**, and **150**. These are detailed in the footnotes to Table I. Changes in the biological activity data upon retesting of the compounds are to be expected in SAR studies using large data sets. These changes do not invalidate the results reported here.

Results and Discussion

To start the descriptor phase of the work, all the automated descriptors supported by ADAPT were generated. Five fragment descriptors provided structural information concerning the number of oxygen atoms, carbon atoms, nitrogen atoms, basis rings, and ring atoms. The number of hydrogens on the α and β carbons was also determined. Five molecular connectivity descriptors provided branching information. These were the simple path one corrected for rings, path one corrected for rings and calculated using the valences of the heteroatoms, path two molecular connectivity, cluster three molecular connectivity, and path-cluster four molecular connectivity. The molar refractivity

(28) Whalen-Pedersen, E. K.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 264–266.

(29) McCabe, G. P., Jr. *Technometrics* **1975**, *17*, 103–109.

(30) Wiessler, M; et. al. *Z. Krebsforsch.* **1974**, *82*, 257–258.

(31) Wiessler, M; et. al. *Z. Krebsforsch.* **1975**, *83*, 205–206.

(32) Wogan, G. N.; Paglialunga, S.; Archer, M. C.; Tannenbaum, S. R. *Cancer Res.* **1975**, *35*, 1981–1984.

(33) Wiessler, M.; et. al. *Z. Krebsforsch.* **1976**, *85*, 47–49.

(34) Wiessler, M.; Schmahl, D. *Z. Krebsforsch.* **1976**, *88*, 25–31.

(35) Lee, K.; Gold, B.; Mirvish, S. S. *Mutat. Res.* **1977**, *48*, 131.

(36) Wiessler, M.; et. al. *Z. Krebsforsch.* **1978**, *91*, 317–322.

(37) Schweinsberg, F.; et. al. *Z. Krebsforsch.* **1978**, *92*, 235–241.

(38) Bertram, B.; et. al. *Cancer Lett.* **1979**, *6*, 57–60.

(39) Preussmann, R.; Habs, M.; Pool, B. L. *J. Natl. Cancer Inst.* **1979**, *62*, 153–156.

Table I. Compounds Comprising the Nitrosamine Data Set

no.	compound	potency[a]	no.	compound	potency[a]
1	N-nitrosopiperidine	+	76	3,5-dimethyl-N-nitrosopiperazine	+
2	3-methyl-N-nitrosopiperidine	+	77	4-keto-3-methyl-N-nitrosopiperidine	+
3	4-methyl-N-nitrosopiperidine	+	78	N,N-dinitrosoperhydropyrimidine	+
4	3-hydroxy-N-nitrosopiperidine	+	79	2-(nitrosomethylamino)pyridine	+
5	4-hydroxy-N-nitrosopiperidine	+	80	N-nitrosoephedrine	+
6	4-keto-N-nitrosopiperidine	+	81	(methoxymethyl)methylnitrosamine	+
7	4-chloro-N-nitrosopiperidine	+	82	(1-methoxyethyl)ethylnitrosamine	+
8	3,4-dichloro-N-nitrosopiperidine	+	83	(methoxymethyl)ethylnitrosamine	+
9	3,4-dibromo-N-nitrosopiperidine	+	84	(1-methoxyethyl)methylnitrosamine	+
10	N-nitroso-1,2,3,6-tetrahydropyridine	+	85	N-nitrosooxazolidine	+
11	3-chloro-N-nitrosopiperidine	+	86	N-nitrosotetrahydro-1,3-oxazine	+
12	3,4-dichloro-N-nitrosopyrrolidine	+	87	N-nitroso-N-ethylbenzylamine	+
13	N-nitrosonornicotine	+	88	N-nitroso-O,N-diethylhydroxylamine	+
14	N-nitrosomorpholine	+	89	(methylbutyroxy)methylnitrosamine	+
15	2,6-dimethyl-N-nitrosomorpholine	+	90	(acetoxylmethyl)methylnitrosamine	+
16	dinitrosopiperazine	+	91	ethyl(acetoxymethyl)nitrosamine	+
17	2-methyldinitrosopiperazine	+	92	n-propyl(acetoxymethyl)nitrosoamine	+
18	2,5-dimethyldinitrosopiperazine	+	93	n-butyl(acetoxymethyl)nitrosamine	+
19	2,6-dimethyldinitrosopiperazine	+	94	N-nitroso-3-methyl-1,2,3,6-tetrahydropyridine	+
20	dinitrosohomopiperazine	+	95	N-nitroso-1,2,3,4-tetrahydropyridine	+
21	N-nitrosoazetidine	+	96	4-(N-methyl-N-nitrosamino)pyridinebutanone	+
22	N-nitrosoheptamethylenimine	+	97	4-(N-methyl-N-nitrosamino)pyridinebutanal	+
23	N-nitrosohexamethylenimine	+	98	2-methyl-N-nitrosomorpholine	+
24	N-nitrosooctamethylenimine	+	99	2-methyl-N-nitrosopiperidine	w
25	N-nitrosodimethylamine	+	100	3,5-dimethyl-N-nitrosopiperidine	w
26	N-nitrosodiethylamine	+	101	4-phenyl-N-nitrosopiperidine[b]	w
27	N-nitrosobis(2-chloroethyl)amine	+	102	4-tert-butyl-N-nitrosopiperidine	w
28	N-nitrosobis(2-methoxyethyl)amine	+	103	N-nitrosopyrrolidine	w
29	N-nitrosobis(2-ethoxyethyl)amine	+	104	N-nitroso-3-pyrroline	w
30	N-nitrosodi-n-propylamine	+	105	N-nitrosothiomorpholine	w
31	N-nitrosobis(2-oxopropyl)amine	+	106	N-nitrosododecamethylenimine	w
32	N-nitroso-N-methylethylamine	+	107	N-nitrosodiisopropylamine	w
33	N-nitroso-N-methylundecylamine	+	108	N-nitrosobis(2-hydroxypropyl)amine	w
34	N-nitroso-N-methyldodecylamine	+	109	2,2-dichloro-di-n-propylnitrosamine	w
35	N-nitroso-N-methyl-2-phenylethylamine	+	110	N-nitroso-N-ethylisopropylamine	w
36	N-nitroso-N-methylneopentylamine	+	111	N-nitrosodiethanolamine[b]	w
37	N-nitroso-N-methylphenylamine	+	112	N-nitroso-N-methyl-1-phenylethylamine	w
38	N-nitroso-N-methylcyclohexylamine	+	113	2,6-dimethyl-N-nitrosopiperidine	–
39	N-nitrosodi-n-butylamine	+	114	2,2,6,6-tetramethyl-N-nitrosopiperidine	–
40	N-nitrosodi-n-pentylamine	+	115	2-carboxy-N-nitrosopiperidine	–
41	N-nitrosodiphenylamine	+	116	4-carboxy-N-nitrosopiperidine	–
42	N-nitroso-N-N-methylvinylamine	+	117	N-nitrosoguvacoline	–
43	N-nitroso-N-methylallylamine	+	118	methyl N-nitrosophenidylate	–
44	N-nitroso-N-methyl-n-propylamine	+	119	2,5-dimethyl-N-nitrosopyrrolidine	–
45	N-nitroso-N-methyl-n-butylamine	+	120	2-carboxy-N-nitrosopyrrolidine	–
46	N-nitroso-N-methyl-n-pentylamine	+	121	2-carboxyl-4-hydroxy-N-nitrosopyrrolidine	–
47	N-nitroso-N-(2-hydroxypropyl)-n-propylamine	+	122	N-nitrosophenmetrazine	–
48	N-nitroso-N-(2-oxopropyl)-n-propylamine	+	123	2,3,5,6-tetramethyl-N,N-dinitrosopiperazine	–
49	N-nitroso-N-methylbenzylamine	+	124	N-nitrosopiperazine	–
50	N,N'-dinitroso-N,N'-dimethylethylenediamine	+	125	4-methyl-N-nitrosopiperazine	–
51	N,N'-dinitroso-N,N'-dimethyl-1,3-propanediamine	+	126	4-benzoyl-3,5-dimethyl-N-nitrosopiperazine[c]	–
52	N-nitroso-N-ethylvinylamine	+	127	N-nitrosobis(2-cyanoethyl)amine	–
53	N,N'-dinitroso-N,N'-diethylethylenediamine	+	128	N-nitrosobis(2,2-diethoxyethyl)amine	–
54	N-nitroso-N-butyl-n-pentylamine	+	129	diisobutylnitrosamine[b]	–
55	N-nitroso-N-ethyl-n-butylamine	+	130	di-sec-butylnitrosamine	–
56	N-nitroso-N-methylaminosulfolane	+	131	N-nitroso-N,O-dimethylhydroxylamine	–
57	N-nitroso-N-ethyl-2-hydroxyethylamine	+	132	N-nitrosodi-n-octylamine	–
58	N-nitrosobis(acetoxyethyl)amine	+	133	N-nitrosodiallylamine	–
59	N-nitroso-N-n-butyl-4-hydroxybutylamine	+	134	N-nitrosodicyclohexylamine	–
60	N-nitroso-N-n-butyl-3-hydroxypropylamine	+	135	N-nitrosodibenzylamine	–
61	N-nitroso-N-n-propyl-4-hydroxybutylamine	+	136	N-methyl-N-heptylnitrosamine[b]	–
62	N-nitroso-N-(2-hydroxyethyl)-n-butylamine	+	137	N-nitroso-4-(methylamino)azobenzene	–
63	N-nitroso-N-(2-oxopropyl)-n-butylamine	+	138	N-nitroso-N-ethyl-tert-butylamine	–
64	N-nitroso-N-methyl-2-chloroethylamine	+	139	N-nitroso-N-n-butyl-3-hydroxypropylamine	–
65	N-nitroso-N-methylaminoacetonitrile	+	140	N-nitrosodiacetonitrile	–
66	N-nitrososarcosine	+	141	N-nitroso-N-(1,1-dimethyl-3-oxobutyl)-methylamine	–
67	N-nitrososarcosine ethyl ester	+			
68	N-nitroso-N-ethyl-4-picolylamine	+	142	4-(N-nitroso-N-methylamino)benzaldehyde	–
69	N-nitroso-N'-carbethoxypiperazine	+	143	N-nitroso-L-proline ethyl ester	–
70	N-nitroso-3,6-dihydro-1,2-oxazine	+	144	trinitrosohexahydro-1,3,5-triazine	–
71	N-nitrosotetrahydro-1,2-oxazine	+	145	N-nitrosoindoline	–
72	N-nitrosodecamethylenimine	+	146	tert-butylmethylnitrosamine	–
73	N,4-dinitroso-N-methylaniline	+	147	3-(nitrosomethylamino)pyridine	–
74	1-[4-(N-methyl-N-nitrosoamino)benzylidene]-indene	+	148	4-(nitrosomethylamino)pyridine	–
			149	N-nitroso-N-methyl-2-(2-phenyl)propylamine	–
75	4-[4-methyl(N-nitrosoamino)styryl]quinoline	–	150	4-acetyl-3,5-dimethyl-N-nitrosopiperazine[b]	–

[a] + = carcinogen; w = weak carcinogen; – = noncarcinogen. [b] More recent investigations show these compounds to be strong carcinogens: Lijinsky, W.; Singer, G. M.; Reuber, M. D. Carcinogenesis 1981, 2, 1045–1048. Lijinsky, W.; Reuber, M. D.; Manning, W. B. Nature (London) 1980, 288, 589–590. Singer, S. S.; et al. Cancer Res. 1981, 41, 1034–1038. Lijinsky, W.; Reuber, M. D. Cancer Lett. 1981, 14, 297–302. Also reported in unpublished research. [c] A more recent investigation reports this compound to be a weak carcinogen. Singer, S. S. et al. Cancer Res. 1981, 41, 1034–1038.

Table II. Twenty-Two Descriptors Giving 97% Classification of the 150 Compound Nitrosamine Data Set

no.	name	descriptor	mean	SD	NNZ[a]	simple R[b]
1	FRAG 4	no. of N atoms	2.31	0.685	150	0.361 (14)
2	MOLC 4	path 2 molecular connectivity	2.55	0.982	150	0.858 (8)
3	GEOM 3	intermediate principal moment	0.138	0.138	150	0.507 (2)
4	GEOM 4	largest PM/intermediate PM	4.23	3.18	150	0.240 (20)
5	ENVR 1	mol. conn. envir. of SS 1	2.35	0.378	150	0.542 (12)
6	SCAV 1	av σ charge of SS 1	-0.0490	0.00684	150	0.192 (19)
7	SCMP 2	most positive σ charge of SS 2	0.00974	0.0232	25	0.811 (13)
8	MREF 1	molar refractivity	41.2	12.4	150	0.858 (2)
9	CSEP 1	charge separation	3.51	1.29	150	0.280 (1)
10	SCMP 3	most positive σ charge of SS 3	0.0442	0.0223	121	0.705 (14)
11	NOHC 2	no. of H on β carbons	3.75	2.81	137	0.388 (13)
12	SSS-2	no. of SS 2 occurrences	0.233	0.595	25	0.813 (13)
13	PATH-2	no. of paths from SS 2	11.7	28.8	25	0.813 (12)
14	SSS-3	no. of SS 3 occurrences	1.46	0.974	121	0.705 (10)
15	PATH-3	no. of paths from SS 3	37.7	26.4	121	0.650 (10)
16	SSS 2	no. of SS 2 occurrences (ring specific)	0.0933	0.335	12	0.382 (11)
17	SSS 3	no. of SS 3 occurrences (ring specific)	0.787	0.909	72	0.802 (18)
18	PATH 3	no. of paths from SS 3 (ring specific)	17.8	20.8	72	0.802 (17)
19	SCMN 5	most negative σ charge of SS 5	-0.00506	0.0112	26	0.395 (10)
20	SSS-4	no. of SS 4 occurrences	0.0733	0.262	11	0.838 (22)
21	SCMP 5	most positive σ charge of SS 5	-0.00026	0.0166	26	0.221 (1)
22	SCAV 6	average σ charge of SS 6	0.00668	0.0202	26	0.838 (20)

[a] Number of nonzero occurrences. [b] Correlation coefficient with descriptor identified in parentheses.

was calculated for each compound. The three-dimensional coordinates of the modeled structures were used to calculate the three principal moments, their three ratios, and the molecular volume. Two descriptors coded the total number of paths in the structures and the average number of paths per atom. The first three substructures shown in Chart I, based on the N-nitroso group, were used to develop substructure counts, environment descriptors, and σ charge descriptors.

This original pool of descriptors was screened to eliminate multicollinearities. This was done by successively considering each descriptor as the dependent variable and regressing it against the remaining descriptor pool. Those which were found to be exact linear combinations of other descriptors, as well as those with high multiple correlation coefficients, were eliminated from consideration. Note that this screening is done without reference to the class labels of the compounds. It only considers the structural descriptors themselves and therefore does not bias the eventual classification results. A set of 15 descriptors passed the screening and was capable of supporting a linear discriminant that correctly classified 140 out of the 150 compounds in the data set (93%). The ten misclassified compounds were numbers **69, 107, 129, 132, 135, 136, 144, 147,** and **148** in Table I.

The misclassified compounds were reviewed to identify structural features that were not being adequately described. Two descriptors were generated to encode aspects of the symmetry of the compounds. Substructure count and path environment descriptors were generated using substructures **2, 3,** and **4** of Chart I. Several adjustable parameters dealing with ring perception and substructure counting were varied during the generation of these environment descriptors. The resulting set of descriptors was evaluated using pattern-recognition methods and the U-statistic routine. Then, two additional substructures, **5** and **6** of Chart I, were developed and used to calculate additional σ charge descriptors. A descriptor was also implemented that coded the distance between the most positive and most negative σ charge in the structure. At all times we recognized the limitations imposed on the number of descriptors that could be considered, and we stayed well within these limits.

After testing of a number of combinations of descriptors, the set of 22 descriptors shown in Table II was found. No

Table III. Mean and Standard Deviations for the 22 Descriptors and the Descriptor Values for 1-[4-(N-Methyl-N-nitrosamino)benzylidene]indene

| no.	name	mean	SD	value
1	FRAG 4	2.31	0.685	2.00
2	MOLC 4	2.55	0.982	4.71
3	GEOM 3	0.138	0.138	0.0329
4	GEOM 4	4.23	3.18	4.50
5	ENVR 1	2.35	0.378	1.87
6	SCAV 1	-0.0490	0.00684	-0.0533
7	SCMP 2	0.00974	0.0232	0.00
8	MREF 1	41.2	12.4	80.31
9	CSEP 1	3.51	1.29	4.31
10	SCMP 3	0.0442	0.0223	0.00
11	NOHC 2	3.75	2.81	2.00
12	SSS-2	0.233	0.595	0.00
13	PATH-2	11.7	28.8	0.00
14	SSS-3	1.46	0.974	0.00
15	PATH-3	37.7	26.4	0.00
16	SSS 2	0.0933	0.335	0.00
17	SSS 3	0.787	0.909	0.00
18	PATH 3	17.8	20.8	0.00
19	SCMN 5	-0.00506	0.0112	-0.0318
20	SSS-4	0.0733	0.262	1.00
21	SCMP 5	-0.00026	0.0166	-0.0245
22	SCAV 6	0.00668	0.0202	0.0196

single additional descriptor could be found that would substantially increase the classifying ability of the discriminants beyond that achieved by this set of 22 descriptors. Table II shows the mean and standard deviation for each descriptor and the number of compounds possessing each descriptor. The mean number of nonzero descriptors per compound was 13.5 with a standard deviation of 1.8. The minimum number of nonzero descriptors for any compound was 8 (for compound **25** in Table I) and the maximum number was 18 (for compounds **13, 80, 82, 97, 110, 112, 118,** and **122**). The final two columns show the largest simple correlation coefficient for each descriptor and the other descriptor with which the correlation exists. An example of the values of these descriptors is shown in Table III for compound **74**, the carcinogen 1-[4-(N-methyl-N-nitrosoamino)benzylidene]-indene, shown in Chart II.

A principal components analysis was performed on the correlation matrix for the 22 descriptors representing the 150 compounds. The first two eigenvalues contribute 37%

Table IV. ADAPT Classification Results for 22 Descriptors

name	analysis method	total no. wrong	% correct +	% correct −	% correct total
TILSQ	iterative least squares (init with Bayes WV)	8	97.32	86.84	94.67
ALS	adaptive least squares	4	99.11	92.11	97.33
BAL	bayes linear function	17	91.96	78.95	88.67
BAQ	bayes quadratic function	11	95.54	84.21	92.67
KNN	first nearest neighbor	28	90.18	55.26	81.33
KNN	three nearest neighbors	31	91.96	42.11	79.33
SIMPLEX	sequential simplex optimization (init with Bayes WV)	23	94.60	55.26	84.67
		12	95.50	81.60	92.00
LDFA	linear discriminant function anal.	17	91.96	78.95	88.67
TMAIN	linear learning machine	4	99.11	92.11	97.33
	predictive ability (29 trainings)		90.91	68.57	85.52
	predictive ability (14 trainings)		86.92	60.61	80.71

Chart II.
1-[4-(N-Methyl-N-nitrosoamino)benzylidene]indene
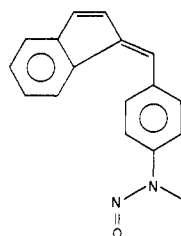
of the cumulative variance. It takes 13 eigenvalues to exceed 95% and 19 to exceed 99% of the cumulative variance. One way to assess the degree of clustering of the points comprising a data set is to plot the projections of the patterns onto the first two principal components for visual examination. Such a plot for these data shows some clustering; however, with only 37% of the cumulative variance represented by these two largest components, the plots contain little of the total information.

Another measure of the clustering tendency of the data set was implemented as follows. For the actual data set, a histogram was constructed of all pairwise interpoint distances (after autoscaling). Then a randomly chosen data set of multivariate normal points, also of 22 descriptors each, was developed, and the same histogram was generated. The random data set was constructed so that it had the identical variance–covariance matrix as the actual data set. The histogram for the actual data set shows less central tendency than the histogram for the random data set. This shows that there is some measure of clustering, but it is not a quantitative measure.

Yet another measure of the degree of clustering of the set of compounds is found in the K-nearest-neighbor distances. These were computed after the data were autoscaled so that each descriptor had a mean of 0 and a standard deviation of 1000. The mean first-nearest-neighbor distance for all 150 compounds was 1790, for 3-NN 2356, for 5-NN 2671, for 7-NN 2886, and for all [(150) (149)/2 = 44 700] interpoint distances it was 6500. The average of the three nearest-neighbor distances for the carcinogens ranged from a minimum of 628 (for compound 10 of Table I) to a maximum of 6778 (for compound 97), with a mean of 2050. The average of the three nearest-neighbor distances for the noncarcinogens ranged from a minimum of 927 (for compound 133) to a maximum of 6599 (for compound 123), with a mean of 3218. Some compounds have near neighbors that are structurally similar, while some have quite dissimilar near neighbors. For example, compound 103, N-nitrosopyrrolidine, has nearest neighbors that are close with the 3-NN average distance of 670 for a tetrahydropyridine (compound 10), a piperi-

dine (compound 1), and a pyrroline (compound 104). N-Nitrosodimethylamine (compound 25) has nearest neighbors that are more distant with the 3-NN average distance of 2857 for three alkyl compounds (42, 43, and 32). Compound 74, the example of Chart II, has quite distant nearest neighbors with the 3-NN average distance of 4765 for a benzaldehyde (compound 142), a diphenyl (compound 41), and an aniline (compound 73).

The ADAPT classification results using the 22 descriptors are listed in Table IV. The number wrong, the percentage correct for each class, and the total percentage correct are given for eight ADAPT classifiers. The adaptive least-squares classifier[40] and the linear learning machine gave the best classification results. Each routine correctly classified 99% of the carcinogenic class and 92% of the noncarcinogenic class. Exclusion of the four compounds, 69, 123, 129, 139, of Table I, resulted in a set of 146 compounds (111 carcinogens and 35 noncarcinogens) that could be completely classified by linear discriminants. The overall classification results for the various methods ranged from 81 to 97%, whereas the classification results for the carcinogenic compounds were much higher, ranging from 90 to 99% correct.

The validity of the descriptors was tested using a procedure called an internal consistency check. This method involved the generation of a series of subsets of the entire data set leaving out a given number of structures to be used as a test set. The linear learning machine was used to develop discriminants based on the subsets. After a discriminant was developed using the subset, the remaining compounds were classified by the discriminant. Twenty-nine training sets were developed excluding the four misclassified compounds and leaving out five compounds from the test set. Each discriminant was formed based on a set of 141 compounds; then it was used to classify the five compounds left out of the analysis. This was repeated for each training set. The "predictive ability" is the average of all the classification results. A second internal consistency check was done with 14 training sets and 10 compounds left out of each set. The results of both tests are presented in Table IV. The predictive ability that would be achieved as a result of random guessing is defined by the square of the fraction of the compounds in class 1 plus the square of the fraction of compounds in class 2. For the nitrosamine data set, this random predictive ability would be 62.2%. The results obtained in this study exceed this figure significantly. The predictive ability results were 85.5% using 29 trainings and 80.7% using 14 trainings. The higher percentage correct classification results of 90.9% and 86.9% for carcinogenic compounds reflect a

(40) Moriguchi, I.; Komatsu, K.; Matsushita, Y. *J. Med. Chem.* 1980, *23*, 20–26.

Table V. Descriptor Contribution by Excluding
Each Feature

no.	descriptor left out	no. wrong +	no. wrong −	% correct
1	FRAG 4	9	15	84.00
2	MOLC 4	17	21	74.67
3	GEOM 3	7	14	86.00
4	GEOM 4	11	18	80.67
5	ENVR 1	15	18	78.00
6	SCAV 1	12	12	84.00
7	SCMP 2	5	12	88.67
8	MREF 1	8	14	85.33
9	CSEP 1	6	13	87.33
10	SCMP 3	18	12	80.00
11	NOHC 2	11	12	84.67
12	SSS 2	12	13	83.33
13	PATH 2	9	13	85.33
14	SSS 3	6	20	82.67
15	PATH 3	8	18	82.67
16	SSS 2	3	12	90.00
17	SSS 3	6	11	88.67
18	PATH 3	10	14	84.00
19	SCMN 5	12	17	80.67
20	SSS 4	7	13	86.67
21	SCMP 5	5	11	89.33
22	SCAV 6	9	10	87.33

greater internal consistency among these compounds. Correct classification of the carcinogenic compounds is more important when this method is used as a screening tool.

The relative contribution of each descriptor to the overall classification scheme was examined. A training set excluding the four misclassified compounds was generated. A discriminant developed by the linear learning machine was used for the analysis. The 146 compounds were classified with one descriptor purposely left out at a time, and the results obtained are shown in Table V. The percentage figures show what fraction of the 150 compounds could be correctly classified with each subset of 21 descriptors. Removal of the molecular connectivity descriptor, MOLC 4, damaged the discriminant the most, with only 112 of the 150, or 75%, of the compounds correctly classified.

Conclusions

Pattern-recognition techniques have been applied to the study of relationships between molecular structure and carcinogenic potential. A linear discriminant was developed that could separate 111 carcinogenic from 35 noncarcinogenic nitrosamines. The fact that this discriminant could be developed, without violating statistical constraints, suggests that information concerning the carcinogenic potential of N-nitroso compounds is contained in their structures. The high predictive ability obtained in the internal consistency checks demonstrates the validity of the method.

The classification results are dependent on the combination of descriptors used in the analysis. The compounds used in these studies were structurally diverse, while maintaining some similarities. Many factors determine whether a compound is carcinogenic or not. The compounds cannot be dissected into active and inactive regions, but they must be considered in their entirety. Therefore, one can not simply look at the compounds and predict their carcinogenicity because no single region of the structure is responsible for the biological activity in all the nitrosamines. The substructure-based descriptors used in these studies are also diverse, encoding many different structural regions. Like the compounds themselves, the descriptors should always be considered as a

set because their individual contributions are relative and depend on the other members of the descriptor set. An evaluation of the amount each descriptor adds to the classification results must be recognized as the contribution relative to the overall set of features. It would be erroneous to conclude that any one descriptor encodes the carcinogenic or mutagenic potential of a compound.

This nitrosamine study does not contradict any of the metabolic theories presented in the literature. Several of the descriptors, used to separate 97% of the data correctly, support the theories, emphasizing the importance of the α and β positions adjacent to the N–NO group as well as the importance of the nitrogen position itself. Four of the substructures used represented the N–NO group and the atoms in the α and β positions. These four substructures were used in 12 of the 22 descriptors. The number of hydrogens on the β carbons was also included in the descriptor set. When this descriptor was deleted from the set, the classification results decreased almost 10%. The single fragment descriptor that survived feature selection contained the number of nitrogens.

The importance of the electronic properties of these compounds is supported by the inclusion of eight electronic descriptors. The molar refractivity and the charge separation descriptors contributed more than 5% to the classification results. Six σ charge descriptors were used, including three most positive, two average, and one most negative partial charges. Two of these descriptors, the most positive σ charge of substructure **3** and the most positive σ charge of substructure **5**, were shown in the descriptor contribution test to contribute about 13% to the observed separation. Electronic properties in molecules not only affect the reactivity of the compounds but can also influence transport properties in a biological system.

Three of the surviving descriptors encoded information concerning the size and shape of the compounds. One was the intermediate principal moment and the other was a ratio of the largest principal moment to the intermediate principal moment. The path two molecular connectivity descriptor is a measure of extent. This descriptor was also observed to individually contribute most, nearly 20%, to the set of features used in this study. The size and shape of compounds influences their transport properties through a biological system as well as their steric hindrance at the reactive site.

The descriptors used in the classification of the nitrosamines span the classes of descriptors and represent a wide diversity of information. The diversity of the descriptors is not suprising, since nitrosamines themselves are so diverse. Also, this diversity supports the proposal of competing pathways.

The results obtained in this study could be improved if more compounds were included in the less populated class. The discriminants developed in this research could be used to predict the carcinogenicity of N-nitroso compounds and thus direct testing toward compounds likely to be benign. When the compounds are more evenly distributed between the two classes, the classification results and the predictive ability are generally improved.

The classification results shown in this study are excellent, and they show that large sets of data can be represented by a limited number of molecular structural descriptors. Errors in the biological activity data are likely when information is collected across many testing procedures and laboratories. The known classifications of the nitrosamines are dependent on animal testing in many different laboratories. Four of the 115 compounds taken from the Chou and Jurs study showed activity different

from that originally assigned after further testing was done. Other activity classifications have changed during the course of this work (as noted in Table I). The inability to achieve 100% separation of the nitrosamines could be due to the lack of appropriate descriptors, but it is also probable that the biological activity data were not completely accurate.