Articles

Computer-Assisted Structure-Activity Studies of Chemical Carcinogens. Aromatic Amines

Kohtaro Yuta and Peter C. Jurs*

Department of Chemistry, The Pennsylvania State University, University Park, Pennsylvania 16802.
Received August 18, 1980

Studies of molecular structure–carcinogenicity relations for a set of 157 aromatic amines are reported. A computer-assisted approach using pattern-recognition methods was used to develop a series of discriminants for aromatic amino carcinogenic potential. The 157 compounds were divided into subsets according to tumor site, route of administration, and activity. Sets of calculated molecular structure descriptors were generated that could support linear discriminant functions able to separate sets of active carcinogens from inactive compounds. Prominent among the important structural descriptors were those coding sizes and shapes of the amines. The pattern-recognition results were not strongly affected by differences in active site, and the study showed that mixed data sets could be used in computer-assisted structure–carcinogenicity studies.

Aromatic amino compounds have been studied and a great deal of data accumulated over the years. The first report on the carcinogenicity of aromatic amines was by Rehn[1] in 1895 who described bladder cancers among employees in a Swiss dye factory. Many studies have followed, in part because aromatic amines are strongly related to industrialization and humans have been exposed to them widely. For example, aniline is employed world-wide in numerous industrial processes, fluorene compounds were originally developed as insecticides, and compounds such as butter yellow have been used as food additives. Primary aromatic amines are commonly used in industrial azo dye syntheses and as antioxidants in rubber products. Thus, the study of the carcinogenic potential of this class of compounds is of widespread interest.

A variety of hypotheses regarding the relationships between structure and activity of aromatic amines have been advanced. These compounds usually produce tumors at a site remote from the site of administration, including the liver, intestine, and bladder. The effects of aromatic amine exposure are species dependent. Activation through metabolic change is an essential step in aromatic amine carcinogenesis. The most prominent theory involves N-hydroxylation of aromatic amines.[2-6] Conversion to *N*-hydroxy compounds appears to be a requirement for carcinogenic activity. However, not all *N*-hydroxy compounds are carcinogenic. Although N-hydroxylation seems to be necessary for carcinogenic activity, it is not sufficient to explain the effect. The "para principle" involves the proposition that an amine should have a long, uninterrupted conjugated system with the amino group attached to one of the para carbons.[7-9] The presence of such a conjugated system lends stability to the ultimate reacting species. However, these theories do not explain all available biological data, and they are still under development. The carcinogenic activity of aromatic amines and some structure–activity relationships have been presented and discussed in several books and reviews.[10-13]

The present work involves the computer-assisted study of a set of 157 aromatic amines using chemical structure information handling and pattern-recognition methods to attempt to develop structure–activity relationships. The methodology of this approach to SAR studies has been described elsewhere and will be summarized here.

Methodology

The fundamental premises involved in applying pattern-recognition methods to SAR studies of chemicals with genetic toxicity are as follows: (1) Molecular structure and biological activity (genetic toxicity) are related. (2) The structures of compounds having genetic toxicity and compounds of similar structural classes that are nontoxic can be adequately represented by a set of molecular structure descriptors. (3) A relationship can be discovered between the structure and activity by applying statistical and/or pattern-recognition methods to a set of tested compounds. (4) The relation can be extrapolated to untested compounds to provide predictive ability.

The structure–activity studies were done using the ADAPT (automatic data analysis using pattern-recognition techniques) computer software system. This system has been developed over the past few years and has been described previously.[14]

The fundamental steps involved in performing an SAR study using the ADAPT system are as follows: (a) Identify, assemble, input, store, and describe a data set of structures for chemicals that have been tested for the biological activity of interest. (b) Develop computer-generated molecular structure descriptors for each of the members of the data set. The descriptors may be derived directly from the stored topological representations of

(1) L. Rehn, *Arch. Klin. Chir.*, **50**, 588 (1895).
(2) J. L. Radomski and E. Brill, *Science*, **167**, 992 (1970).
(3) J. L. Radomski and E. Brill, *Arch. Toxikol.*, **28**, 159 (1971).
(4) G. M. Conzelman, Jr., A. A. Rey, and E. Brill, *J. Natl. Cancer Inst.*, **50**, 989 (1973).
(5) F. F. Kadlubar, J. A. Miller, and E. C. Miller, *Cancer Res.*, **37**, 805 (1977).
(6) E. K. Weisburger, *Annu. Rev. Pharmacol. Toxicol.*, **18**, 395 (1978).
(7) H. Druckrey, *Arzneim. Forsch.*, **2**, 503 (1952).
(8) N. P. Buu-Hoi, *Arzneim. Forsch.*, **4**, 531 (1954).
(9) J. C. Arcos and M. F. Argus, "Chemical Induction of Cancer", Volume 2B, Academic Press, New York, 1974.
(10) J. C. Arcos and M. Arcos, in "Progress in Drug Research", E. Jucker, Ed., Wiley-Interscience, New York, 1962.
(11) W. C. Hueper and W. D. Conway, "Chemical Carcinogenesis and Cancers", Charles C. Thomas, Springfield, Ill., 1964.
(12) D. B. Clayson and R. C. Garner, in "Chemical Carcinogens", C. W. Searle, Ed., American Chemical Society, Washington, D.C., 1976.
(13) J. L. Radomski, *Annu. Rev. Pharmacol. Toxicol.*, **19**, 129 (1979).
(14) A. J. Stuper, W. E. Brugger, and P. C. Jurs, "Computer Assisted Studies of Chemical Structure and Biological Function", Wiley-Interscience, New York, 1979.

0022-2623/81/1824-0241$01.00/0

the structures or they may require the development of three-dimensional molecular models. (c) Using pattern-recognition methods, develop classifiers to discriminate between active and inactive compounds based on the sets of molecular descriptors. (d) Systematically reduce the set of molecular structure descriptors employed to the minimum set sufficient to retain discrimination between the active and inactive compounds. (e) Test the predictive ability of these discriminants on compounds of unknown activity.

Descriptor Generation. The heart of SAR studies lies in the development of molecular structure descriptors. One of the major premises of the approach is that one can find an "adequate" set of descriptors to represent the compounds of interest. The classes of descriptors employed in this work are topological (those derived from the connection table) and geometrical (those derived from the three-dimensional model of the molecule). The descriptors used were from the following classes.

Fragment Descriptors. These include counts of the number of atoms of each type, the number of bonds of each type, the molecular weight, the number of basis rings, and the number of ring atoms.

Substructure Descriptors. Each of the structures comprising a set of compounds under study is searched for the presence of the substructure of interest. If it is present, then the number of occurrences is computed. If not, then the descriptor is given the value of zero. The substructures to be used are dependent on the problem under investigation, and they must be found through the application of experience by the researcher.

Molecular Connectivity Descriptors. The molecular connectivity[15] of a molecule is a measure of the degree of overall branching of the structure. The path 1 molecular connectivity is formed by summing contributions for each bond in the structure, where the contribution of each bond is determined by the connectivity of the atoms that are joined by that bond. Higher order molecular connectivities can also be computed by considering all paths of length 2, 3, etc. These descriptors have been shown in several published reports[16] to be correlated with a number of physicochemical parameters, such as partition coefficients and steric parameters.

Environment Descriptors. The information present in the fragment and substructure descriptors indicates the components of the molecular structure. However, the manner of interconnection is missing. Environment descriptors supply information about the connections by coding the immediate surroundings of substructures. To generate an environment descriptor, the molecule being coded is searched for the presence of the substructure fragment that forms the heart of the environment being sought. If no match is found, the descriptor is given the value of zero. If the substructure is found, then the descriptor is computed by performing a path 1 molecular connectivity calculation on the atoms comprising the substructure, as imbedded within the structure, and, in addition, the first nearest-neighbor atoms. Thus, the value of the path 1 molecular connectivity represents the immediate surroundings as imbedded within the molecule being coded.

Geometric Descriptors. Given a three-dimensional model of the structures being coded, one can calculate descriptors designed to represent the shape of the molecules. The three principal moments of inertia and their ratios and the molecular volume have been used in this work.

Pattern-Recognition Analysis. Once each compound in a data set has been represented by a set of molecular structure descriptors, then the analysis phase of the SAR study begins. The object of the analysis phase is to find discriminants that separate subsets of the data into the proper categories. That is, one is trying to find mathematical discriminants that will classify compounds as belonging to the toxic or nontoxic compound subset based on the molecular structure descriptors available. This phase of SAR studies must be guided by the user in a highly interactive manner in order to search through the available descriptors for the best set.

From the mathematical point of view, each compound in the set under investigation is represented by a point in an n-dimensional space. For a particular compound, which is represented by a particular point, the value of each coordinate is the numerical value of one of the molecular structure descriptors comprising the representation. The expectation is that the points representing compounds of common biological activity (e.g., toxic compounds) will lie in one limited region of the space, while the points representing the compounds of different biological activity (e.g., nontoxic compounds) will be found elsewhere in the space. Pattern recognition[17,18] consists of a set of methods for investigating data represented in this manner to assess the degree of clustering and general structure of the data space.

Parametric methods of pattern recognition attempt to find classification surfaces or clustering definitions based on statistical properties of the members of one or both classes of points. Examples of parametric classifiers include Bayesian discriminants[18] and discriminants developed by linear discriminant function analysis procedures such as that used by BMD04M.[19] Nonparametric methods attempt to find discriminants by using the data themselves directly, without computing statistical measures. Examples of nonparametric methods include error-correction feedback linear learning machines[17] (threshold logic units or perceptrons) and simplex optimization methods[20] of searching for discriminants.

The pattern-recognition methods discussed here develop discriminants that separate compounds drawn from two classes from one another. Geometrically, a linear discriminant can be thought of as a hyperplane dividing the space containing the data points into two regions. The objective is to place such a linear discriminant in the space so as to completely separate the points representing toxic compounds from the points representing nontoxic compounds. Such linear discriminants are *not* models for either the active compounds alone or the inactive compounds alone. The discriminant developed is dependent on the identities of the compounds of both the active and the inactive classes. Furthermore, the compounds need only be tagged as to category (toxic vs. nontoxic) rather than having a quantitative dependent variable such as LD_{50}. This type of study based on linear discriminants should not be confused with the results generated by multiple linear regression analysis where a quantitative model is developed in order to model the degree of activity of a set of compounds.

Once discriminants have been found that do separate the data set into the appropriate subsets, then these discriminants can be validated or checked for internal consistency. This is usually done by a round-robin procedure involving leaving out a small number of data set members to act as "unknown" compounds. A much more stringent and realistic test of such a discriminant would involve the prediction of true unknowns.

The final output of this type of SAR study is the identity of the descriptors shown to be correlated with the biological activity of interest and the discriminants developed. Study of these can lead to further insights into the biological activity of interest.

Application to SAR Studies. The methodology described here has been applied to a variety of structure–activity relationship studies.[14,21] Several studies have appeared in which these methods were used to study sets of genotoxic compounds, including a heterogeneous data set,[22] polycyclic aromatic hydrocarbons,[23,24]

(15) M. Randic, *J. Am. Chem. Soc.*, **97**, 6609 (1975).
(16) L. B. Kier and L. H. Hall, "Molecular Connectivity in Chemistry and Drug Research", Academic Press, New York, 1976.
(17) N. J. Nilsson, "Learning Machines", McGraw-Hill, New York, 1965.
(18) J. T. Tou and R. C. Gonzalez, "Pattern Recognition Principles", Addison-Wesley, Reading, Mass., 1974.
(19) W. J. Dixon, Ed., "BMD Biomedical Computer Programs", University of California Press, Berkeley, Calif., 1973.
(20) G. L. Ritter, S. R. Lowry, C. L. Wilkins, and T. L. Isenhour, *Anal. Chem.*, 47, 1951 (1975).
(21) G. L. Kirschner and B. R. Kowalski, In "Drug Design", Vol. VIII, E. J. Ariens, Ed., Academic Press, New York, 1979, p 73.
(22) P. C. Jurs, J. T. Chou, and M. Yuan, *J. Med. Chem.*, **22**, 476 (1979).
(23) M. Yuan and P. C. Jurs, *Toxicol. Appl. Pharmacol.*, **52**, 294 (1980).
(24) B. Norden, U. Edlund, and S. Wold, *Acta Chem. Scand., Ser. B*, **32**, 602 (1978).

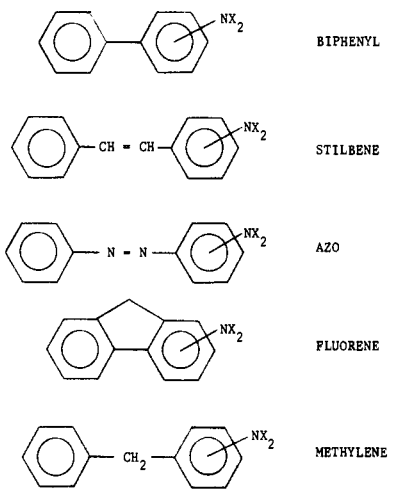


Figure 1. Basic structures of aromatic amines.

N-nitroso compounds,[25,26] and 4-nitroquinoline 1-oxides.[27]

The Data Set. The aromatic amines used in this study were taken from a published compilation of tested compounds[12] and a second compilation was used for confirmation.[28] A complete list of the 157 compounds employed is given in Table I. To be selected for inclusion in the data set a compound was required to have biological activity data reported for at least three organ sites, including the breast, had to be an aromatic amine (nitro aromatics and nitroso aromatics lacking amine moieties were excluded), and had to belong to a common structural class. A small number of compounds that appear in the published compilation were excluded from this study because of obscure or ambiguous nomenclature. However, this selection was all completed prior to the pattern-recognition studies, and no further selection was done after the studies had begun. The five structural classes present in this data set are characterized in Figure 1. All the test results were for rats. The route of administration was specified with one of two labels: oral or other. The site of action was specified by one of five labels: breast, earduct, liver, other, or all sites. Each compound was considered to be either active or inactive. In addition to the names of all the compounds studied, Table I shows the route of administration and the active site or sites. The final column of Table I contains the designation for each compound for a final set of studies in which only compounds that were negative in all sites or had at least three active sites were considered. Table II shows a breakdown of the set of compounds by structural class, active site, route of administration, and activity. The distribution of compounds among the numerous possible subsets of the data allowed studies to be performed using the following ten groupings: breast, oral; earduct, oral; liver, oral; other sites, oral; all sites, oral; breast, all routes; earduct, all routes; liver, all routes; other sites, all routes; all sites, all routes. Additional studies were done with a subset of the data consisting of the compounds that gave negative tests in all sites or gave positive tests in at least three sites. Each of these studies is done by selecting the proper subset of the 157 compounds. A particular compound can be a member of the active class for some studies and a member of the inactive class for other studies.

Results and Discussion

A very large number of molecular structure descriptors could have been developed for each of the 157 compounds in the data set. However, it is well known that the number of compounds forming the data set should exceed three times the number of descriptors studied in order to avoid chance separations.[18] It has also been shown[29] that the probability of chance separation is negligible if the number of compounds in the least populated activity category of the data set exceeds the number of descriptors under investigation. These two conditions have been met in all the pattern-recognition analyses performed during this work. Additionally, a descriptor was required to be present in at least 10 to 20% of the compounds forming the data set. No descriptors were included in analyses that were found by regression to be involved in multicollinear relationships with other descriptors with multiple correlation coefficients exceeding 0.9. Application of these rules led to the selection of the 31 descriptors shown in Table III as the starting set of descriptors. This set was used as the starting point in each of the individual pattern-recognition studies with the subsets of the data.

The descriptors generated fall into two classes: (a) those generated just once for the members of the data set and (b) those generated using a substructure as a starting point (environment descriptors). Descriptors in class a include the fragment descriptors (18 possible), total path counts (2), molecular connectivities (8), geometric descriptors (6), and molecular volume (1). All of these 35 class a descriptors were generated and then subjected to the 10% appearance and $R < 0.9$ rules stated above. This decreases the number of descriptors to the point where substructure environment descriptors can be included in the analysis without violating the 3:1 ratio criterion. The environment descriptors are substructure driven, and their number is therefore limited only by the user's ability to devise relevant substructures. An attempt is made to identify substructures within the compounds forming the data set that will be related to the functional groups involved in the structural classes under study, will be related to hypotheses of mode of action, or will be related to hypotheses of metabolic activation, etc. Small numbers of such environment descriptors are included in the pattern-recognition analyses at any given time to ensure that the 3:1 criterion is never approached. The total number of environment descriptors generated that were included in any pattern-recognition study was not large.

The molecular connectivity environment descriptors based on the substructures shown in Table III code important structural features of the aromatic amines. Substructures 7 through 13 in Table III include masked atoms, denoted by asterisks. During substructure searching a substructure atom that is masked will be matched against any structure atom of the same type with equal or additional bonding to the substructure atom. Thus, substructure 7 with the nitrogen masked for connectivity will be found imbedded within compounds containing either a primary, secondary, or tertiary aromatic amino group. Substructure 12 codes for para-amino biphenyls whether or not they have substitutents adjacent to the single bond joining the rings. Substructures 1 and 2 include information about aromatic primary and secondary amines, and substructure 7 is an even more general coding of nitrogen functionalities. Substructures 3 and 8 refer to ortho substitutions, and substrates 9 and 10 pick up meta and para substitutions, respectively. Substructures 4, 5, and 6 were used to distinguish between bridging groups present in the data set.

Substructures can be chosen as the basis for environment descriptors because they relate to one or more structure–activity theories. Substructures 6, 12, and 13 are related to the theory of coplanarity.[30,31] The ortho-

(25) J. T. Chou and P. C. Jurs, *J. Med. Chem.*, 22, 792 (1979).
(26) W. J. Dunn III and S. Wold, *Bioorg. Chem.*, in press.
(27) W. J. Dunn III and S. Wold, *J. Med. Chem.*, 21, 1001 (1978).
(28) "Survey of Compounds Which Have Been Tested for Carcinogenic Activity", U.S. Department of Health, Education, and Welfare, U.S. Public Health Service, National Institutes of Health, 1968–1969.
(29) E. K. Whalen-Pedersen and P. C. Jurs, *J. Chem. Inf. Comput. Sci.*, 19, 264 (1979).
(30) D. B. Clayson, *Br. J. Cancer*, 1, 460 (1953).

Table I. Aromatic Amine Compounds Studied

index no.	compounds	route of administration[a]	active sites[b]					
			breast	ear-duct	liver	other	all	mixed data[c]
1	4-biphenylacetamide	po	+	+	−	+	+	+
2	4-biphenyldimethylamine	po	+	+	−	+	+	+
3	4-biphenylacethydroxamic acid	po	+	+	−	+	+	+
4	2-fluoro-4-phenylaniline	po	+	−	−	−	+	
5	3′-fluoro-4-phenylaniline	po	+	−	−	−	+	
6	2-methyl-4-phenylaniline	po	−	+	−	+	+	
7	3-methyl-4-phenylaniline	po	−	−	−	−	−	−
8	2′-methyl-4-phenylaniline	po	−	−	−	−	−	−
9	4′-methyl-4-phenylacetanilide	po	−	−	−	−	−	−
10	3,2′-dimethyl-4-biphenylamine	po	+	+	−	+	+	+
11	*o*,*o*′-dianisidine	po	−	na	−	−	−	−
12	3,3′-dichlorobenzidine	po	+	+	−	+	+	+
13	3,3′-dihydroxybenzidine	po	−	−	−	−	−	−
14	2-methyldiacetylbenzidine	po	+	+	−	+	+	+
15	4,4′-methylenedianiline	po	−	−	na	−	−	−
16	4,4′-methylenebis(2-methylaniline)	po	+	−	+	+	+	+
17	4-*N*-stilbenamine	po	+	+	−	−	+	
18	*N*-hydroxy-4,*N*-stilbenylacetamide	po	+	+	−	+	+	+
19	4-stylbenyl-*N*,*N*-dimethylamine	po	−	+	−	−	+	
20	*N*,*N*,2′-trimethyl-4-stilbenamine	po	−	−	−	−	−	−
21	*N*,*N*,3′-trimethyl-4-stilbenamine	po	−	+	−	−	+	
22	*N*,*N*,4′-trimethyl-4-stilbenamine	po	−	na	−	−	−	−
23	2′-fluoro-4-stilbenyl-*N*,*N*-dimethylamine	po	+	+	+	−	+	+
24	4′-chloro-4-stilbenyl-*N*,*N*-dimethylamine	po	−	+	−	−	+	
25	4′-nitro-4-stilbenyl-*N*,*N*-dimethylamine	po	−	na	−	−	−	−
26	9-oxo-2-fluorenylacetamide	po	+	na	+	−	+	
27	2-fluorenyldimethylamine	po	+	+	+	+	+	+
28	2-fluorenyldiethylamine	po	−	−	−	+	+	
29	2-fluorenylmethylamine	po	+	+	+	+	+	+
30	2-fluorenyldiacetamide	po	+	+	+	+	+	+
31	*N*,2-fluorenylformamide	po	+	+	−	−	+	
32	*N*,2-fluorenylsuccinamic acid	po	−	−	+	−	+	
33	*N*-2-fluorenyl-*p*-toluenesulfonamide	po	−	−	−	−	−	−
34	*N*-(2-fluorenyl)-2,2,2-trifluoroacetamide	po	+	+	+	−	+	+
35	*N*-3-glycylaminofluorene	po	−	−	+	−	+	
36	1-fluoro-2-fluorenylacetamide	po	+	+	−	−	+	
37	3-fluoro-2-fluorenylacetamide	po	+	+	+	−	+	+
38	4-fluoro-2-fluorenylacetamide	po	+	+	−	−	+	
39	5-fluoro-2-fluorenylacetamide	po	+	+	−	−	+	
40	6-fluoro-2-fluorenylacetamide	po	+	+	+	+	+	+
41	7-fluoro-2-fluorenylacetamide	po	+	+	+	−	+	+
42	8-fluoro-2-fluorenylacetamide	po	+	+	+	−	+	+
43	7-fluoro-2-*N*-fluorenylacethydroxamic acid	po	+	+	+	+	+	+
44	7-chloro-2-fluorenylacetamide	po	−	−	na	−	−	−
45	3-iodo-2-fluorenylacetamide	po	−	+	−	−	+	
46	7-iodo-2-fluorenylacetamide	po	−	−	−	−	−	−
47	1-methoxy-2-fluorenylamine	po	−	na	+	+	+	
48	3-methoxy-2-fluorenylamine	po	−	−	−	−	−	−
49	3-methoxy-2-fluorenylacetamide	po	−	−	−	−	−	−
50	7-methoxy-2-fluorenylacetamide	po	+	+	+	−	+	+
51	2,7-diaminofluorene	po	+	−	−	−	+	
52	2,5-fluorenylenebisacetamide	po	+	−	−	−	+	
53	3-methyl-2-naphthylamine	po	+	−	−	+	+	
54	3-nitro-2-naphthylamine	po	−	na	−	+	+	
55	9,10-dihydro-2-phenanthramine	po	+	+	−	+	+	+
56	4-(phenylazo)aniline	po	−	−	−	+	+	
57	4-(phenylazo)acetanilide	po	−	−	−	−	−	−
58	4-(phenylazo)diacetanilide	po	−	−	−	−	−	−
59	4-(phenylazo)-*N*-phenylacethydroxamic acid	po	−	−	−	−	−	−
60	4-(phenylazo)-*o*-anisidine	po	−	+	+	+	+	+
61	4-[(*p*-methoxyphenyl)azo]-*o*-anisidine	po	−	na	+	+	+	
62	4-(*m*-tolylazo)aniline	po	−	−	−	−	−	−
63	4-(*m*-tolylazo)acetanilide	po	−	−	−	−	−	−
64	4′-fluoro-4-stilbenyl-*N*,*N*-dimethylamine	po	−	+	−	−	+	
65	4-(*o*-tolylazo)-*m*-toluidine	po	−	−	−	−	−	−
66	4-(*o*-tolylazoxy)-*o*-toluidine	po	−	−	−	−	−	−
67	4′-fluoro-*p*-phenylaniline	po	−	−	−	−	−	−
68	1-(phenylazo)-2-naphthylamine	po	−	−	−	−	−	−
69	5-acetamido-3-(5-nitro-2-furyl)-6*H*-1,2,4-oxadiazine	po	−	−	+	+	+	
70	5-nitro-2-furamidoxime	po	−	−	−	−	−	−
71	4,6-diamino-2-(5-nitro-2-furyl)-*s*-triazine	po	+	−	−	−	+	
72	*N*,*N*′-[6-(5-nitro-2-furyl)-*s*-triazine-2,4-diyl]bisacetamide	po	+	−	−	−	+	

Table I (*Continued*)

index no.	compounds	route of administration[a]	active sites[b] breast	ear-duct	liver	other	all	mixed data[c]
73	2-hydrazino-4-phenylthiazole	po	–	–	–	–	–	–
74	3-amino-*s*-triazole	po	–	–	+	+	+	
75	3-carbazolylacetamide	po	–	–	–	–	–	–
76	3-dibenzofuranylacetamide	po	+	+	–	–	+	
77	2-dibenzothiophenylacetamide	po	+	+	–	+	+	+
78	3-dibenzothiophenylacetamide	po	+	+	–	+	+	+
79	2-methoxy-3-benzofuranylamine	po	+	+	–	+	+	+
80	4-biphenylamine	po	+	–	na	–	+	
81	4,4′-methylenebis(2-chloroaniline)	po	–	–	+	+	+	
82	suramine	po	–	–	+	+	+	
83	4-*N*-stilbenylacetamide	po	+	+	–	–	+	
84	2-[4-(*N*,*N*-dimethylamino)styryl]quinoline	po	–	+	–	–	+	
85	2,7-fluorenylbisacetamide	po	+	+	+	+	+	+
86	1-naphthylhydroxylamine	po	–	–	–	+	+	
87	2-naphthylamine	po	–	–	–	–	–	–
88	1-phenanthrylacetamide	po	–	–	–	–	–	–
89	9-phenanthrylamine	po	+	–	–	–	+	
90	3-phenanthrylamine	po	+	–	–	–	+	
91	2-phenanthrylamine	po	+	–	–	–	+	
92	1-phenanthrylamine	po	+	–	–	–	+	
93	9-phenanthrylacetamide	po	–	–	–	–	–	–
94	2-phenanthrylacetamide	po	+	+	–	+	+	+
95	2-(*p*-tolylazo)-*p*-toluidine	po	–	–	–	–	–	–
96	4-(*p*-tolylazo)-*o*-toluidine	po	–	–	–	–	–	–
97	4-(*p*-tolylazo)-*m*-toluidine	po	–	–	–	–	–	–
98	4-(*m*-tolylazo)-*m*-toluidine	po	–	–	–	–	–	–
99	2-(*o*-tolylazo)-*p*-toluidine	po	–	–	+	–	+	
100	*N*-[4-(5-nitro-2-furyl)-2-thiazolyl]formamide	po	+	–	–	+	+	
101	2-hydrazino-4-(4-nitrophenyl)thiazole	po	+	–	–	–	+	
102	2-hydrazino-4-(5-nitro-2-furyl)thiazole	po	+	–	–	–	+	
103	formic acid, 2-[4-(5-nitro-2-furyl)-2-thiazolyl]hydrazide	po	+	–	na	+	+	
104	2-(2,2-dimethylhydrazino)-4-(5-nitro-2-furyl)thiazole	po	+	–	–	+	+	
105	4-(hydroxyamino)quinoline 1-oxide	po	+	–	–	+	+	
106	1-(*o*-tolylazo)-2-naphthylamine	po	–	–	–	–	–	–
107	6-[(1-methyl-4-nitroimidazol-5-yl)thio]purine	po	–	+	–	+	+	
108	3,6-bis(dimethylamino)acridine	po	–	–	+	–	+	
109	2-chloro-4-phenylaniline	sc	–	–	–	–	–	–
110	4′-fluoro-4-biphenylamine	sc	–	–	+	+	+	
111	3,3′-dimethyl-4-biphenylamine	sc	–	–	–	+	+	
112	3,2′,5′-trimethyl-4-biphenylamine	sc	–	–	+	+	+	
113	3,2′,4′,6′-tetramethyl-4-biphenylamine	sc	–	–	+	+	+	
114	3-methoxy-4-biphenylamine	sc	–	–	–	+	+	
115	benzidine	sc	–	+	+	+	+	+
116	*o*,*o*′-tolidine	sc	+	+	–	+	+	+
117	*N*-(4-styrylphenyl)hydroxylamine	sc	–	na	–	+	+	
118	*N*-acetoxy-4-*N*-stilbenylacetamide	sc	–	+	–	+	+	
119	4-stilbenyl-*N*,*N*-diethylamine	sc	–	+	–	–	+	
120	2-methyl-4-stilbenamine	sc	–	–	–	+	+	
121	3-methyl-4-stilbenamine	sc	–	–	–	+	+	
122	4-(2,5-dimethoxy)stilbenamine	sc	–	+	–	–	+	
123	4′-fluoro-4-stilbenamine	sc	–	+	+	+	+	+
124	4,4′-diaminostilbene	sc	–	–	+	–	+	
125	3,3′-dichloro-4,4′-diaminostilbene	sc	–	–	–	+	+	
126	2,2′-dichloro-4,4′-diaminostilbene	sc	–	–	+	–	+	
127	2-cyano-4-stilbenamine	sc	–	–	+	+	+	
128	2-fluorenamine	vr	+	+	+	+	+	+
129	2-fluorenylacetamide	vr	+	+	+	+	+	+
130	*N*-acetoxyfluorenylacetamide	sc	–	na	–	+	+	
131	*N*-(benzoyloxy)fluorenylacetamide	sc	–	na	–	+	+	
132	fluorenyl-2-acethydroxamic acid	vr	+	+	+	+	+	+
133	2-fluorenylhydroxylamine	sc	+	+	–	+	+	+
134	*N*-fluorenyl-2-benzamide	ip	–	–	–	–	–	–
135	*N*-fluorenyl-2-benzohydroxamic acid	ip	+	–	–	+	+	
136	*N*-hydroxy-*N*-fluorenylbenzenesulfonamide	ip	+	–	–	+	+	
137	*N*-fluorenyl-2-benzenesulfonamide	ip	–	–	–	–	–	–
138	1-fluorenylacetamide	ip	+	–	–	–	+	
139	1-fluorenylacethydroxamic acid	ip	+	–	–	–	+	
140	3-fluorenylacethydroxamic acid	ip	+	–	–	–	+	
141	3-fluorenylacetamide	ip	+	–	–	–	+	
142	1-naphthylacethydroxamic acid	ip	–	–	–	–	–	–
143	2-naphthylhydroxylamine	ip	–	–	–	+	+	
144	1-anthramine	tp	–	–	–	–	–	–

Table I (*Continued*)

index no.	compounds	route of administration[a]	active sites[b] breast	ear-duct	liver	other	all	mixed data[c]
145	2-anthramine	tp	+	−	−	+	+	
146	9-anthramine	sc	−	−	−	−	−	−
147	2-anthranylacetamide	sc	−	−	−	−	−	−
148	2-phenanthrylacethydroxamic acid	sc	+	−	−	+	+	
149	*N*-acetoxy-4-phenanthrylacetamide	sc	−	−	−	+	+	
150	4-(phenylazo)-*N*-phenylhydroxylamine	sc	−	−	−	−	−	−
151	4-(*o*-tolylazo)-*o*-toluidine	na	−	−	+	−	+	
152	4-aminoquinoline 1-oxide	sc	−	−	−	−	−	−
153	3-(hydroxyamino)quinoline 1-oxide	sc	−	−	−	−	−	−
154	5-(hydroxyamino)quinoline 1-oxide	sc	−	−	−	−	−	−
155	4-(hydroxyamino)pyridine 1-oxide	sc	−	−	−	−	−	−
156	*N*-[4-(5-nitro-2-furyl)-2-thiazolyl]acetamide	na	+	−	−	+	+	
157	*N*-acetoxy-4-biphenylacetamide	sc	−	−	−	+	+	

[a] Abbreviations used: po, oral; sc, subcutaneous injection; tp, topical; ip, intraperitoneal injection; na, not available; vr, various routes. [b] +, carcinogen; −, noncarcinogen; na, data not available. [c] A plus sign appears for compounds with either four or five active sites; a minus sign appears for completely inactive compounds.

Table II. Distribution of the Aromatic Amine Compound Data Set among Structural Subsets

structure classes	active site: administration:	breast po	breast other	breast all	earduct po	earduct other	earduct all	liver po	liver other	liver all	other sites po	other sites other	other sites all	all sites po	all sites other	all sites all	total
biphenyl	act.	9	1	10	7	2	9	0	4	4	7	7	14	10	7	17	54
	not act.	7	7	14	8	6	14	15	4	19	9	1	10	6	1	7	64
stilbene	act.	4	0	4	7	4	11	1	4	5	1	7	8	7	11	18	46
	not act.	6	11	17	1	6	7	9	7	16	9	4	13	3	0	3	56
azo	act.	0	0	0	2	0	2	3	1	4	3	0	3	5	1	6	15
	not act.	16	2	18	13	2	15	13	1	14	13	2	15	11	1	12	74
fluorene	act.	18	10	28	16	4	20	15	3	18	8	8	16	23	12	35	117
	not act.	10	4	14	10	8	18	13	11	24	20	6	26	5	2	7	89
methylene	act.	1	0	1	0	0	0	3	0	3	3	0	3	3	0	3	10
	not act.	3	0	3	4	0	4	0	0	0	1	0	1	1	0	1	9
misc	act.	19	3	22	8	0	8	3	0	3	16	6	22	26	6	32	87
	not act.	15	11	26	25	14	39	29	14	43	18	8	26	8	8	16	150
total		108	49	157	101	46	147	104	49	153	108	49	157	108	49	157	771

hydroxylation theory[32] led us to include substructures 3 and 8. Many other descriptors which were expected to support SAR theories were generated, but the various criteria of appearance times, correlation coefficients, and the 3:1 criterion precluded thier use. Descriptors which were thought to support the most dominant N-hydroxylation theory[2-6] were deleted because of their low number of appearances. Descriptors that were based on the theory of para principles[7-9] were dropped because of high correlation coefficients. To use these descriptors would require a larger data set.

Table IV gives the mean and standard deviation for each of the 31 descriptors for the entire 157 compound data set. The final two columns give the values for compound 13, 3,3′-dihydroxybenzidine (a compound negative in all tests) and compared 40, 6-fluoro-2-fluorenylacetamide (a compound positive in all tests) as examples of descriptor values.

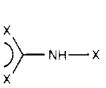
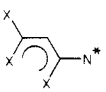
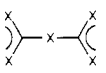
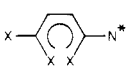
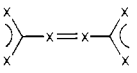
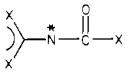
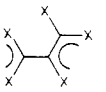
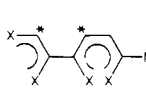
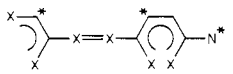
Using these 31 molecular structure descriptors as the starting set and using the pattern-recognition methods and feature selection methods previously described, the 11 studies were done with the 11 subsets of the data. In each study a slightly different set of final descriptors was found to be the best minimal subset. The identities of the 11 final subsets of the 31 descriptors are shown in Table V. The number of descriptors forming the best subset for an individual problem ranges from a low of 12 (mixed) to a high of 23 (other sites, all routes) with a mean number of 18.4.

For each of the studies, a selection of pattern-recognition methods was applied. We do not report the individual results but rather summarize them here for brevity. The methods used included (a) Bayesian quadratic discriminant, (b) Bayesian linear discriminant, (c) *K* nearest-neighbor classification, (d) an iterative least-squares linear discriminant development routine, (e) a simplex algorithm for development of linear discriminants, and (f) linear learning machine. All of these methods have been described in detail previously.[14] In general, the KNN method was the least successful for this study, having classification success in the range of 60 to 70%. The Bayesian methods depend on assuming that the data can be well represented by the covariance matrix. Many of the descriptors used in this study take on a few discrete values and are therefore multimodal, making the covariance matrix a poor representation of the data. Therefore, the results from Bayesian discriminants were used only for comparison. The Bayesian quadratic discriminants were consistently inferior in classification ability to the linear discriminants; this demonstrates that the data do not consist of a cluster of points corresponding to the active compounds surrounded by a random scattering of points corresponding to inactive compounds. Of the linear discriminant development

(31) A. L. Walpole, M. H. C. Williams, and D. C. Roberts, *Br. J. Ind. Med.*, 9, 255 (1952).
(32) E. C. Miller, R. B. Sandin, J. A. Miller, and H. P. Rusch, *Cancer Res.*, 16, 525 (1956).

Table III. Set of 31 Starting Descriptors[a]

descr no.	descriptor identity
1	no. of carbon atoms
2	no. of oxygen atoms
3	no. of nitrogen atoms
4	no. of single bonds
5	no. of double bonds
6	no. of aromatic bonds
7	no. of basis rings
8	no. of ring atoms
9	path 3 molecular connectivity
10	path 4 molecular connectivity
11	total no. of paths
12	molecular connectivity environment: SS1
13	molecular connectivity environment: SS2
14	molecular connectivity environment: SS3
15	molecular connectivity environment: SS4
16	molecular connectivity environment: SS5
17	molecular connectivity environment: SS6
18	largest principal moment
19	intermediate principal moment
20	smallest principal moment
21	ratio of largest to smallest principal moment
22	ratio of intermediate to smallest principal moment
23	molecular connectivity environment: SS7
24	molecular connectivity environment: SS8
25	molecular connectivity environment: SS9
26	molecular connectivity environment: SS10
27	molecular connectivity environment: SS11
28	molecular connectivity environment: SS12
29	molecular connectivity environment: SS13
30	molecular volume
31	no. of F + Cl + I + S

SS no.	substructure	SS no.	substructure
1		7	
2		8	
3		9	
4		10	
5		11	
6		12	
		13	

[a] SS = substructure.

routines, the iterative least-squares program enjoyed the most success while attempting to classify the entire data sets—classification success rates in the neighborhood of 90% were attained for most of the subsets. This iterative least-squares routine was used to identify the minimal subset of compounds that had to be neglected in order to develop completely separable subsets. These linearly separable subsets were then studied using linear learning machines.

In the study of each data subset some compounds could not be correctly classified even when the entire set of 31 descriptors was used. A reasonable compromise was sought between the maximum number of correct classifications and the minimum number of descriptors in each final set. Table VI shows the number of compounds that had to be deleted from each subset of the data in order to proceed with the study. Set A for each study contains the entire subset of the data relevant to that study. Set B contains the subset of the data that remained at the end of each study. The numbers of excluded compounds on the right side of Table VI show the absolute number of compounds excluded from each study. In percentage terms the number of compounds ranged from a low of 6.5% to a high of 14.6% with an average of 9.3%. The data set with the labels (other sites, all routes) was the most difficult set to study. The number of excluded compounds (23) was larger than in the other studies, and convergence was slow during pattern-recognition experiments. These facts may suggest that this subset of the data would be better described by some other set of descriptors or an expanded set of descriptors or that this activity class is not comparible to the other subsets due to possible biological differences.

Table VII presents the compound numbers for those compounds excluded from each study. Figure 2 shows the structures of the five compounds that were most often excluded. The most often deleted compound is 2-phenanthrylacetamide (94). This compound was deleted in many subset studies except those involving the liver as the active site. This suggests that this compound is not well represented by the starting 31 descriptors or that it may be subject to special metabolic effects in the rat liver. This would have to be studied further by biological testing.

Investigation of the trends of descriptor appearances in Table V provides some clues as to important characteristics of aromatic amines. Descriptor 7, the number of basis rings, was retained in all but one of the individual studies.

Table IV. Descriptor Statistics and Sample Descriptor Values

descr no.	mean	SD	value for 13	value for 40
1	13.943	3.007	12.00	15.00
2	1.070	1.246	2.00	1.00
3	1.732	1.242	2.00	1.00
4	5.624	2.943	5.00	7.00
5	1.166	1.501	0.00	1.00
6	11.662	3.565	12.00	12.00
7	2.439	0.603	2.00	3.00
8	12.465	2.074	12.00	13.00
9	2.706	1.321	2.394	2.981
10	9.589	0.999	8.034	10.85
11	572.497	372.805	349.0	839.0
12	0.798	0.700	2.207	0.000
13	0.901	3.041	0.00	2.874
14	1.769	1.626	2.282	3.545
15	1.165	1.763	0.0	3.794
16	1.074	1.788	0.0	0.0
17	1.384	1.975	0.0	4.146
18	11.071	4.316	10.50	11.31
19	1.419	0.570	1.382	1.339
20	0.105	0.271	0.8048E-6	0.1684
21	630.920	429.776	1000.0[a]	67.16
22	516.415	485.492	1000.0[a]	7.947
23	2.375	0.799	2.207	2.874
24	0.801	1.372	2.282	0.0
25	1.658	2.134	0.0	4.652
26	2.725	1.906	3.593	4.658
27	0.997	1.412	0.0	2.724
28	2.230	2.731	4.826	6.006
29	1.333	2.380	0.0	0.0
30	75.013	13.724	69.77	76.74
31	0.274	0.583	0.0	1.00

[a] These values are used to code the ratios of principal moments for flat compounds.

Table V. Descriptors Remaining in Each Study of a Subset of the Aromatic Amines

descr no.	data sets											appearance in mixed administration	appearance in oral administration	total
	BO	EO	LO	OO	AO	B	E	L	O	A	M			
1			*b*					*b*	X		*c*	1	0	1
2				X	X	X	X	X	X	X	*c*	5	2	7
3		X	X	X	X	X	X	X	X	X		5	4	9
4			X	X		X	X	X		X	X	4	2	7
5			X				X	X	X	X		4	1	5
6		X	*b*	X	X		X		X	X	*c*	3	3	6
7	X	X	X		X	X	X	X	X	X	X	5	4	10
8	X	X	X	X		X	X	X	X			4	4	8
9	X	X	X			X	X	X	X	X		5	3	8
10				X				X	X	X	X	3	1	5
11	X		*b*	X		X	X		X	X		4	2	6
12	X		X		X		X	X	X		X	3	3	7
13	X	X		X	X			X	X	X	X	3	4	8
14	X	X	X	X	X	X		X	X			3	5	8
15		X	X		X		X	X	X	X		4	3	7
16			*b*	X		X		*b*			*c*	1	1	2
17	X			X		X		X	X	X		4	2	6
18		X		X	X	X	X	X	X		X	4	3	8
19	X		X	X			X	X	X	X	X	4	3	8
20	X		*b*	X	X	X	X		X	X		4	3	7
21	X	X	X	X		X	X	X		X	X	4	4	9
22	X		*b*	X	X	X				X	X	2	3	6
23	X	X	*b*	X	X	X	X			X		3	4	7
24	X		X	X		X	X	X	X	X		5	3	8
25			*b*	X	X			*b*			*c*	0	2	2
26	X	X	X		X			X	X	X	X	3	4	8
27			X	X	X	X		X	X		*c*	3	3	6
28		X	*b*				X		X		*c*	2	1	3
29	X	X	*b*		X		X		X	X	X	3	3	7
30	X	X			X	X		X				2	3	5
31	X	X	X		X	X		X		X	X	3	4	8
no. in final set	18	16	15	20	18	19	19	21	23	21	12			

[a] Data sets: BO, breast and oral; EO, earduct and oral; LO, liver and oral; OO, other sites and oral; AO, all sites and oral; B, breast and all routes; E, earduct and all routes; L, liver and all routes; O, other sites and all routes; A, all sites and all routes; M, mixed. [b] Deleted from the starting descriptor set because of population distribution of the data. [c] Deleted in order to keep total number of descriptors less than one-third the number of compounds.

Table VI. Compositions of Subsets of the Data Used for Individual Studies

	set A			set B			excluded compds		
	act.	not act.	total	act.	not act.	total	act.	not act.	total
breast, oral	51	57	108	47	51	98	4	6	10
earduct, oral	40	61	101	36	58	94	4	3	7
liver, oral	25	79	104	21	75	96	4	4	8
other sites, oral	37	71	108	31	69	100	6	2	8
all sites, oral	74	34	108	70	31	101	4	3	7
breast, all routes	65	92	157	57	86	143	8	6	14
earduct, all routes	50	97	147	43	90	133	7	7	14
liver, all routes	37	116	153	30	110	143	7	6	13
other sites, all routes	65	92	157	55	79	134	10	13	23
all sites, all routes	111	46	157	106	34	140	5	12	17

This suggests that the number of rings (related to molecular volume or bulk) is an important descriptor relating aromatic amino structure to carcinogenic potential. This is consistent with recent reports by Cain and co-workers[33] showing that 9-aminoacridine derivatives with two rings bind more strongly to DNA than 4-aminoquinoline derivatives with one ring. In a previous SAR study of polycyclic aromatic hydrocarbons,[23] the number of rings and other size information were shown to be important information relating structure to carcinogenic potential.

Other descriptors that show a large number of appearances are numbers 2 (number of oxygen atoms), 12 (molecular connectivity environment for substructure 1), 20 (smallest principal moment), 23 (molecular connectivity environment for substructure 7), and 29 (molecular connectivity environment for substructure 13). Descriptors 1 and 25 were found to be unimportant in these studies. The other descriptors show appearance times in the range of four to six.

We break down the set of descriptors into three types: (1) high number of appearances, (2) medium number of

(33) B. F. Cain, B. C. Baguley, and W. A. Denny, *J. Med. Chem.*, 21, 658 (1978).

Table VII. Identities of Compounds that Were Excluded from the Individual Studies

breast, oral	act.	17, 53, 83, 94
	not act.	32, 67, 73, 86, 87, 107
earduct, oral	act.	12, 14, 85, 94
	not act.	4, 20, 28
liver, oral	act.	23, 60, 69, 99
	not act.	28, 39, 46, 51
other sites, oral	act.	18, 40, 56, 60, 61, 94
	not act.	87, 102
all sites, oral	act.	10, 60, 61, 83
	not act.	22, 67, 87
breast, all routes	act.	17, 30, 52, 80, 94, 105, 136, 148
	not act.	32, 58, 69, 73, 107, 157
earduct, all routes	act.	6, 12, 60, 94, 107, 115, 116
	not act.	20, 28, 44, 49, 63, 135, 157
liver, all routes	act.	23, 37, 60, 69, 99, 108, 112
	not act.	7, 28, 51, 96, 120, 135
other sites, all routes	act.	56, 60, 61, 74, 94, 110, 115, 129, 145, 148
	not act.	4, 5, 23, 31, 32, 35, 50, 51, 95, 99, 102, 109, 137
all sites, all routes	act.	60, 61, 81, 83, 94
	not act.	20, 22, 58, 62, 67, 73, 87, 109, 142, 144, 146, 152
mixed (12 descr)	act.	94, 115, 116
	not act.	67
mixed (9 descr)	act.	12, 60, 94
	not act.	11, 13, 67, 134
mixed (14 descr)	act.	10, 60, 94
	not act.	48, 49, 67

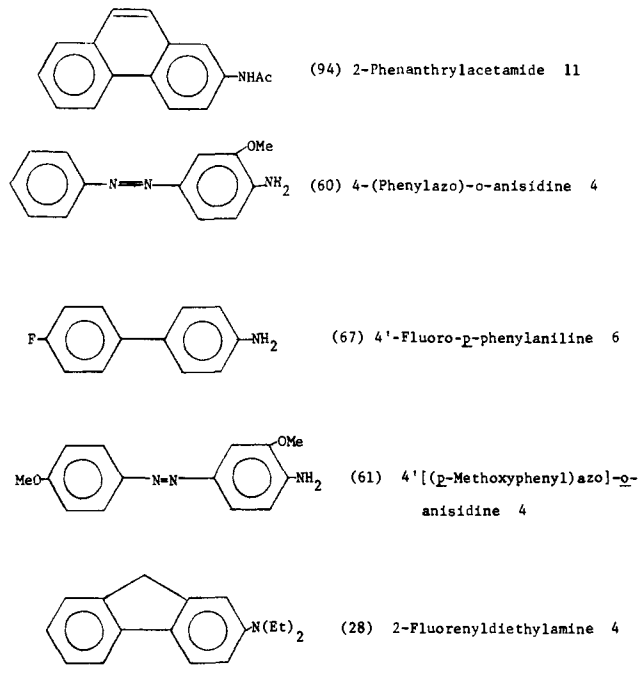


Figure 2. Structures of the compounds most often excluded from the individual studies.

appearances, and (3) low number of appearances. Type 1 descriptors are the most important for representation of the data set, and type 2 descriptors may contribute somewhat. Type 3 descriptors do not contribute.

Among the descriptors with high numbers of appearances, there are several that relate directly to size and shape of the amines. These types of considerations have been found to be important in relating carcinogenic potential to other types of compounds to structures, e.g., in a study by Hansch and Fujita.[34] While the size/shape descriptors plus the number of basis rings descriptor are important, they are insufficient to separate the active compounds from the inactives. It should be emphasized that for each of the studies reported, the set of descriptors must be considered as a *set* and not individually.

Table VIII shows three somewhat different sets of descriptors that were generated during studies of the overall, mixed data set of 79 compounds. These are the compounds that show activity in at least three sites to be considered active or show no activity in any site to be considered inactive. Descriptor set I contains a 12 descriptor subset of the 31 descriptors used for the ten studies described above. For the study of the 79 compound data set, a number of additional descriptors were generated and tested. Descriptor set II and descriptor set III shown in Table VIII contain some of these new descriptors in addition to many of those from the original pool of 31 descriptors. The pattern-recognition results obtained with these three sets of descriptors are shown in Table IX. We have shown results for a variety of pattern-recognition methods to illustrate the different methodologies that are available.

The results for internal consistency checking shown in Table X were obtained from a series of tests as follows: randomly choose two compounds from the current data set to be held out during development of the linear discriminant function; generate the discriminant function using the remaining compounds; use the discriminant to predict the activity class of the two compounds held out; average over as many sets of two compounds as there are in the data set. Each compound in the data set is predicted just once. The overall percentage correctly classified during the entire procedure is termed "predictive ability", and it is a measure of the validity of the discriminants used.

While the pattern-recognition results obtained with these three sets of descriptors are similar, the internal consistency check results shown in Table X do show some differences. These tables demonstrate that there are many potential sets of descriptors that, taken as sets, will support linear discriminant functions that can separate carcinogens from noncarcinogens. Starting with a set of descriptors containing as much relevant information as the user can devise, the feature selection steps of a pattern-recognition study allow the user to discard nonessential descriptors. This is an iterative, user-directed procedure, and it results in a final minimal subset of descriptors. The detailed steps used during feature selection will determine the final subset selected, and it may be difficult to choose among alternative subsets. The "predictive ability" results shown in Table X provide one method for choosing which descriptor subset is best, but this measure has some random fluctuation associated with it that is difficult to quantify. Nevertheless, predictive abilities in the range of 85 to 90% do show that these sets of descriptors contain structural information relating amino structures to carcinogenic potential. The discriminants developed using these methods would be able to make predictions as to the carcinogenic potential of truly unknown compounds of similar structural type. The true test of the overall veracity of the method would be in using the discriminant functions in conjunction with biological testing programs to determine the accuracy

(34) C. Hansch and T. Fujita, *J. Am. Chem. Soc.*, 86, 1616 (1964).

Table VIII. Three Sets of Descriptors Developed for the Mixed Data Set

	descriptor set I	descriptor set II	descriptor set III
1	no. of single bonds	no. of single bonds	no. of oxygen atoms
2	no. of basis rings	no. of double bonds	no. of single bonds
3	mol connectivity environment: SS 1	mol connectivity environment: SS 2	no. of double bonds
4	mol connectivity environment: SS 3	mol connectivity environment: SS 4	no. of basis rings
5	path 4 mol connectivity	path-cluster 4 mol connectivity	mol connectivity environment: SS 5
6	largest principal moment	mol connectivity environment: SS 6	path 2 mol connectivity
7	intermediate principal moment	ratio of largest to smallest principal moment	path 4 mol connectivity
8	ratio of largest to smallest principal moment	ratio of intermediate to smallest principal moment	all paths
9	ratio of intermediate to smallest principal moment	mol connectivity environment: SS 7	mol connectivity environment: SS 6
10	mol connectivity environment: SS 8		smallest principal moment
11	mol connectivity environment: SS 9		mol connectivity environment: SS 10
12	no of F + Cl + I + S		σ charge average: SS 11
13			σ charge average: SS 5
14			σ charge average: SS 12

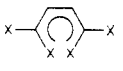
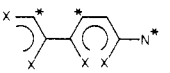
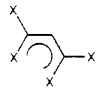
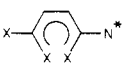
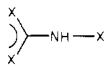
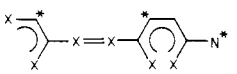
SS no.	substructure	SS no.	substructure[a]	SS no.	substructure	SS no.	substructure
1		4		7		10	
2		5		8		11	
3		6		9		12	

Table IX. Pattern-Recognition Results Obtained for the 81 Compound Aromatic Amine Data Sets

		descriptor set I: % correctly classified			descriptor set II: % correctly classified			descriptor set III: % correctly classified		
classifier	set	act.	not act.	total	act.	not act.	total	act.	not act.	total
Bayes (quadratic)	A	90.9	87.0	88.6	90.9	80.4	84.8	93.9	82.6	87.3
	B	96.7	93.3	94.7	100.0	85.7	91.7	96.7	88.4	91.8
Bayes (linear)	A	75.8	84.8	81.0	93.9	71.7	81.0	84.9	84.8	84.8
	B	83.3	91.1	88.0	96.7	85.7	90.3	93.3	93.0	93.2
KNN										
$K = 1$	A	66.7	73.9	70.9	78.8	73.9	76.0	75.8	76.1	76.0
$K = 1$	B	73.3	82.2	78.7	83.3	85.7	84.7	80.0	83.7	82.2
$K = 3$	A	66.7	82.6	76.0	81.8	78.3	79.8	72.7	78.3	76.0
$K = 3$	B	73.3	86.7	81.3	86.7	78.6	81.9	80.0	79.1	79.5
iterative least square	A	84.9	97.8	92.4	90.9	91.3	91.1	87.9	89.1	88.6
	B	96.7	97.8	97.3	100.0	97.6	98.6	100.0	100.0	100.0
simplex	A	81.8	89.0	84.8	87.9	82.6	84.8	87.9	87.0	87.3
	B	93.3	93.3	93.3	96.7	97.6	97.2	93.3	93.0	93.2
linear learning machine	B	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
set:		set A			set A			set A		
composition:		act. 33			act. 33			act. 33		
		not act. 46			not act. 46			not act. 46		
		total 79			total 79			total 79		
		set B			set B			set B		
		act. 30			act. 30			act. 30		
		not act. 45			not act. 42			not act. 43		
		total 75			total 72			total 73		

Table X. Results for Internal Consistency Checks Obtained for the 81 Compound Aromatic Amine Data Sets[a]

	descriptor set I			descriptor set II			descriptor set III		
	act.	not act.	total	act.	not act.	total	act.	not act.	total
no. predicted	30	44	74	30	42	72	30	42	72
no. correct	25	38	63	26	37	63	25	36	61
% correct	83.3	86.4	85.1	86.7	88.1	87.5	83.3	85.7	84.7

[a] Linear learning machine results by leave two out method.

of predictions of true unknowns.

Conclusions

A data set consisting of 157 aromatic amines has been studied using computer-assisted structure–activity methods. The 157 compounds were divided into subsets according to tumor sites, routes of administration, and activity. Differences in active site did not affect the results to a large extent. However, different pattern-recognition methods showed markedly different classification and predictive abilities.

In spite of the dissimilarity of the data sets' positive and negative compound distribution, there were some molecular structure descriptors that consistently were chosen as important. Prominent among these important descriptors were those coding size and shape information, e.g., number of rings and principal moments. These descriptors must be representing common factors important in the different subset studies. Final or conclusive meanings that attach to these descriptors relevant to structure–activity relationships would have to be determined by biological experiments.

These results are specific to the data sets used. They will be general to the extent that this data set mimics the universe of aromatic amino compounds. If this set of 157 compounds is a good representation of all aromatic amines, then the results should generalize. If the data set is not representative of the universal set, then the results are applicable only to the immediate compounds.

The results show that similar results can be obtained by using data of a mixed nature, that is, compounds which were tested using various protocols. It is not necessary to limit a study to one site, one route of administration, etc., in order to proceed. Of course, the results will be somewhat dependent on such factors, but mixed data sets can be used in computer-assisted SAR studies.

Acknowledgment. This research was supported by the National Cancer Institute through Contract N01 CP 75926. The computer used for this work was purchased with partial financial support of the National Science Foundation.

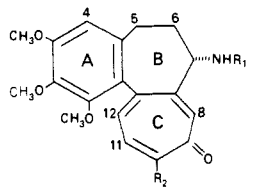
Quantitative Structure–Activity Relationships of Colchicines against P388 Leukemia in Mice

Frank R. Quinn* and John A. Beisler

Laboratory of Medicinal Chemistry and Biology, Division of Cancer Treatment, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20205. *Received January 7, 1980*

A quantitative structure–activity relationship (QSAR) was derived for colchicine and 14 analogues acting against P388 lymphocytic leukemia in mice. Twelve additional compounds were synthesized to reinforce and confirm the correlation. The final correlation indicates that there is a parabolic dependence of antitumor potency on the partition coefficient with log $P_0 = 1.17$. When an amino nitrogen is present on the B ring, increased potency is favored by acylation of that nitrogen. The most potent compound of the series was the 7-fluoroacetamido analogue. Strong electron-withdrawing groups substituted at the 10 position of the tropolone ring destroy activity. Electron-releasing groups at position 10 improve potency slightly but have a limited effect.

Colchicine (1) is a potent mitotic inhibitor which occurs



1 (colchicine), $R_1 = CH_3CO$; $R_2 = CH_3O$

naturally in the autumn crocus, *Colchicum autumnale L.* The antitumor property of colchicine has been recognized for over 3 millennia.[1] The antitumor effect has recently been shown by Kram and Schmitt[2] to be due to a binding of colchicine around a cysteine residue on the tubulin polypeptide chain. They conclude that this binding results from an interaction of the aromatic rings of colchicine with hydrophobic domains of tubulin.

The colchicines continue to find limited and sporadic use in the treatment of neoplasms.[3] The lack of interest in the colchicines among clinicians apparently arises from the clear superiority of the vinca alkaloids in the management of advanced lymphomas and Hodgkin's disease.[4] The extreme toxicity of colchicine is another factor which works against its acceptance in the treatment of human cancer.[5]

Many attempts have been made to discover more effective and less toxic analogues of colchicine. This search has not been without some success. For example, deacetyl-*N*-methylcolchicine (colcemid, 4) has proven to be less toxic than colchicine and has been used in the treatment of chronic granulocytic leukemia.[6]

In the search for improved colchicines, a number of useful but sometimes contradictory qualitative structure–activity postulates have been developed. These are scattered in the literature and we have summarized them here. In their original monograph, Eigsti and Dustin[1] listed the folllowing then known requirements for the antimitotic activity of colchicine derivatives: (1) at least one methoxy group on the A ring; (2) the amino group on the B ring should be acylated; (3) ring C should be seven membered;

(1) Eigsti, O. J.; Dustin, P. "Colchicine in Agriculture, Medicine, Biology and Chemistry"; Iowa State College Press: Ames, Iowa, 1955; Chapter 1.
(2) Kram, R.; Schmitt, H. *J. Supramol. Struct.* **1978**, *Suppl. 2*, 328.
(3) Gomez, G. A.; Sokal, J. E.; Aungst, C. W. *Proc. Am. Assoc. Cancer Res.* **1977**, *18*, 194.
(4) Creasy, W. A. "Antineoplastic and Immunosuppressive Agents II"; Sartorelli, A. C.; Johns, D. G., Eds.; Springer-Verlag: Berlin, 1975; Chapter 67.
(5) Dowling, M. D.; Krakoff, I. H.; Karnofsky, D. A. "Chemotherapy of Cancer"; Cole, W. H.; Ed.; Lea and Febiger: Philadelphia, 1970; Chapter 1.
(6) Spiers, A. S.; Kaur, J.; Richards, H. G. *Clin. Oncol.* **1975**, *1*(4), 285.