(5) M. Meselson and K. Russell in "Origins of Human Cancer", H. H. Hiatt, J. D. Watson, and J. A. Winstein, Eds., Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y., 1977, p 1473.
(6) J. Ashby and J. A. Styles, *Nature (London)*, 271, 452 (1978).
(7) B. N. Ames and K. Hooper, *Nature (London)*, 274, 20 (1978).
(8) C. Hansch, *Acc. Chem. Res.*, 2, 232 (1969).
(9) C. Hansch, *J. Med. Chem.*, 19, 1 (1976).
(10) G. J. Hatheway, C. Hansch, K. H. Kim, S. R. Milstein, C. L. Schmidt, R. N. Smith, and F. R. Quinn, *J. Med. Chem.*, 21, 563 (1978).
(11) C. Hansch, G. J. Hatheway, F. R. Quinn, and N. Greenberg, *J. Med. Chem.*, 21, 574 (1978).
(12) R. Preussmann, S. Ivankovic, C. Landschuetz, J. Gimmy, E. Flohr, and O. Griesbach, *Z. Krebsforsch. Klin. Onkol.*, 81, 285 (1974).
(13) B. N. Ames, W. E. Durston, E. Yamasaki, and F. D. Lee, *Proc. Natl. Acad. Sci. U.S.A.*, 70, 2281 (1973).
(14) H. J. Vogel and D. M. Bonner, *J. Biol. Chem.*, 218, 97 (1956).
(15) C. Hansch and T. Fujita, *J. Am. Chem. Soc.*, 86, 1616 (1964).
(16) Y. C. Martin and C. Hansch, *J. Med. Chem.*, 14, 777 (1971).
(17) E. Bueding and R. P. Batzinger in "Origins of Human Cancer", H. H. Hiatt, J. D. Watson, and J. A. Winstein, Eds., Cold Spring Harbor Laboratory, 1977, Cold Spring Harbor, N.Y., p 445.
(18) (a) C. Malaveille, G. F. Kolar, and H. Bartsch, *Mutat. Res.*, 36, 1 (1976); (b) G. F. Kolar and J. Schlesiger, *Cancer Lett.*, 1, 43 (1975); (c) G. F. Kolar and J. Schlesiger, *Chemotherapy*, 8, 91 (1976); (d) G. F. Kolar and J. Schlesiger, *Chem.-Biol. Interact.*, 14, 301 (1976); (e) G. F. Kolar, R. Fahrig, and E. Vogel, *ibid.*, 9, 365 (1974).
(19) Y. C. Martin, "Quantitative Drug Design", Marcel Dekker, New York, N.Y., 1978, Chapter 5.
(20) A. Cammarata and K. S. Rogers in "Advances in Linear Free Energy Relationships", N. B. Chapman and J. Shorter, Eds., Plenum Press, London, 1972, p 412.

Computer-Assisted Structure–Activity Studies of Chemical Carcinogens. A Heterogeneous Data Set

Peter C. Jurs, J. T. Chou, and M. Yuan

Department of Chemistry, The Pennsylvania State University, University Park, Pennsylvania 16802.
Received November 13, 1978

A structure–activity relations study has been performed on a heterogeneous set of organic compounds to develop predictive ability for carcinogenic potential. The compounds employed came from more than 12 structural classes and numbered 130 carcinogens and 79 noncarcinogens. A set of 28 calculated molecular structure descriptors was identified that supported a linear discriminant function able to completely separate 192 compounds into the carcinogenic and noncarcinogenic classes. A predictive ability of 90% for carcinogens and 78% for noncarcinogens was obtained in randomized testing. The results demonstrate that pattern-recognition methods can be used to analyze a diverse set of compounds each represented by calculated molecular structure descriptors for a common biological activity.

The attempt to rationalize the connection between the molecular structures of organic compounds and their biological activities comprises the field of structure–activity relations (SAR) studies. Correlations between structure and activity are important for the development of pharmacological agents, herbicides, pesticides, and chemical communicants (olfactory and gustatory stimulants) and the investigation of chemical toxicity and mutagenic and carcinogenic potential. Practical importance attaches to these studies because the results can be used to predict the activity of untested compounds. In addition, SAR studies can direct the researcher's attention to molecular features that correlate highly with biological activity, thus suggesting mechanisms or further experiments. SAR studies have been used to some extent in the pharmaceutical and agricultural industries. The methods are beginning to be applied to the important problems of chemical toxicity and chemical mutagenesis and carcinogenesis.

Evidently, chemical carcinogenesis poses a public health problem of enormous magnitude. A dilemma confronts regulatory agencies and chemically related industry, namely, how to test the enormous numbers of compounds that are produced or could be produced in order to avoid exposure to toxic materials which could lead to adverse effects among the population. While the most satisfactory approach is to use rodent testing, this is uneconomical for the large number of compounds involved. Thus, a new set of short-term tests have been applied to the problem.[1,2] Recently, a set of techniques drawn from SAR studies in pharmacology have been applied to sets of chemical carcinogens in an attempt (a) to develop predictive capability for unknown compounds and (b) to further fundamental understanding of the structural features of molecules that can lead to carcinogenic potential.

The superior way to develop predictive capability is to understand, at the molecular level, the mechanisms that lead to the undesired carcinogenic initiation reaction. Unfortunately, this knowledge is not yet available for most classes of chemical carcinogens. Furthermore, the progress made through a living system by a carcinogen or its precursors is not usually known. Thus, two choices are presented: study the mechanisms for a very few compounds to develop fundamental information for those few compounds, or use empirical methods to study larger sets of compounds with correlative methods. The latter method comprises an SAR approach to the study of chemical carcinogenesis. Thus, one has available a set of compounds that has been tested in rodent tests and the observations that resulted from the tests. One can then search for correlations between the structures of the compounds tested and the biological observations reported. One is actually modeling the entire process of uptake, transport, metabolism, cell penetration, binding, etc.

The discovery and design of biologically active compounds (drug design) is a field that has been subject to widespread and well-documented[3-9] changes in the past decade. A host of new techniques and perspectives have evolved. While these techniques have been used largely for the development of pharmaceuticals, they can also be applied to the rationalization of structure–activity relations among sets of toxic, mutagenic, or carcinogenic compounds.

0022-2623/79/1822-0476$01.00/0

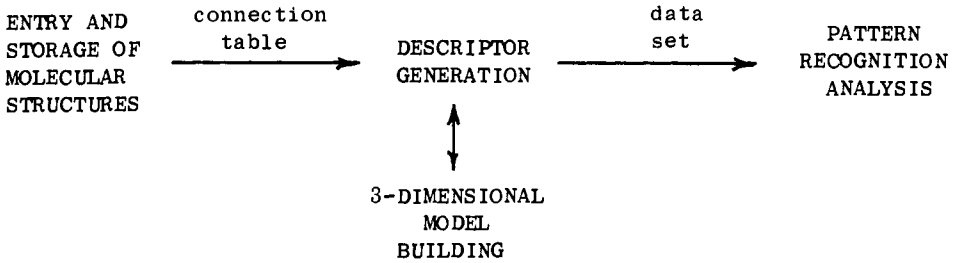


Figure 1. Flow chart of steps involved in structure–activity studies using chemical structure information handling and pattern-recognition methods.

Several approaches to SAR have been reported: the semiempirical linear free energy (LFER) or extrathermodynamic model proposed by Hansch and co-workers;[10-12] the additivity or Free–Wilson model;[13] quantum mechanically based models;[14,15] and pattern-recognition methods.[16] Reviews are cited that describe the progress made using each of the approaches.

Chemical and SAR Applications of Pattern Recognition. Pattern recognition is a subfield of artificial intelligence comprised of a set of nonparametric techniques used to study data sets that may not conform to well-characterized probability density functions. A voluminous literature describes the field.[17-20]

Most of the pattern-recognition methods share a set of common properties. The data to be analyzed, here molecular structures of carcinogens and noncarcinogens of interest, are represented by points in a high-dimensional space. For a given compound, which is represented by a given point, the value of each coordinate is just the numerical value for one of the molecular structure descriptors comprising the representation. The expectation is that the points representing carcinogenic compounds will cluster in one limited region of the space, while the points representing the noncarcinogenic compounds will cluster elsewhere. Pattern recognition 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. For example, Bayesian classification surfaces are developed using the mean vectors for the members of the classes and the covariance matrices for the classes. If the statistical properties cannot be calculated or estimated, then nonparametric methods are used. Nonparametric methods attempt to find clustering definitions or classification surfaces by using the data themselves directly, without computing mean vectors, convariance matrices, etc. Examples of nonparametric methods would include error-correction feedback linear learning machines (threshold logic units or perceptrons) and simplex optimization methods of searching for separating classification surfaces. Complete descriptions of these methods can be found in standard texts (e.g., see ref 20).

Applications of pattern-recognition methodology to chemical problems were first reported in the 1960's[21,22] with studies of mass spectra. Since then, papers have described work in a variety of areas,[23,24] including mass spectrometry, infrared spectroscopy, NMR spectroscopy, electrochemistry, materials science and mixture analysis, and the modeling of chemical experiments. Diagnoses of pathological conditions from sets of measurements made on complex biological mixtures, e.g., serum, have been reported.[25] The successes in these areas have led to the belief that these methods should prove useful in the development of structure–activity relations.

Table I. Heterogeneous Data Set of Chemical Carcinogens

	+	−	weak	?
aromatic amines	23	10	1	10
alkyl halides	13	2		4
polycyclic aromatics	23	6	3	4
esters, epoxides, carbamates	11	7	5	5
NO_2 aromatics & heterocycles	14	3		3
misc	5	7		8
nitrosamines	19	2		1
fungal toxins & antibiot	3	1	2	1
misc	2	2	1	6
misc N compds	8	1		
azo dyes and diazo compds	9	2	2	4
naturally occurring compds		36		
total	130	79	14	46

A number of studies of the application pattern recognition to the problem of searching for correlations between molecular structure and biological activity have been reported. A large fraction of the effort in this area must be devoted to the generation of appropriate descriptors from the molecular structures available. Areas of study include drug structure–activity relations, studies of chemical communicants, etc. Applications of pattern recognition to drug design have been reviewed by Kirschner and Kowalski.[16]

Recently, papers have begun to appear reporting the application of SAR methods to sets of chemical carcinogens. One study reports correlations between the number of carbon atoms per compound and carcinogenic activity for a set of 47 carcinogenic nitrosamines.[26] Correlations of carcinogenic activity and liposolubility for small sets of cyclic nitrosamines have been reported.[27] Carcinogenic activity was correlated with theoretical reactivity indices for 25 representative polycyclic aromatic hydrocarbons.[28] A quantitative relationship was reported that correlated carcinogenic activity with the hydrophobic parameter π and an electronic constant for 21 nitrosamines.[29] Correlations were reported between nitrosamine carcinogenic activity and molecular connectivity indices.[30] Another study of cyclic nitrosamines found correlations between carcinogenic potency and aqueous and octanol solvation free energies.[31] A paper has appeared describing the application of the SIMCA method of pattern recognition to the SAR study of 4-nitroquinoline 1-oxides.[32]

Experimental Section

The SAR studies of chemical carcinogens reported here were done using the ADAPT computer software system,[33] which was run on the Department of Chemistry MODCOMP II/25 16-bit digital computer. The fundamental steps involved in performing an SAR study using this system are shown in Figure 1. The individual steps are as follows: (a) Identify, assemble, input, store, and describe a data set of structures for chemicals that have been tested for carcinogenic activity. (b) Develop computer-generated

Table II. Chemical Carcinogen Data Set

compds	carcinogenicity[a]
aromatic amines	
N-acetoxy-2-(acetylamino)fluorene	+
2-aminofluorene	+
2-(acetylamino)fluorene	+
N-hydroxy-2-(acetylamino)fluorene	+
N-hydroxy-2-aminofluorene	+
2-nitrosofluorene	+
2,7-bis(acetylamino)fluorene (D)	+
2,7-diaminofluorene (D)	+
6-aminochrysene	+
acridine orange	+
2-aminoanthracene	+
2-naphthylamine	+
2-naphthylhydroxylamine	+
1-naphthylhydroxylamine	+
2-nitrosonaphthalene	+
4-aminobiphenyl	+
2′,3-dimethyl-4-aminobiphenyl	+
benzidine	+
4-amino-*trans*-stilbene	+
4-(dimethylamino)-*trans*-stilbene	+
4,4′-methylenebis(2-chloroaniline)	+
2,4-diaminotoluene	±
auramine	+
7-hydroxy-2-(acetylamino)fluorene (D)	−
1-hydroxy-2-(acetylamino)fluorene	−
3-hydroxy-2-(acetylamino)fluorene	−
5-hydroxy-2-(acetylamino)fluorene	−
4-(acetylamino)fluorene	−
1-naphthylamine	−
3,3′,5,5′-tetramethylbenzidine (D)	−
p-aminodiphenylamine (D)	−
p-aminophenol	−
o-toluidine	−
alkyl halides	
bis(chloromethyl) ether	+
methylbis(2-chloroethyl)amine	+
uracil mustard	+
cyclophosphamide	+
isophosphamide	+
chlornaphazin	+
melphalan	+
ICR-10	+
ICR-170	+
dimethylcarbamyl chloride	+
captan	±
DDE	+
dieldrin	+
chloroacetic acid	−
folpet	−
nitro aromatics and heterocycles	
4-nitrobiphenyl	+
2-nitronaphthalene	+
5-nitroacenaphthene	±
2-nitrofluorene	±
4-nitroquinoline 1-oxide	+
4-(hydroxyamino)quinoline 1-oxide	±
metronidazole	+
1,2-dimethyl-5-nitroimidazole	+
niridazole	+
1,2-dihydro-2-(5-nitro-2-thienyl)-quinazolin-4(3*H*)-one	+
4-[(2-hydroxyethyl)amino]-2-(5-nitro-2-thienyl)quinazoline	+
4-bis[(2-hydroxyethyl)amino]-2-(5-nitro-2-thienyl)quinazoline	+
2-(2-furyl)-3-(5-nitro-2-furyl)acrylamide	+
N-[4-(5-nitro-2-furyl)thiazolyl]-formamide	±
5-nitro-2-furamidoxime (D)	−
1-[(5-nitrofurfurylidene)amino]-hydantoin (D)	−
5-nitro-2-furoic acid	−
miscellaneous	
ethionine	±
safrole (D)	+
1′-hydroxysafrole (D)	+
1′-acetoxysafrole	+
diethylstilbesterol	+
acetone	−
acetic acid	−
ethanol	−
ethylene glycol	−
dimethyl sulfoxide	−
1-naphthyl isothiocyanate	−
12-*O*-tetradecanoylphorbol 13-acetate	−
miscellaneous nitrogen compounds	
propylenimine	+
ethylenimine	+
tris(1-aziridinyl)phosphine sulfide	+
1,2-dimethylhydrazine	+
N-(2-hydroxyethyl)hydrazine	+
natulan	+
1-phenyl-3,3-dimethyltrizine	+
1-(4-chlorophenyl)-3,3-dimethyltriazine	+
maleic hydrazide	−
***N*-nitroso compounds**	
dimethylnitrosamine	+
diethylnitrosamine	+
di-*n*-propylnitrosamine	+
di-*n*-butylnitrosamine	+
di-*n*-pentylnitrosamine	+
N-nitrosopyrrolidine	+
N-nitrosomorpholine	+
N-nitrosopiperidine	+
N-methyl-*N*′-nitro-*N*-nitrosoguanidine	+
N-ethyl-*N*′-nitro-*N*-nitrosoguanidine	+
N-propyl-*N*′-nitro-*N*-nitrosoguanidine	+
N-butyl-*N*′-nitro-*N*-nitrosoguanidine	+
N-isobutyl-*N*′-nitro-*N*-nitrosoguanidine	+
N-pentyl-*N*′-nitro-*N*-nitrosoguanidine	+
N-nitrosomethylurea	+
N-nitrosoethylurea	+
N-nitroso-*N*-methylurethane	+
streptozotocin	+
cycasin	+
dibenzylnitrosamine	−
diphenylnitrosamine	−
polycyclic aromatics	
dibenzo[*a*,*e*]pyrene	+
dibenzo[*a*,*i*]pyrene	+
benzo[*a*]pyrene	+
6-(hydroxymethyl)benzo[*a*]pyrene	±
benzo[*a*]pyrene 4,5-oxide	+
7,8-dihydrobenzo[*a*]pyrene	+
dibenz[*a*,*j*]acridine	+
dibenz[*a*,*c*]anthracene	±
dibenz[*a*,*h*]anthracene	+
3-methylcholanthrene	+
benz[*a*]anthracene	±
7-methylbenz[*a*]anthracene	+
7,12-dimethylbenz[*a*]anthracene	+
7-(hydroxymethyl)-12-methylbenz[*a*]-anthracene	+
7-chloromethyl-12-methylbenz[*a*]-anthracene	+
7-(chloromethyl)benz[*a*]anthracene	+
7-(bromomethyl)-12-methylbenz[*a*]-anthracene	+
7,9-dimethylbenz[*c*]acridine	±
7,10-dimethylbenz[*c*]acridine	+
9,10-(dichloromethyl)anthracene	+
10-(chloromethyl)-9-methylanthracene	+
10-(chloromethyl)-9-chloroanthracene	+
10-(bromomethyl)anthracene	+
dibenz[*a*,*h*]anthracene 5,6-oxide (D)	−
pyrene	−
anthracene	−
phenanthrene	−

Table II (*Continued*)

compds	carcinogenicity[a]
polycyclic aromatics	
fluorene	−
naphthalene	−
esters, epoxides, and carbamates	
methyl methanesulfonate	+
β-propiolactone	+
β-butyrolactone	+
1,3-propane sultone (D)	+
2,3-epoxypropionaldehyde	+
1,2:3,4-diepoxybutane	+
1,2:7,8-diepoxyoctane	+
ethyl carbamate	+
1-phenyl-1-(3,4-xylyl)-2-propynyl *N*-cyclohexylcarbamate	+
1,1-diphenyl-2-butynyl *N*-cyclohexylcarbamate	+
1,1-diphenyl-2-propynyl *N*-cyclohexylcarbamate	+
ethyl methanesulfonate	−
vinyl acetate	−
n-propyl *p*-hydroxybenzoate ester	−
malathion	−
styrene oxide	−
1-naphthyl *N*-methylcarbamate	−
diphenylethinylcarbinol	−
azo dyes and diazo compounds	
o-aminoazotoluene	+
3-methoxy-4-aminoazobenzene	+
N-methyl-4-aminoazobenzene	+
N,N-dimethyl-4-aminoazobenzene	+
3′-methyl-4-(dimethylamino)azobenzene	+
N-(benzoyloxy)-4-methylaminoazobenzene	+
azaserine	+
diazoacetylglycinamide	+
diazoacetylglycinehydrazide	+
N,N-diethyl-4-aminoazobenzene (D)	−
methyl orange (D)	−
naturally occurring compounds	
adenosine (D)	−
cytidine	−
nicotinamide	−
naturally occurring compounds	
d-pantothenic acid	−
riboflavin	−
L-ascorbic acid	−
thioctic acid	−
β-L-arabinose	−
α-D-galactose	−
L-fucose	−
α-D-glucose	−
inositol	−
α-D-ribose	−
α-D-glucosamine	−
maltose	−
lactose	−
sucrose	−
D-(−)-gluconic acid	−
diaminopimelic acid	−
L-glutamic acid	−
L-asparagine	−
glycine	−
L-methionine	−
L-phenylalanine	−
L-tyrosine	−
L-tryptophan	−
L-lysine	−
glutathione	−
ethyl acetate	−
citric acid	−
glycerol	−
propylene glycol	−
dl-tartaric acid	−
indole	−
spermidine (D)	−
putrescine (D)	−
fungal toxins	
aflatoxin B1	+
aflatoxin G1	+
sterigmatocystin	+
aflatoxin P1	−
miscellaneous heterocycles	
amitrole	+
hycanthone methanesulfonate	+
5-iododeoxyuridine	−
nicotine (D)	−

[a] Abbreviations used: +, carcinogenic; −, noncarcinogenic.

molecule descriptors for each of the members of the data. The descriptors may be derived directly from the stored topological representations of the structures or they may require the development of three-dimensional models. (c) Using pattern-recognition methods, develop classifiers to discriminate between carcinogens and noncarcinogens based on the sets of molecular descriptors. (d) Test the predictive ability of these discriminants of compounds of unknown activity. (e) Systematically reduce the set of molecular structure descriptors employed to the minimum set sufficient to retain discrimination between the carcinogens and noncarcinogens and to retain high predictive ability.

Data Set. The data used in this study were taken from a published compilation of tested compounds.[34] All 269 compounds that had reported carcinogenic activity in the publication were entered into the ADAPT disk files. Entry of the structures was accomplished by sketching them on the screen of a graphics display terminal under the control of an interactive program. This can be done in 30 s to several minutes per compound, depending on structural complexity. The structure files are stored permanently on disk files for further processing by the other modules of ADAPT. A table of statistics describing this data set is shown in Table I. A subset of the total data set was investigated; it consists of the carcinogens (column 1 of Table I labeled +) and the noncarcinogens (column 2 in Table I labeled −) and includes 209 compounds in all. They are spread over more than 12 structural classes. Procarcinogens, proximate carcinogens, and ultimate carcinogens are represented among the structures. The identities of each of the 209 compounds are listed in Table II. There is a total of 130 carcinogens and 79 noncarcinogens.

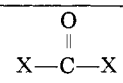
Descriptor Generation. After entry of the data set to disk files and after a set of stored compounds has been chosen for study, then the next step in an SAR study is the generation of molecular structure descriptors.[33]

There are three general classes of descriptors: topological, geometrical, and physicochemical. Topological descriptors are derived from the topological representation of the structure, the connection table. The geometrical descriptors are derived from the three-dimensional model of the molecule. Physicochemical descriptors may be measured experimentally, calculated using a mathematical model, or represented by linearly correlated calculable descriptors. The descriptors that are currently available in ADAPT are as follows.

(a) 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.

(b) Substructure Descriptors. ADAPT has a substructure searching routine that can be used to develop 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 problem dependent and must be found through the application of common sense and experience by the researcher.

Table III. Environment Descriptor Examples



compd structure	connectivity values	compd name
$H_2NC(=O)N(CH_3)N=O$	1.61	*N*-nitrosomethylurea
$NH_2C(=O)Ph$	1.56	nicotinamide
$CH_3C(=O)OCH_2CH_3$	1.32	ethyl acetate
$CH_3C(=O)CH_3$	1.20	acetone
$H_2NC(=O)OCH_2CH_3$	1.11	ethyl carbamate
$CH_3C(=O)OH$	0.93	acetic acid

Table IV. Descriptors Used with 209 Compound Heterogeneous Data Set

no.	fragment descriptors
1	no. of nitrogen atoms
2	no. of chlorine atoms
3	no. of double bonds
4	no. of basis rings
no.	**molecular connectivity descriptors**
5	path 1 molecular connectivity
6	path 2 molecular connectivity
7	path 4 molecular connectivity
no.	**geometric descriptor**
8	smallest principal moment

(c) **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 substructural 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 of the substructure as imbedded within the molecule being coded. Table III shows the values obtained for the environment descriptor based on the carbonyl group substructure as imbedded in a variety of simple compounds.

(d) **Molecular Connectivity Descriptors.** The molecular connectivity of a molecule is a measure of the branching of the structure. It 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. This is the path-1 molecular connectivity. 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 to be correlated with a number of physicochemical parameters, such as partition coefficients and steric parameters.[35,36]

(e) **Geometric Descriptors.** Given a three-dimensional model of the structures being coded, one can calculate descriptors designed to represent the shape of the molecules. We calculate the three principal moments of inertia and their ratios and the molecular volume.

In addition to the individual descriptor generation routines, ADAPT has several other supporting routines. There is a general-purpose descriptor file-management routine that allows the review of any stored descriptor, for example. There is a routine that allows mathematical manipulation of descriptors, such as addition, multiplication, logarithmic transformation, exponentiation, autoscaling, etc. There is a routine that calculates the linear correlation coefficient of any or all pairs of descriptors.

For each of the 209 compounds, a large number of descriptors was developed and tested. The set of 26 descriptors that was the finally selected set and that provides the best performance found is that shown in Tables IV to VI. Included are 4 fragment descriptors, 3 molecular connectivity descriptors, 1 geometric descriptor, 13 environment descriptors, and 5 combined descriptors.

Table V. Environment Descriptors for 209 Compound Heterogeneous Data Set

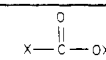
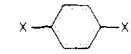
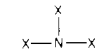
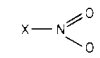
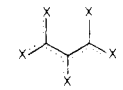
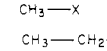
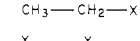
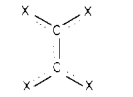
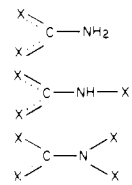
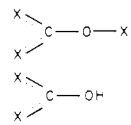
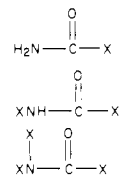
no.	substructure	no. of occurrences	no.	substructure	no. of occurrences
9		21	17		19
10		53	18		21
11		54			
12		22	19		28
13					
14		22	20		88
15		88			
16		22 47	21		49

Table VI. Combined Descriptors for 209 Compound Heterogeneous Data Set

no.	substructures	no. of occurrences
22	no. of ring atoms/no. of basis rings	
23		55
24		21
25		44
26	largest principal moment × intermediate principal moment = "area"	

Some statistics regarding these 26 descriptors are given in Table VII. The mean value and standard deviation are shown for each descriptor. Within the data set of 209 compounds, the minimum number of nonzero descriptors for any compound is 4, the maximum number is 18, and the mean number is 10.9 with a standard deviation of 2.8. The average linear correlation coefficient over all pairs of the descriptors is 0.17. The final column of Table VII shows the descriptor values calculated for the compound 2-acetylaminofluorene (2-AAF). 2-AAF has 15 nonzero

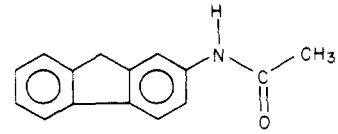


Table VII. Means and Standard Deviations for the 26 Descriptors and the Descriptor Values for 2-Acetylaminofluorene

descriptor no.	mean value	SD	value for 2-AAF
1	1.35	1.37	1
2	0.225	0.755	0
3	1.17	1.35	1
4	1.87	1.54	3
5	4.14	1.96	4.84
6	3.40	1.88	4.39
7	8.04	6.50	12.23
8	0.152	0.218	0.167
9	0.234	0.714	0
10	0.503	0.888	2.17
11	0.671	1.17	0
12	0.137	0.404	0
13	0.134	0.391	0
14	0.560	0.672	1.35
15	0.183	0.541	0
16	0.812	1.52	0
17	0.268	0.859	0
18	0.335	1.01	0
19	0.580	1.49	0
20	0.829	1.03	1.45
21	0.734	1.34	3.39
22	3.79	2.26	4.33
23	0.723	1.28	2.87
24	0 243	0.792	0
25	0.476	1.13	2.02
26	14.4	22.3	11.0

descriptors. Only one double bond is counted because aromatic bonds are considered separately. Thus, while each compound in the data set is represented by a point in a 26-dimensional space, a reasonable fraction of the 26 descriptors have values of zero.

Pattern-Recognition Analysis. Once each of the compounds comprising the data set was represented by a set of descriptors, then pattern-recognition methods were used to analyze the data. The goal was to identify invariant properties among those coded in the data set and to develop discriminants that would separate the carcinogens from the noncarcinogens. A variety of heuristic pattern-recognition methods were employed.

During the course of these studies, no set of descriptors was found that would support a discriminant capable of completely separating all 130 carcinogens from all 79 noncarcinogens. The best sets of descriptors produced recognition percentages of 90–95% for the best linear discriminants. Therefore, a set of experiments was performed to identify a subset of the compounds that would be linearly separable.

The study was started using the best set of descriptors available as far as could be determined by experience working with the data set. In the present case, there were approximately 30 descriptors worthy of inclusion. Then the iterative least-squares routine was used to train the best linear discriminants possible; such a discriminant incorrectly classified approximately 18 compounds out of 209 for a recognition rate of 91.4%. Then the 10 compounds causing the most trouble to the iterative least-squares algorithm during its attempt to find a separating linear discriminant were identified. A training set with these 10 compounds excluded was generated. Then this training set of 199 compounds was provided to the iterative least-squares algorithm, and another attempt was made to develop a linear discriminant. This time, approximately 10 more compounds were preventing the algorithm from finding a separating discriminant. These were excluded and the process was repeated. After several iterations, a set of 17 compounds was identified that were preventing linear separation. They are marked as deleted (D) in Table II. A training set, called training set A, was generated that excluded these 17 compounds and included all 192 others. This training set was then used to determine which of the descriptors under investigation were worthy of retention. After a number of trial studies, a set of 26 descriptors was identified that was sufficient to support linear separability for the training set of 192 compounds. These are the 26 descriptors shown in Tables IV to VI.

Table VIII shows some results obtained using a variety of pattern-recognition techniques with the entire data set of 209 compounds and the training set of 192 compounds. Of course, the linear learning machine achieves a recognition rate of 100% on training-set A. The iterative least-squares algorithm cannot quite find a linear discriminant that fully separates the carcinogens from the noncarcinogens for training-set A. All other algorithms produce less satisfactory results, even for training-set A, which is known to be linearly separable.

After the set of 26 descriptors had been identified, then their predictive ability was assessed. Table IX shows the results obtained. Out of the total of 209 compounds in the data set, the 17 previously mentioned were excluded. Training set/prediction set combinations were chosen using random selection out of the

Table VIII. Pattern Recognition Results for 209 Compound Heterogeneous Data Set and 26 Descriptors

		no. correctly classified[a]		% correctly classified[a]		
		+	−	+	−	total
Bayes (quadratic)	all data	115	75	88.5	94.9	90.9
	set A	116	65	92.8	92.8	94.3
Bayes (linear)						
all	all data	100	65	76.9	82.3	79.0
− class	all data	86	75	66.2	94.9	77.0
+ class	all data	115	49	88.5	62.0	78.5
all	set A	101	63	80.8	94.0	85.4
− class	set A	90	65	72.0	97.0	80.7
+ class	set A	113	49	90.4	73.1	84.4
learning machine	all data	108	50	83.1	63.3	75.6
	all data	107	56	82.3	70.9	78.0
	set A	125	67	100.0	100.0	100.0
iterative least squares	all data	121	65	93.1	82.3	89.0
	set A	124	66	99.2	98.5	99.0
K nearest neighbor						
$K = 1$	all data	105	49	80.8	62.0	73.7
$K = 1$	set A	102	45	81.6	67.2	76.6

set A composition: no. in + class = 125; no. in − class = 67; total 192

[a] Abbreviations used: +, carcinogen class; −, noncarcinogen class.

Table IX. Predictive Ability Tests for 209 Compound Heterogeneous Data Set and 26 Descriptors[a]

	carcinogens	noncarcinogens	total
A. populations of training sets and prediction sets			
training set	119	63	182
prediction set	6	4	10
total	125	67	192
	no. predicted	no. correct	% correct
B. predictive ability			
carcinogens	180	162	90.0
noncarcinogens	120	94	78.3
total	300	256	85.3

[a] Number of sets employed = 10; number of compounds excluded = 17.

remaining 192 compounds. Each prediction set contained 10 randomly chosen members—6 from the carcinogen class and 4 from the noncarcinogen class. These numbers were chosen because they reflect the overall populations in these two classes. A total of 10 such training set/prediction set pairs were chosen. These 10 training set/prediction set pairs were used in 30 separate studies to determine predictive ability. For each study, the training set of 182 compounds was used to develop a linear discriminant, and then without further change this discriminant was used to predict the class of the 10 members of the prediction set. The results reported in Table IX are the averages for 30 such trials involving a total of 300 individual predictions. An overall predictive ability of 85.3% was obtained. The predictive success for the carcinogens was substantially higher at 90% than that for noncarcinogens at 78%. This means that when incorrect predictions are made, the error is more likely to arise because a noncarcinogen is predicted to be a carcinogen than the reverse case. If errors are committed, this is the preferable way to commit them.

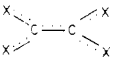
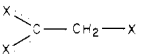
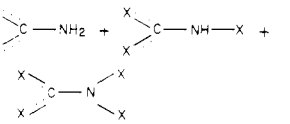
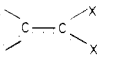
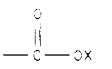
Multiple Linear-Regression Analysis. The ADAPT system has a stepwise MLR program interfaced to it so that any set of descriptors stored on disk can be tested in a convenient manner. Several MLR studies were run on the heterogeneous data set. The dependent variable used was derived from the reported number of regressions per nanomole in the original paper.[34] This measure of mutagenic potency covers a wide range, from zero to multiples of 10^4. Therefore, it was felt that a logarithmic transformation was desirable. If Y represents the value reported in the original paper for reversions per nanomole for a compound, then the final dependent variable used in the regression analysis was $1/[\log_{10}(Y + 2)]$. The value of 2 was added to Y before the log operation to avoid later division in zero. The new dependent variable has a mean of 2.17 and a standard deviation of 1.24 for the 209 compounds in the data set. A reversion value of zero converts to a dependent variable value of 3.32; a value of 50 for Y converts to a value of 0.58. In general, a small value for the dependent variable indicates more mutagenic potency.

The data set of 209 compounds, each represented by the same descriptors used for pattern-recognition studies, was presented to the stepwise MLR program, with the results shown in Table X. The variables are listed in the sequence chosen by the program; the identity of each variable is shown in the continuation of Table X. With 14 variables chosen, the multiple correlation coefficient was 0.653 for this analysis. Thus, only 42.6% of the total variation about the mean was explained by the regression equation. The standard error was 0.974. Thus, it is seen that MLR analysis using this set of descriptors developed for the pattern-recognition studies is not particularly successful. Additionally, it should be remembered that the descriptors being used were developed for the separation of carcinogens from noncarcinogens, whereas the MLR analysis dependent variable is mutagenic potency.

A second stepwise MLR analysis of the same data set using the same set of descriptors was done. The dependent variable used was in $(Y + 1)$. This variable has a mean of 1.66 and a standard deviation of 2.33. With 16 variables entered into the regression equation, the multiple correlation coefficient was 0.619 and the standard error was 1.90. This equation explained only 38.3% of the variance in the dependent variable, and therefore the details of the analysis are not presented.

Table X. Linear Regression Analysis Results for 209 Compound Heterogeneous Data Set[a]

no.	index	variable name	coefficient
1	5	FRAG 17	−0.238
2	16	ENVR 16	−0.491
3	21	ENVR 24	−0.297
4	1	FRAG 4	−0.222
5	10	ENVR 1	−0.109
6	25	MATH	0.182
7	20	ENVR 23	−0.173
8	11	ENVR 2	−0.479
9	22	ENVR 26	−0.139
10	18	ENVR 21	−0.195
11	29	MATH	0.007
12	3	FRAG 13	−0.214
13	24	ENVR 31	0.227
14	13	ENVR 8	−0.117
	const		3.457

variable name	variable identity
FRAG 17	no. of basis rings
ENVR 16	
ENVR 24	
FRAG 4	no. of N atoms
ENVR 1	
MATH	
ENVR 23	
ENVR 2	X—Cl
ENVR 26	
ENVR 21	CH_3—X
MATH	GEOM 1 × GEOM 2 = "area"
FRAG 13	no. of dbl bonds
ENVR 31	
ENVR 8	

[a] $N = 209$; $R = 0.653$; SE = 0.974.

Discussion

The fundamental premises involved in applying pattern-recognition methods to SAR studies of chemical carcinogens have been confirmed by the results obtained in this study. First, chemical carcinogens and compounds of similar structural classes that are noncarcinogens can be adequately represented by calculated molecular structure descriptors. In this context, adequately means that linear discriminants can be found that separate large numbers of carcinogens from noncarcinogens based on these calculated descriptors. Second, it has been demonstrated that discriminants can be found based on calculated descriptors that can completely separate large numbers of known chemical carcinogens from large sets of noncarcinogens. Third, these discriminants have been shown to possess substantial predictive ability. While the

predictive ability was not assessed on truly unknown compounds, a statistical procedure that produces unbiased estimates of predictive ability was used. Thus, it has been demonstrated that pattern-recognition procedures may have utility for the prediction of carcinogenic activity of compounds.

Acknowledgment. This research was sponsored 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.

References and Notes

(1) B. A. Bridges, *Nature (London)*, **261**, 195 (1976).
(2) I. F. H. Purchase, E. Longstaff, J. Ashby, J. A. Styles, D. Anderson, P. A. Lefevre, and F. R. Westwood, *Nature (London)*, **264**, 624 (1976).
(3) E. J. Ariens, *Drug Des., 1971–1978*, 1–8 (1971–1978).
(4) A. Burger, "Medicinal Chemistry", Part I, Wiley-Interscience, New York, 1970.
(5) B. Bloom and G. E. Ullyot, Eds., "Drug Discovery", American Chemical Society, Washington, D.C., 1971.
(6) Wade Van Valkenburg, Ed., "Biological Correlations—The Hansch Approach", American Chemical Society, Washington, D.C., 1972, p 252.
(7) W. P. Purcell, G. E. Bass, and J. M. Clayton, "Strategy of Drug Design", Wiley-Interscience, New York, 1973.
(8) Y. C. Martin, "Quantitative Drug Design", Marcel Dekker, New York, 1978.
(9) Science Information Services Department, Franklin Institute Research Laboratories, "Structure Activity Correlation Bibliography: With Subject and Author Index", PB-240 658/5 GA, Mar 1975.
(10) W. J. Dunn, *Annu. Rep. Med. Chem.*, 8, 313 (1973).
(11) R. D. Cramer, *Annu. Rep. Med. Chem.*, 11, 301 (1976).
(12) C. Hansch, in "Advances in Linear Free Energy Relationships", Vol. 2, N. R. Chapman and J. Shorter, Eds., Plenum Press, New York, in press.
(13) P. N. Craig, in ref 6, p 115.
(14) W. G. Richards and M. E. Black, *Prog. Med. Chem.*, 11, 67 (1975).
(15) R. E. Christoffersen, in "Quantum Mechanics of Molecular Conformations", B. Pullman, Ed., Wiley, New York, 1976.
(16) G. L. Kirschner and B. R. Kowalski, *Drug Des., 1978*, 8, (1978).
(17) N. J. Nilsson, "Learning Machines", McGraw-Hill, New York, 1965.
(18) E. A. Patrick, "Fundamentals of Pattern Recognition", Prentice-Hall, Englewood Cliffs, N.J., 1972.
(19) H. C. Andrews, "Introduction to Mathematical Techniques in Pattern Recognition", Wiley-Interscience, New York, 1972.
(20) J. T. Tou and R. C. Gonzalez, "Pattern Recognition Principles", Addison-Wesley, Reading, Mass., 1974.
(21) V. L. Tal'roze, V. V. Raznikov, and G. D. Tantsyrev, *Dokl. Akad. Nauk SSSR*, **159**(1), 182 (1964).
(22) P. C. Jurs, B. R. Kowalski, and T. L. Isenhour, *Anal. Chem.*, 41, 21 (1969).
(23) P. C. Jurs and T. L. Isenhour, "Chemical Applications of Pattern Recognition", Wiley-Interscience, New York, 1975.
(24) B. R. Kowalski, *Anal. Chem.*, **47**, 1152A (1975).
(25) M. L. McConnell, G. Rhodes, U. Watson, and M. Novotny, *J. Chromatogr.*, in press.
(26) J. S. Wishnok and M. C. Archer, *Br. J. Cancer*, **33**, 307 (1976).
(27) G. M. Singer, H. W. Taylor, and W. Lijinsky, *Chem.-Biol. Interact.*, **19**, 133 (1977).
(28) I. A. Smith, G. D. Berger, P. G. Seybold, and M. P. Serve, *Cancer Res.*, **38**, 2968 (1978).
(29) J. S. Wishnok, M. C. Archer, A. S. Edelman, and W. M. Rand, *Chem.-Biol. Interact.*, **20**, 43 (1978).
(30) L. B. Kier, R. J. Simons, and L. H. Hall, *J. Pharm. Sci.*, **67**, 725 (1978).
(31) A. J. Hopfinger and G. Klopman, *Chem.-Biol. Interact.*, in press.
(32) W. J. Dunn III and S. Wold, *J. Med. Chem.*, **21**, 1001 (1978).
(33) A. J. Stuper, W. E. Brugger, and P. C. Jurs, in "Chemometrics: Theory and Application", B. R. Kowalski, Ed., American Chemical Society, Washington, D.C., 1977, p 165.
(34) J. McCann, E. Choi, E. Yamasaki, and B. N. Ames, *Proc. Natl. Acad. Sci. U.S.A.*, **72**, 5135 (1975).
(35) L. B. Kier and L. H. Hall, "Molecular Connectivity in Chemistry and Drug Research", Academic Press, New York, 1976.
(36) W. J. Murray, *J. Pharm. Sci.*, **66**, 1352 (1977).

Multivariate Analysis and Quantitative Structure–Activity Relationships. Inhibition of Dihydrofolate Reductase and Thymidylate Synthetase by Quinazolines

Bor-Kuan Chen, Csaba Horváth,*

Chemical Engineering Group, Department of Engineering and Applied Science, Yale University, New Haven, Connecticut 06520

and Joseph R. Bertino

Department of Pharmacology, School of Medicine, Yale University, New Haven, Connecticut 06510.
Received November 27, 1978

Quantitative structure–activity relationships (QSAR) have been established for the inhibition of dihydrofolate reductase and thymidylate synthetase by 2,4-diaminoquinazoline–glutamic acid analogues. For dihydrofolate reductase from both human acute lymphocytic leukemia cells and murine L1210R cells, QSAR's obtained with 50 quinazolines were similar. On the other hand, for the inhibition of thymidylate synthetase from murine L1210S cells and from *Lactobacillus casei*, QSAR's formulated on the basis of data measured with 33 compounds were different, indicating that the two enzymes are dissimilar. The use of multivariate statistics including cluster analysis, factor analysis, and discriminant analysis is shown to facilitate the formulation of a satisfactory correlation equation. The procedure is demonstrated by the development of QSAR for the inhibition of thymidylate synthetase.

Folate antagonists continue to be useful drugs in the treatment of certain neoplastic diseases. A great number of compounds have been synthetized and tested for biological activity in order to find more potent and less toxic anticancer agents. All clinically useful folate antagonists are inhibitors of the enzyme dihydrofolate reductase[1] (EC 1.5.1.3), and the screening of potential drugs usually commences with the measurements of the inhibitory

0022-2623/79/1822-0483$01.00/0