Computer-Assisted Studies of Molecular Structure–Biological Activity Relationships

PETER C. JURS,* TERRY R. STOUCH, MARIA CZERWINSKI, and JAVIER N. NARVAEZ

Department of Chemistry, The Pennsylvania State University, University Park, Pennsylvania 16802

Received February 1, 1985

Computer-assisted methods can be used to investigate the relationships between the molecular structures of compounds and their biological activity. A number of approaches have been reported in the literature, including correlations of activity with substituent constants, conformational analysis and display, quantum mechanical methods, and methods relying on discriminant development and pattern-recognition techniques. Application areas for this technology include drug design, agricultural chemical design, and studies of chemical toxicity and genetic toxicity (mutagenic or carcinogenic potential). These structure–activity methods are introduced, and citations are given. Several current structure–activity relationship (SAR) studies using pattern recognition are presented as examples of typical projects that are feasible with this approach. These include the investigation of a set of 122 antiinflammatory steroids, a study of 153 retinoids for cancer prevention, and a study of chemicals that have been tested in a sister chromatid exchange mutagen screen.

INTRODUCTION

The rationalization of connections between molecular structure and biological activity comprises the field of structure–activity relationship (SAR) studies. Correlations between structure and activity are important for both the understanding and the rational development of pharmaceutical drugs, agricultural agents, and chemical communicants (olfactory and gustatory stimulants) and for the investigation of chemical and genetic toxicity (mutagenic, carcinogenic, teratogenic potential). These studies have practical importance because they provide the potential for prediction of the activity of untested or even hypothetical compounds. In addition, the insights generated by SAR studies can focus attention on molecular features that are important to the biological activity of interest, thus confirming or contradicting mechanisms of action or suggesting further experiments. The techniques of SAR have been applied for more than 20 years to the development of pharmaceuticals (drug design) and agricultural chemicals. More recently, the methods have been used to study structure–activity relations of compounds showing chemical and genetic toxicity.

The most desirable way to predict activities of untested compounds is to exploit the molecular level theory of action or mechanism. However, this knowledge is not yet available for many classes of biologically active compounds, and so, this approach is often not possible. Another approach is to use correlative methods to look for relationships between the molecular structures of tested compounds and their experimentally observed biological activities. Given a set of compounds that have been tested in a standard bioassay, SAR methods can be used to seek correlative relationships within the data. This approach might attempt to simultaneously take into account the processes of uptake, transport, distribution, metabolism, cell penetration, binding, excretion, etc. by correlating the structure of the administered compounds with the final observed biological activity. While this is an oversimplification, it often provides valuable information and can guide the direction of further study.

The discovery and design of drugs is an active and well-documented field.[1–6] While the techniques of SAR have been applied mostly to drug design, the same methods have been used in agricultural chemical applications and studies of toxic compounds as well. Several complementary approaches to SAR have been reported: (a) the correlative approach proposed by Hansch and co-workers[2,5,7,8] that provides quantitative statistical models, (b) methods relying on conformational analysis with graphical display for the intensive study of few molecules at a time,[9–11] (c) quantum mechanical methods that attempt to explain activity on the basis of electronic parameters,[12–14] and (d) qualitative methods relying on the generation of discriminants through multivariate statistics and/or pattern recognition.[15,16]

CORRELATIONS WITH SUBSTITUENT CONSTANTS—HANSCH ANALYSIS

The Hansch approach[2,7,8] to SAR studies has generated the most widespread interest in the past 20 years. This method is used to study congeneric sets of compounds. The physicochemical factors that govern the transport and receptor site interaction are usually factored into hydrophobic, electronic, and steric components. Each contribution is modeled with a substituent constant, and the activities of a set of congeneric compounds are fit to a multidimensional linear model of the form

$$\log(1/C) = -a(\log P)^2 + b\log P + c\sigma + dE_s + e \qquad (1)$$

where P is the 1-octanol/water partition coefficient, σ is the substituent electronic parameter of Hammett, E_s is the Taft steric constant for the substituent, and C is the molar concentration of the compound that elicits a standard biological response (such as EC_{50}, MED, or LD_{50}). The constants a, b, c, d, and e are fit to a set of data by multiple linear regression. In addition to the substituent constants above, many others have been reported. Hydrophobic parameters used include $\log P$ but also π and R_m. Hammett σ constants have been extended and generalized in many forms. Steric effects of substituents have been represented by molar refractivity (MR) and Verloop's STERIMOL constants.[17]

There are three fundamental assumptions of the correlational analysis approach. The first is that molecular properties related to biological activity can be separated and quantified. These molecular properties are assumed to be representable by physicochemical parameters, either measured or calculated. This approach grew out of the physical organic chemistry paradigm due to Hammett that was applied to chemical reactivity. The second important assumption is that the biological activity of interest can be measured quantitatively. It is implicit that the compounds of the study all have the same mode of action. The third assumption is that the relationship between the structures, represented by physicochemical pa-

0095-2338/85/1625-0296$01.50/0

rameters, and the biological activities can be described by a simple mathematical relationship. The linear model has been used most often. Statistically based methods are used, so confidence intervals and standard errors are available to characterize the models generated.

These assumptions are best fulfilled when the set of compounds under investigation has the following characteristics. The compounds should be structural analogues that act by the same mechanism. Reliable physicochemical substituent parameters must be available for all of the compounds. The data set should be large enough so that statistical methods are applicable. The variations in potency should be greater than the measurement errors. The physicochemical parameters should have substantial variation across the set of compounds, but collinearity among the physicochemical parameters must be avoided. These factors have been discussed in detail by Martin[2] and Unger,[7] among others.

CONFORMATIONAL ANALYSIS AND DISPLAY

Biological activity often arises as a consequence of interactions with a biological receptor. A number of drug–receptor complexes have been characterized by X-ray crystallography, the most accurate way of quantifying geometry. If the biological response due to drug–receptor interactions is dependent on physical interactions between the molecules, then it is logical to develop methods for viewing these interactions. For example, Langridge and co-workers[11] have developed a computer-based color graphics system that allows a chemist to interactively work with macromolecule–ligand interactions. When a receptor is assumed but not known, an alternative method is to investigate the drug molecules and attempt to deduce important features directly from their structures. Comparisons of the volumes occupied by active analogues compared to inactive ones can be done to deduce the topography of the receptor site.[9] In the absence of X-ray coordinates for the drug molecules, three-dimensional models can be built by molecular mechanics.[18-20] Modeling systems that incorporate quantum mechanical considerations have also been used in SAR of drugs (e.g., reference 21).

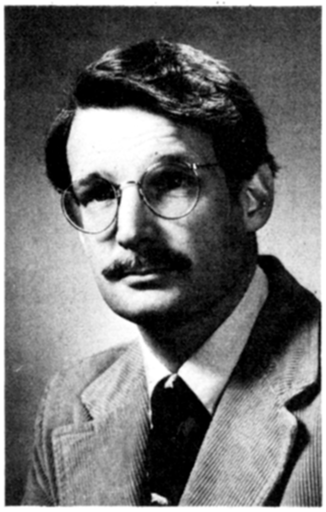
QUANTUM MECHANICAL APPROACHES

Quantum mechanical methods have been applied to SAR studies for many years.[12-14,21] Two approaches have dominated the work in this area. The first is the generation of electronic parameters to be used in correlation analysis along with the substituent constants.[22] The objective is to codify as fundamental a level of information as possible in order to elucidate the mechanism of action.[23,24] The second is the use of quantum mechanical methods in conjunction with conformational analysis of biologically active molecules to be compared with antagonists or mimic compounds.[25]

DISCRIMINANT DEVELOPMENT AND PATTERN RECOGNITION

In our laboratory we have developed an approach to SAR that involves pattern recognition and discriminant development methods. This approach allows the study of structure–activity problems for which the correlative approach or the conformational analysis approach is not feasible. Some characteristics of such data sets include the following: large number of compounds, perhaps hundreds; substantial structural diversity; qualitative or semiquantitative biological activity data.

The steps that are taken in performing a structure–activity study of this type are as follows (see Figure 1): (a) Input and store a set of molecular structures for which the property or biological activity of interest is available. (b) Generate three-dimensional molecular models. (c) Develop molecular structure descriptors for each of the members of the data set. The descriptors can be derived directly from the stored topological representations of the structures, or they can be derived from the three-dimensional molecular models. (d) By pattern recognition or statistical methods, develop classifiers or mathematical models to discriminate between the classes of data on the basis of the sets of molecular structure descriptors. (e) Test the predictive ability of these discriminant functions or models with compounds of unknown property values. (f) Systematically focus on which of the molecular



Peter C. Jurs is Professor of Chemistry at The Pennsylvania State University. He received his B.S. in chemistry from Stanford University in 1965 and his Ph.D. in chemistry from the University of Washington in 1969. He then joined the faculty of The Pennsylvania State University, where he has been Professor of Chemistry since 1978. He served as Program Director for Chemical Analysis in the Chemistry Division of the National Science Foundation during 1983–1984. Jurs' research interests include the application of pattern-recognition and statistical methods to chemical and biologically related problems. He has done research in analytical data interpretation with pattern-recognition methods, structure–activity studies of biologically active compounds such as pharmaceuticals and genotoxic compounds, carbon-13 NMR chemical shift prediction, and a number of other computer application areas in chemistry.

Terry Stouch received his Ph.D. in chemistry from The Pennsylvania State University in the summer of 1985. His B.S. in biochemistry was awarded at the same institution in 1980. He is a member of the American Chemical Society, its Medicinal and Computational divisions, Phi Lambda Upsilon, Phi Kappa Phi, and the Society of Environmental Toxicologists and Chemists. Stouch's research has included computer-aided studies of structure–activity relationships, the development of computer software for chemical applications, the investigation of pattern-recognition methodology and applications, molecular modeling studies, and the synthesis of potential antiinflammatory agents.

Maria Czerwinski received her B.S. degree in chemistry from The Pennsylvania State University in 1985. Her undergraduate research interests emphasized the structure–biological activity relationships of antiinflammatory steroids with the ADAPT software system. She will enter graduate school in the fall of 1985 to continue research in QSAR, molecular modeling, and drug design. Career goals also include medical school.

Javier N. Narvaez is a graduate student in chemistry at The Pennsylvania State University. He received his B.S. in chemistry from The State University of New York at Stony Brook in 1982. His present research activity is in the area of structure–activity relationship studies of biologically active compounds. In addition to the study of sister chromatid exchange, he is also involved in a collaborative study of musk odorants.

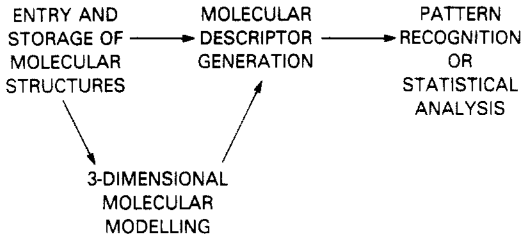

ENTRY AND STORAGE OF MOLECULAR STRUCTURES
MOLECULAR DESCRIPTOR GENERATION
PATTERN RECOGNITION OR STATISTICAL ANALYSIS
3-DIMENSIONAL MOLECULAR MODELLING

Figure 1. Schematic diagram of the steps involved in a structure–activity study.

structure descriptors are the most useful in developing the separating discriminant functions.

This approach to SAR studies has been implemented in an interactive computer software system called ADAPT (Automated Data Analysis using Pattern recognition Techniques.[15,26] The ADAPT system currently consists of more than 90 modular FORTRAN IV routines that implement all the steps necessary to perform an SAR study. It is designed to be run on a time-sharing minicomputer and requires guidance and interaction by the user throughout the SAR study. The ADAPT system can be viewed as a set of software tools with which to perform SAR experiments.

The system has been implemented in modules. This provides the user with the freedom to choose sequences of operations in a flexible way suitable to the problem at hand. It allows the developers and/or users to develop and use new modules without having to alter other parts of the system. For example, the development of a new descriptor generation routine can be done with no knowledge of either the structural storage or descriptor analysis parts of the system.

The system's flexibility allows the user to attack his SAR problem with whichever set of software tools he finds are appropriate. A set of compounds for which a quantitative biological activity variable is available can be studied by multiple linear regression analysis to form quantitative models for prediction of level of activity. A set of compounds tagged by activity class (e.g., active, weakly active, inactive) can be studied by using discriminants to separate the classes.

The heart of the ADAPT system is the set of descriptor development modules. They fall into four classes: topological, geometrical, electronic, and physicochemical.

Topological. These descriptors are derived from the connection table representation of the structure and include such information as atom and bond counts, substructure counts, molecular connectivity, and substructure environment descriptors.

Geometrical. These are derived from the three-dimensional molecular models and include such information as the principal moments of inertia, molecular volume, and surface area.

Electronic. These are quantities characterizing the structure with partial charges, dipole moments, bond strengths, etc. derived either from del Re σ charge computations or extended Hueckel calculations.

Physicochemical. Parameters such as log P (the logarithm of the partition coefficient of a compound between water and 1-octanol) are often correlated with biological activity. Descriptors that are either estimations of such quantities or calculable measures that are simply related to the physicochemical descriptors can be used.

ADAPT includes a routine that allows the user to input any additional descriptors from outside the system. This allows the user to study laboratory data (e.g., retention indices, spectroscopic data) along with the calculated descriptors.

The analysis portion of ADAPT has routines implementing statistical methods and pattern recognition methods of a variety of types. Linear and nonlinear regression analysis are available. In the pattern recognition area, all of the major types of pattern-recognition methods are available: mapping and display; discriminant development; clustering; class modeling.

Once a set of data is prepared for analysis, the user can easily analyze the exact same set of data with a wide variety of techniques while seeking the most effective analysis. Just as with the descriptor development routines, new analysis routines can be inserted into the system with no need to be concerned about the structure of the remainder of the system.

PATTERN RECOGNITION IN STRUCTURE–ACTIVITY RELATION STUDIES

A number of studies of the application of pattern recognition to the problem of searching for relationships between molecular structure and biological activity have been reported. The types of activity studied include pharmaceuticals (drug design), agricultural chemicals, chemical communicants (olfactory stimulants), and toxicity (chemical toxicity and mutagenic and carcinogenic activity). A large fraction of this type of research is involved with the generation of appropriate descriptors from the molecular structures available. Early applications of pattern recognition to drug design have been reviewed by Kirschner and Kowalski.[16] This area of application of pattern recognition is briefly mentioned by Varmuza.[27] A book describing one approach to SAR research has appeared.[15]

A few representative, published SAR studies are as follows: a study of 200 drugs for anticancer activity,[28] a study of 9-anilinoacridines for antitumor selectivity,[29] studies of drugs of accepted therapeutic value,[30] structure–carcinogenic potential,[31] olfactory quality of organic compounds,[32] structure–carcinogenic potential of PAH.[33]

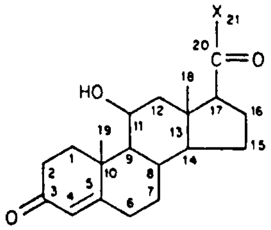
In order to give some examples of the types of SAR studies that can be done with the pattern-recognition approach, the following presents three example studies that are in progress. They involve a study of antiinflammatory steroids, a study of retinoids for prevention of cancer, and an investigation of the mutagenic activity of compounds as measured by a sister chromatid exchange assay. These projects are included because they indicate the range of problems that are accessible to pattern-recognition studies in SAR, but they are incomplete and should be considered as examples rather than definitive studies.

SAR of Antiinflammatory Steroids. Currently, we are investigating the SAR of some antiinflammatory steroids by pattern-recognition methodology. These compounds are interesting for several reasons: they have pharmaceutical applications, they all have a similar backbone, this backbone is reasonably rigid, and there are a good deal of data available in the literature. These last three reasons make these compounds very attractive as a data set for testing methodology. A good deal of SAR work has been done on this class of compounds;[34] however, few of these methods have involved the application of pattern recognition, and fewer still utilized linear discriminants. The data that was used for the studies reported here were taken from one such linear discriminant study.[35] This paper listed the names and activities of 122 steroids, most of which were assayed in the McKenzie–Stoughton human vasoconstrictor test. In that previous study, the authors achieved 100% correct classification of the 122 compounds by using a set of 22 descriptors. Those results appeared very promising, and we are attempting to extend and verify that study.

Recent work has shown that the levels of random correct classifications increase dramatically as the number of variables in a study increases.[36] The simple variation in the structures and the intuitive approach taken by the authors of the steroid study support their approach, but the results of the random classification studies indicate that for the 122 compounds divided into two classes, 74/78, and described by 22 variables the random classifications could range around 85%–90%. While the results of the previous steroid study were better than this, it is conceivable that the clustering of a few very similar structures could lead to this increase in classifications. The fact that random effects could account for such high levels of classifications does not necessarily invalidate those results, but it does mean that further verification of those results is warranted.

A large number of variables were used to develop the final discriminant in the previous study. A different variable was used to describe the occurrence of each different substituent at each position on the steroid backbone. However, the chemistry and binding of these compounds is determined not by the identity of the substituents but rather by their physicochemical effects—their size, shape, lipophilicity, electronic properties, and chemical reactivity. The approach that we have taken was to develop variables that would code for these effects at the various sites of substitution. In this way, we hoped to reduce the number of variables that were required to explain the SAR and generate discriminants that had a lower probability of being due to random chance.

In the first stage of analysis, in order to avoid the effects of conformational variability, the 122 compounds were separated into those that were and those that were not unsaturated at the 1,2-position:



Many more unsaturated compounds were available (28–93), so these were used in the first study. The vasoconstrictor data were quantitative, ranging from zero to almost 2000. The 93 compounds were divided into two classes: an active class containing 58 compounds and an inactive class containing 35. The classification criterion was the same as in the previous study: the activity of hydrocortisone 17-butyrate, which had an activity of 50. Any compounds that did not have a listed quantitative activity were excluded from consideration, even though the other authors assigned them as either potent or nonpotent. Their reasoning for doing this was not clear, and we thought it best to use only those data for which actual experimental data were available. This reduced the number of compounds in the first study to 73, of which 28 were nonpotent and 47 were potent.

Descriptor generation was done as follows. The five positions on the ring having structural variations were 6, 9, 16, 17, and 21. The hydrophobic substituent constant for each substitution was either obtained from tables or calculated by the method of Hansch and Leo.[37] The substitutions at positions 17 and 21 were more complex than those at the other positions and ranged from methyl groups to seven-member oxygenated chains. From a cursory examination of the data, the length and complexity of these chains were found to be very important to activity. In order to include that information, simple molecular connectivity indices[38] were calculated: path 0 valence and path 1 valence. Electronic effects could be important in the transport and binding of these compounds, and these were coded by using some simple del Re σ charges. The most positive and negative charges that occurred in the substituents at positions 17 and 21 were calculated. This brought the total number of descriptors to 12.

The first pattern recognition trials were performed with all 12 of these descriptors. By use of an initial discriminant that was generated by a parametric method based on the Bayes theorem,[39] an iterative adaptive least-squares method[40] generated a linear discriminant that classified above 95% of the patterns correctly. A nearest-neighbor clustering approach also showed promising results with classification around 80%. The nearest-neighbor approach yields high classifications only when the classes are well separated. Many examples can readily be given where a linear discriminant would yield better classifications than this simple approach. Other clustering results showed loose clusters of compounds of like activity.

Table I. Final Set of Six Descriptors Used in the Steroid Study

descriptor	position	relative ranking[a] no. wrong	% correct	rank
log *P*	6	6	92	3
molecular connectivity				
path 1 valence	17	5	93	2
	21	3	96	1
σ charge				
most negative	17	34	54	6
	21	8	89	4
most positive	21	16	78	5

[a] Classification results when that element of the discriminant is removed. Rank is 1 (least important) to 6 (most important).

This 12-variable set and the linearly separable set of compounds were subjected to variance feature selection[41] in order to eliminate useless variables and see if these classifications could be obtained with a subset of the 12 descriptors. One variable was eliminated at a time until classification results decreased appreciably after the elimination of six of the original 12. By use of the remaining six variables, a discriminant was generated that could correctly classify all but three of the 75 compounds, an overall 96% classification. The misclassified compounds had activities that were relatively close to the arbitrary cutoff of 50. One had an activity of 90, and another had an activity of 85. The third had an activity of 16, which was higher than all but a few of the inactive compounds. The ratio of descriptors to observations and the distribution of patterns between the classes could account for less than 75% random classifications. Nearest-neighbor classifications were between 87% and 94%, and cluster analysis showed loose clustering within the two classes. Leave-two-out internal validation achieved 96% correct classification of the excluded compounds. These results all support the validity of the final discriminant. A list of the final six descriptors and their relative ranking is presented in Table I. While the only sure test of a discriminant is the correct prediction of the activities of unknowns, these results are encouraging.

The pattern-recognition results presented here support those presented previously. The physicochemical variables used should code much the same information that was inherent in that previous study, but do so in a more parsimonious manner. The descriptors used here have an additional advantage over those used in the previous pattern-recognition study. In that study, no compound could be used if it contained a unique substituent because the descriptors were simple substructure counts. The use of physicochemical parameters allows the inclusion of a wider variation of substituents. This could provide information valuable to a truly useful SAR.

These results are preliminary. Those compounds that were not unsaturated at the 1,2-position must still be examined. Also, closer examination of the misclassified compounds indicates that the description of the steric factors at positions 17 and 21 was probably not adequately coded. Future work will extend this study in both of those directions.

SAR of Retinoids for Cancer Prevention. Retinoids, vitamin A analogues, elicit a variety of biological responses including the prevention of some forms of cancer in laboratory animals and the prevention in vitro of malignant transformation of cells.[42] These observations have led to an extensive effort to

Table II. The 152 Compounds Used in the Retinoid Study

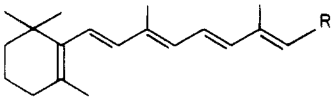
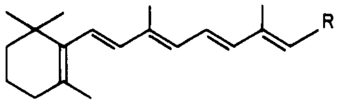
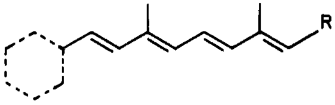
			
no.	activity	R	other modifications
1	+	COOH	
2	+	CH_3	
3	+	$CONC_6H_5COOC_2H_5$	
4	+	COOH	3-demethyl
5	+	$COC(CH_3)_3$	
6	+	CO_2	8,9-didehydro
7	+	CH_2SCH_3	
8	+	$CHCH(CN)_2$	
9	+	$CH_2OCH_2CH_2CH_3$	
10	+	$CH_2OCH_2CH_3$	
11	+	$CH_2NCOC_6H_5$	
12	+	$CH_2NCOOCH_2CH_3$	
13	+	CH_2OH	
14	+	CH_2OCOCH_3	
15	+	CH_2OCH_3	
16	+	CHO	
17	+	$CHNNHCOCH_3$	
18	+	CHNOH	
19	+	$CH_2NHCOCH_3$	
20	+	$COOCH_3$	
21	+	$COOCH_2CH_3$	
22	+	$CONCH_2CH_3$	
23	+	COOH	1′,2′-dihydro; 2′,3′-didehydro
24	+	$COCOCH_3$	1′,2′-dihydro; 2′,3′-didehydro
25	+	COOH	1′,2′-dihydro
26	+	$CO_2CH_2CH_3$	2-F
27	+	$CO_2CH_2CH_3$	4-F
28	+	$CO_2CH_2CH_3$	4,5-cis; 4-F
29	+	$CO_2CH_2CH_3$	6-F
30	+	CO_2	1,6-epoxide
31	+	CO_2CH_3	3′-oxo
32	+	$CH_2OC_6H_5OCH_3$	
33	+	CONHCONHOH	
34	+	CONHCO-aziridinyl	
35	+	$CNNHCOC_6H_5$	
36	+	$CONHC_6H_5pOH$	
37	+	$CONHC_6H_5pOCOCH_3$	
38	+	$CONHC_6H_5pOCOCH_2CH_3$	
39	+	$CONHCH_2CH_2OH$	
40	+	$CONHCH_2CH_2CH_3$	
41	+	$CONH(CH_2)_3CH_3$	
42	+	$CHC(COCH_3)_2$	
43	+	$CHC(COCH_2CH_3)_2$	
44	+	$CHC(COOCH_3)_2$	
45	+	CH_2NHCH_3	
46	+	$CHN(CH_2)_3CH_3$	
47	+	$CONHCH_2COOH$	
48	+	$CONHCH_2CH_2OCH_2CH_2CH_3$	
49	+	CHO	3-demethyl
50	+	CHCHCO	
51	+	CHCHCOOH	
52	+	CO_2CH_3	1′,2′-epoxide
53	+	CH_2OCOCH_3	1′,2′-epoxide
54	+	CO_2CH_3	1′,2′-epoxide; 6,7-epoxide
55	+	retinoyl-L-pantolactone	
56	+	retinoylimidazole	
57	+	2-retinylidene-1,3-cyclohexanedione	
58	+	2-retinyl-5,5-dimethyl-1,3-dioxin	
59	+	1,3-cyclopentenedione adduct of retinol	
60	+	barbituric acid adduct of retinol	
61	+	isopropylidene malonate adduct of retinol	
62	+	2-retinylidene-1,3-cycloheptanedione	
63	+	5′,6′-epoxide of 5,5-dimethyl-2-retinylidene-1,3-cyclohexanedione	
64	−	CH═CH—COOH	3-demethyl; 2-methyl
65	−	CH_2CH_3	
66	−	$CHCHCO_2CH_3$	
67	−	CH_2OCOCH_3	8,9-didehydro
68	−	$CH_2OCH_2C_6H_5$	
69	−	CH_3	2-methyl
70	−	$COCH_3$	
71	−	$CH_2O(CH_2)_3CH_3$	
72	−	COOH	8,9-dihydro
73	−	COOH	6,7-dihydro

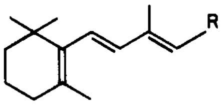
Table II (Continued)



no.	activity	R	other modifications
74	–	COOH	4,5-dihydro
75	–	COOH	2,3-dihydro
76	–	CH$_2$OC$_6$H$_5$	
77	–	CONHC(CH$_3$)$_3$	
78	–	CHCHCOCH$_3$	
79	–	CHC(COCH$_2$CH$_2$CH$_3$)$_2$	
80	–	C(CH$_3$)$_2$OH	
81	–	CHCHCOCH$_2$CH$_3$	
82	–	COCH$_2$CH$_3$	
83	–	CH$_2$OCCH	
84	–	*cis*-COOCH$_3$	4-COOCH$_3$
85	–	*cis*-CO	4-CH$_2$OH
86	–	CH$_2$SC$_6$H$_5$	
87	–	COCH$_2$SOCH$_3$	
88	–	CH$_2$SOCH$_3$	
89	–	*N*-retinylphthalimide	
90	–	retinyl-3,4-methylenedioxyphenyl ether	
91	–	2-retinylidene-1,3-cyclopentanedione	



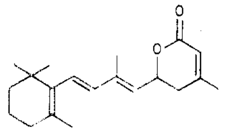
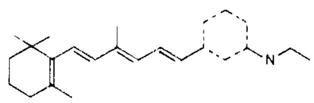
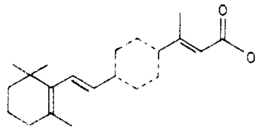
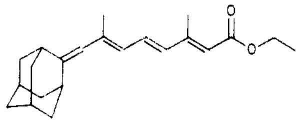
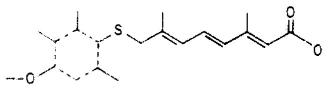
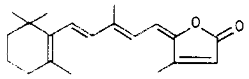
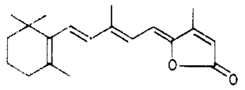
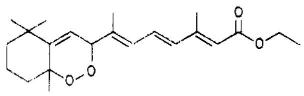
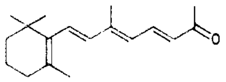
no.	activity	R	other modifications	ring substitutions
92	+	COOH		2,3,6-trimethyl; 4-OCH$_3$
93	+	COOH		2,3,6-trimethyl
94	+	CONH$_2$		2,3,6-trimethyl; 4-OCH$_3$
95	+	COOH		3,6-dimethyl; 4-OCH$_3$; 2-Cl
96	+	COOCH$_2$CH$_3$		3,6-dimethyl; 4-OCH$_3$; 2-Cl
97	+	CO$_2$CH$_2$CH$_3$	2-F	2,3,6-trimethyl; 4-OCH$_3$
98	+	CO$_2$CH$_2$CH$_3$	4-F	2,3,6-trimethyl; 4-OCH$_3$
99	+	CO$_2$CH$_2$CH$_3$	3-CF$_3$	2,3,6-trimethyl; 4-OCH$_3$
100	–	COOCH$_2$CH$_3$		2,3,6-trimethyl; 4-OCH$_3$
101	–	COOCH$_2$CH$_3$		2,3,6-trimethyl; 4-OH
102	–	COOH		2,3,4,5,6-pentamethyl; 4-OCH$_3$
103	–	CONHC$_2$H$_5$		2,3,6-trimethyl; 4-OCH$_3$
104	–	CONHC$_2$H$_5$	8,9-dihydro	2,3,6-trimethyl; 4-OCH$_3$
105	–	COOH	4,5-dihydro	2,3,6-trimethyl; 4-OCH$_3$
106	–	COOH		2,3,6-trimethyl; 4-OC$_2$H$_5$
107	–	COOC$_2$H$_5$		2,3,6-trimethyl; 4-OC$_2$H$_5$
108	–	COOH		2,6-dimethyl; 3,5-diethyl; 4-OCH$_3$
109	–	COOC$_2$H$_5$		2,3,6-trichloro; 4-OCH$_3$
110	–	COOC$_2$H$_5$		2,3-dimethyl; 4-OCH$_3$; 6-chloro
111	–	COOC$_2$H$_5$		2,6-dichloro; 4-OCH$_3$
112	–	COOH		2,4,6-trimethyl; 5-chloro
113	–	CH$_2$OH		2,3,6-trimethyl; 4-OCH$_3$
114	–	CH$_2$OCH$_3$		2,3,6-trimethyl; 4-OCH$_3$
115	–	CH$_2$O(CH$_2$)$_3$CH$_3$		2,3,6-trimethyl; 4-OCH$_3$
116	–	CO-*N*-morpholinyl		2,3,6-trimethyl; 4-OCH$_3$
117	–	CO$_2$CH$_3$	6-F	2,3,6-trimethyl; 4-OCH$_3$
118	–	CO$_2$CH$_3$	8-F	2,3,6-trimethyl; 4-OCH$_3$



no.	activity	R
119	–	CH═CH—(3-NCOCH$_3$—C$_6$H$_4$)
120	–	CH═CH—COOH
121	–	CH═CH—CO—CH$_2$OH
122	–	CH═CH—C(CH$_2$)—CH$_2$OH
123	–	CHO
124	–	CH═CH—CHO
125	–	CH═CH—CH═CH—CHO
126	–	CH═CH—(2,3-didehydrobutyrolacton-3-yl)
127	–	*o*-cyano-C$_6$H$_4$

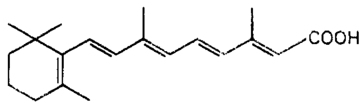
no.	activity	name or structure
128	–	abscisic acid
129	–	4-methyl-2,3,6-trimethylphenyl analogue of 10-fluoro-13-*cis*-retinoic acid
130	–	6-hydroxy-7-methyl-9-(2,6,6-trimethyl-1-cyclohexen-1-yl)-2,4,7-nonatrienoic acid
131	–	2,4,5-trimethylthienyl analogue of retinoic acid ethyl amide
132	–	1,3-indanedione adduct of retinal

Table II (Continued)

no.	activity	name or structure
133	–	2-norbornenyl analogue of retinoic acid ethyl ester
134	–	juvenile hormone I
135	–	juvenile hormone II
136	–	juvenile hormone III
137	–	mycophenolic acid
138	–	2,4,5-trimethylthienyl analogue of retinoic acid ethyl ester
139	–	3,7-dimethyl-9-(2-methyl-3-thienyl)-2,4,6,8-nonatetraenoic acid
140	–	3,7-dimethyl-9-(3-methyl-2-thienyl)-2,4,6,8-nonatetraenoic acid
141	–	3,7-dimethyl-9-(4,5-dimethyl-3-thienyl)-2,4,6,8-nonatetraenoic acid
142	+	9-(5′,5′-dimethyl-2′-acetyl-1-cyclopentenyl)-2,4,6,8-nonatetraenoic acid
143	+	9-[5′,5′-dimethyl-2′-(1-methoxyethyl)-1-cyclopentenyl]-2,4,6,8-nonatetraenoic acid
144	+	
145	+	
146	–	
147	–	
148	–	
149	–	
150	–	
151	–	
152	–	

develop an effective agent for the prevention of common forms of cancer.[43] During the course of this program, hundreds of new retinoids were synthesized and tested in a well-established in vitro assay. This assay, the hamster tracheal organ culture assay, measures the ability of a compound to control epithelial cell differentiation, and its results are thought to correlate well with cancer prophylaxis. The assay data for many of these compounds are available through a publication of the National Cancer Institute.[43]

The structure of retinoic acid is



The compounds developed in this program were modified at the ring, the unsaturated chain, and the polar terminus. Some of the compounds were so altered that they bore little resemblance to retinoic acid and could be considered retinoids only in the broadest sense. This structural variation made the data untractable for SAR methods, which require homologous series of compounds. Because of this, most of the SAR that has been done has involved little more than visual examination of the data. We thought that pattern-recognition analysis might be valuable in the investigation of the SAR of these data. The nature of the experimental data also supported this approach. Each compound was assayed at several concentrations from 10^{-6} to 10^{-11} M. The activity of the compounds was listed as "active" or "inactive" for each concentration tested. Such semiquantitative data are readily amenable to the classification methods of pattern recognition.

The preliminary investigation reported here was performed in an attempt to identify structural features important to the activity of the retinoids. The compounds were divided into two classes: those active at 10^{-8} M and those inactive at this concentration. Only compounds that were assayed at this concentration were used. This provided a set of 152 compounds, 75 active and 77 inactive, which are listed in Table II. Descriptors were developed that coded for the structural variations at the three positions indicated above. Some coded for variations in the conjugation in the conjugated chain or for the presence of substituents in this intermediate portion

Table III. Set of 21 Descriptors Used in the Retinoic Study

	SS	relative ranking[b]		
descriptor	no.[a]	no. wrong	% correct	rank
no. of aromatic bonds		34	76	17
no. of nitrogen atoms		31	78	16
molecular connectivity				
path 3		21	85	11
cluster 3		42	70	20
substructure				
simple count	1	3	96	3
environment[c]	2	30	79	15
	3	17	88	8
	4	15	89	6
	5	12	91	4
	6	15	89	7
	7	14	90	5
	8	27	81	14
	9	34	76	18
	10	19	87	9
	11	19	87	10
	12	26	82	13
	13	3	98	2
path[c]	14	3	98	1
	15	24	83	12
	16	60	58	21
	17	39	73	19

[a] Substructure number—see Figure 2 for the identity of the substructures. [b] Classification results when that element of the discriminant is removed. Rank is 1 (least important) to 6 (most important). [c] See text.

of the molecule. Others coded for variations in the polar terminus. Some of these latter variables also coded for variation in electron density at that location. Changes in the ring portion were present in the data set, and descriptors were developed that coded for this variation, also. Substructural descriptors were used heavily in order to account for the wide structural variation. A total of approximately 50 descriptors was generated.

After systematic screening of the pool of descriptors described above, a set of 21 was found that provided a data space in which 94% of the 152 compounds could be correctly classified by a linear discriminant. Internal validation results from a leave-10-out approach classified 86% of the left-out compounds correctly. Table III shows these descriptors, and Figure 2 contains the substructures used. Two whole molecule molecular connectivity indices were used as were two simple fragment counts (number of aromatic bonds and number of nitrogen atoms). The remaining 17 descriptors were based on substructures. One of these was simply the number of occurrences of the substructure. Four were counts of the number of paths that originated at a pseudostructure composed of the substructure and its nearest-neighbor atoms. These variables encode information pertaining to the complexity of the region surrounding the substructure. The remaining 13 substructural descriptors encoded similar information about the environment of the substructures. The pseudostructure was formed as above, but path 1 molecular connectivity indices were calculated for the pseudostructure. These variables encode information pertaining more to the immediate environment than do the path count descriptors.

Some of the substructures in this set are very general and code for the presence and surroundings of simple alkyl chains. Several dealt with various carbonyl moieties, which, in most of the compounds, occurred at the polar terminus. The structural environment of benzene rings was also used. Benzene rings were often found in place of the cyclohexene moiety of retinoic acid.

The work on the retinoids is still in progress. Current work involves the evaluation of the validity of the discriminant and the 21-descriptor set. Also, new variables are being investigated that may allow for the correct classification of all of the 152 compounds. More rigorous studies will then be built on these results. Future work includes further subdivision of the retinoids into finer classes of activity and the inclusion of more compounds. Another future study involves the SAR of the toxicity of the retinoids. Those compounds that show the greatest anticancer potential also tend to be the most toxic. Since the most useful pharmacological agents are those with high activity and low toxicity, a computerized system for the evaluation of toxicity as well as activity would be valuable for evaluation of the pharmacological potential of the retinoids.

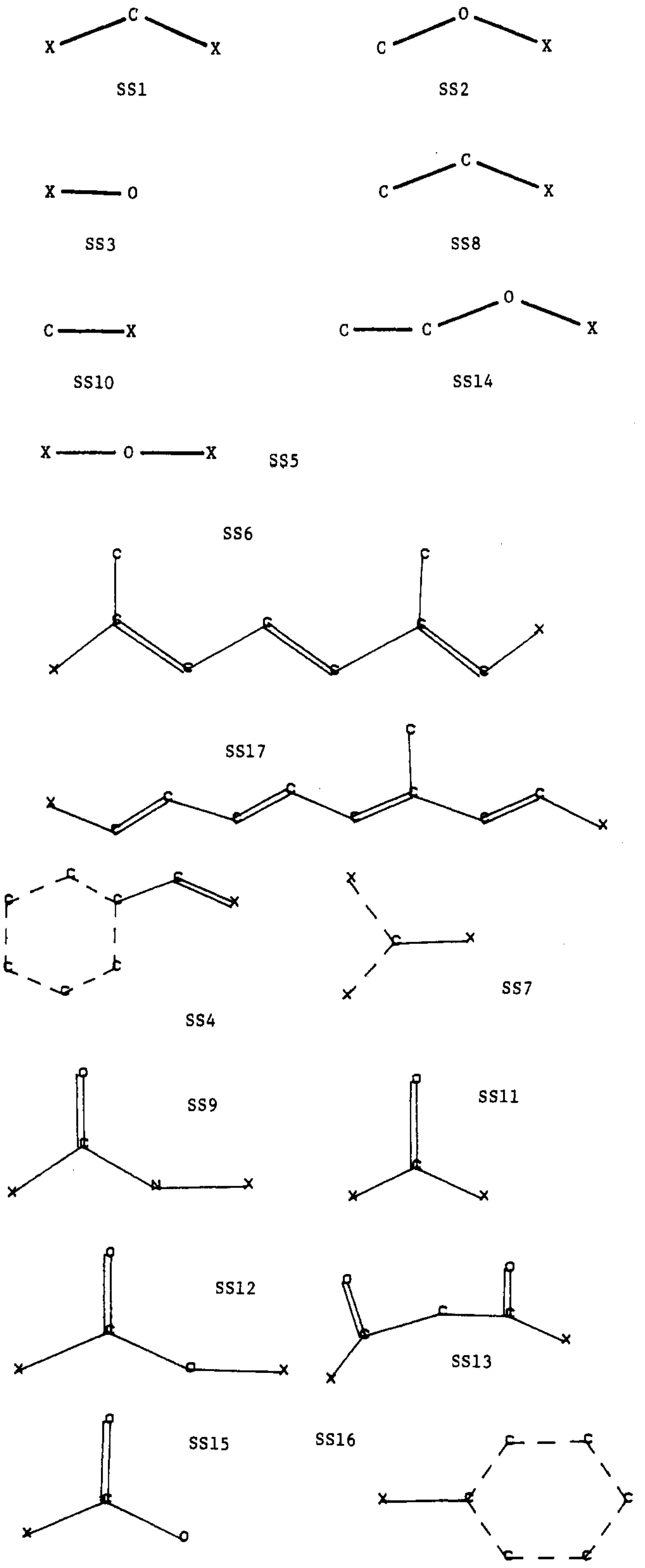


Figure 2. Substructures used for the development of environment descriptors in the retinoid study.

Table IV. The 120 Compounds Used in the SCE Study

compound	CAS Registry No.	class[a]	test system[b]
acetone	67-64-1	−	t, DON
(acetylamino)fluorene	53-96-3	−	t, CHO
acrylamide	79-06-01	−	v, M (DDY)
aminopyrene	58-15-1	−	t, DON
anthracene	120-12-7	−	v, CH
barbital	57-44-3	−, P	t, DON
benzene	71-43-2	−	t, HL
bilirubin	635-65-4	−	t, HL
butanol	71-36-3	−	t, CHO
N-(*n*-butyl)urethane	591-62-8	−	t, DON
ϵ-caprolactone	502-44-3	−	t, DON
cyclophosphamide	50-18-0	−	t, HL
dichlorvos	62-73-7	−, P	t, HL
(dibutylmethyl)phenol	30587-81-6	−	t, DON
diethylnitrosamine	55-18-5	−	t, HL
diethylstilbestrol	56-53-1	−	t, DON
dimethyl sulfoxide	67-68-5	−	t, CHO
ethanol	64-17-5	−	t, HL
hydroxyurea	127-07-1	−	t, V79
maleic hydrazide	123-33-1	−, P	t, CHO
6-mercaptopurine	60-44-6	−	t, A(T1)C1-3
methanol	67-56-1	−	t, CHO
3-methylcholanthracene	56-49-5	−	t, V79
N-methylurea	598-50-5	−	t, DON
4-*O*-(methyltetradecanoyl)phorbol-13-acetate	57716-89-9	−	t, V79
ozone	10028-15-6	−	v, CH
perylene	198-55-0	−	t, V79
phenanthrene	85-01-8	−	t, V79
propanol	71-23-8	−, P	t, CHO
psoralen (8-methoxy)	298-81-7	−	t, HL
pyrene	129-00-0	−	v, M (C3H)
quinoline	91-22-5	−	t, DON
acetic acid	127-09-3	−	t, HL
sodium dehydroacetate	4418-26-2	−	t, DON
Δ^9-tetrahydrocannabinol	1972-08-3	−	t, HF
tilorone	27591-97-5	−	t, A(T1)C1-3
toluene	108-88-3	−	t, HL
xylene	1330-20-7	−	t, HL
acetaldehyde	75-07-0	+	t, HL
2-aminofluorene	153-78-6	+, P	t, CHO
acetoxy-2-(acetylamino)fluorene	6098-44-8	+	t, HF
N-hydroxy-2-(acetylamino)fluorene	53-95-2	+	t, CHO
acridine orange	65-61-2	+	t, V79
adriamycin	23214-92-8	+	t, HL
aflatoxin B1	1162-65-8	+	t, CHO
alkeran	148-82-3	+	t, HL
aminoacridine C829	65094-73-7	+	t, HL
aminoacridine C846	63710-43-0	+, P	t, HL
4-aminoquinoline oxide	2508-86-3	+	t, DON
aniline	142-04-1	+	t, DON
ascorbate	50-81-7	+	t, CHO
5-azacytidine	320-67-2	+	t, A(T1)C1-3
7,12-dimethylbenz[*a*]anthracene	57-97-6	+	t, DON
7-methylbenz[*a*]anthracene (7-MBA)	2541-69-7	+	t, CHO
trans-1,2-dihydro-1,2-dihydroxy-7-MBA	64521-13-7	+	t, CHO
trans-3,4-dihydro-3,4-dihydroxy-7-MBA	62641-70-7	+	t, CHO
trans-8,9-dihydro-8,9-dihydroxy-7-MBA	64521-15-9	+	t, CHO
benzo[*a*]pyrene (B[*a*]P)	50-32-8	+, P	t, CHO
trans-7,8-dihydro-7,8-dihydroxy-B[*a*]P	57404-88-3	+	t, CHO
trans-9,10-dihydro-9,10-dihydroxy-B[*a*]P	58886-98-9	+	t, CHO
bredinin	50924-49-7	+	t, L5178Y
5-bromodeoxyuridine	59-14-3	+	t, HL
busulphan	55-98-1	+	t, HL
N′-(*n*-butyl)-*N*-nitrosourea	869-01-2	+	t, DON
N′-(*n*-butyl)-*N*-nitrosourethane	6558-78-7	+	t, DON
N-(*n*-butyl)urea	592-31-4	+	t, DON
caffeine	58-08-2	+, P	t, HF
carofur	24998-17-2	+	t, HL
chlorambucil	305-03-3	+	t, HL
chlorpromazine	50-53-3	+	t, V79
chlorpropamide	94-20-2	+	t, V79
chlorprothixene	113-59-7	+	t, V79
daunomycin	20830-81-3	+	v, M (C57B1/6)
dibromomannitol	488-41-5	+	v, M (C57B1/6)
dibutylamine	111-92-2	+	t, DON
diepoxybutane	1464-53-5	+, P	t, CHO
3,3-dimethyl-1-phenyltriazine	7227-91-0	+	v, M (CBA)

Table IV (Continued)

compound	CAS Registry No.	class[a]	test system[b]
dimethyl sulfate	77-78-1	+	t, HF
diphenyl	92-52-4	+	t, DON
ethyl methanesulfonate	62-50-0	+	t, HL
N-ethyl-*N'*-nitro-*N*-nitrosoguanidine	4245-77-6	+	t, HLB
ethylnitrosourea	759-73-9	+	t, CHO
gentian violet	548-62-9	+	v, CE
hoechst 33258	23491-45-4	+	t, CHO
hycanthone	3105-97-3	+	t, A(T1)C1-3
isopropyl methanesulfonate	926-06-7	+	t, HL
malathion	121-75-5	+	t, HF
(methylazoxy)methanol acetate	592-62-1	+	t, DON
melphalan	148-82-3	+	t, A(T1)C1-3
methotrexate	59-05-2	+	t, A(T1)C1-3
methylene blue	61-73-4	+	t, V79
2-methyl-4-(dimethylamino)azobenzene	54-88-6	+	t, DON
methyl methanesulfonate	66-27-3	+, P	t, HL
N-methyl-*N'*-nitro-*N*-nitrosoguanidine	70-25-7	+	t, V79
methylnitrosourea	684-93-5	+	t, DON
mitomycin C	50-07-7	+	t, DON
neutral red	553-24-2	+	v, MM
4-nitro-*o*-phenylenediamine	99-56-9	+	t, CHO
2-nitro-*p*-phenylenediamine	5307-14-2	+	t, CHO
nitrogen mustard	51-75-2	+	t, CHO
4-nitroquinoline 1-oxide	56-57-5	+	t, HL
procarbazine	366-70-1	+	v, CH
proflavin	92-62-6	+, P	t, CHO
promazine	58-40-2	+	t, V79
propane sulfone	1120-71-4	+	t, DON
β-propiolactone	57-57-8	+	t, DON
pyridine	110-86-1	+	t, DON
quinacrine mustard	4213-45-0	+	t, HL
RO-10	54350-48-0	+	t, HF
sodium benzoate	532-32-1	+	t, DON
saccharin	81072	+, P	t, DON
streptonigrin	3930-19-6	+	v, M (C57B1/6)
12-*O*-Tetradecanoylphorbol-13-acetate	16561-29-8	+	t, V79
thio-TEPA	52-24-4	+	t, A(T1)C1-3
thymidine	50-89-5	+	t, CHO
trenimon	68-76-8	+	t, V79
tris(2,3-dibromopropyl) phosphate	126-72-7	+	t, V79
urethane	51-79-6	+	t, DON
vitamin A acid	302-79-4	+	t, HF
vitamin A palmitate	79-81-2	+, P	t, HF

[a] (−) SCE negative; (+) SCE positive; P, member of prediction set. [b] t, in vitro [Abbreviations used: A(T1)C1-3, a Chinese hamster cloned line; CHO, Chinese hamster ovary; DON, Chinese hamster diploid lung; HF, human fibroplast; HL, human lymphocytes; L5178V, mouse lymphoma; V79, Chinese hamster lung fibroblast]; v, in vivo [Abbreviations used: CE, chicken embryo; CH, Chinese hamster; M, mouse (strain in parentheses); MM, mudminnow].

SAR of Chemical Mutagens. The development of short-term tests for carcinogenicity is an active field receiving a great deal of attention. A starting point for these tests lies in the assumption that a mutagenic event is a prerequisite for neoplasia. The sister chromatid exchange (SCE) assay is one of these tests. It is readily performed, and it denotes changes at the DNA level.[44] An SCE is an expression of the exchange of segments between the two chromatids of a chromosome.[45] The process presumably involves DNA breakage and reunion, although little is known about its molecular basis.[46]

A great many compounds have been tested with the SCE assay. This has prompted a pattern recognition based SAR study. The main challenges of the study are provided by the extreme structural diversity of the compounds and the lack of knowledge about the number and type of mechanisms involved.

The structures and SCE activities used in this study cam efrom a 1981 report of the GENE-TOX Program in which the effects of 163 substances on SCE frequencies are reviewed.[46] As the data are drawn from 216 publications, many compounds have results from different laboratories and for varying test conditions. The authors present tables for negative, marginal, and positive SCE-tested compounds, and many compounds appear in more than one table. Since these ambiguously classified compounds are too numerous to ignore, general criteria were set up to place them into two classes. In vitro studies were given priority over in vivo which were given priority over those requiring metabolic activation. After removing compounds that could not be classified definitely as positive or negative, or whose structure could not be handled by the ADAPT system, 120 compounds remained. They are listed in Table IV. 13 compounds were randomly selected to form a separate prediction set. The remaining compounds comprised a training set of 73 active (mutagens) and 34 inactive compounds (nonmutagens).

The structures were entered into the ADAPT system and modeled by a strain-minimizing procedure. A total of 122 descriptors, falling into the following categories, were generated: (1) geometric, (2) electronic, (3) physicochemical, and (4) topological (this last class is potentially the largest and includes fragment counts, substructure counts, and substructure environment descriptors, among others). Descriptors that had less than 10% nonzero values in either the active or inactive class were discarded. Descriptors with unacceptably high multiple linear correlations with other descriptors were eliminated. Of the remaining descriptors available, no more than 35 were considered at a time. Before being used in the pattern-recognition analysis, the descriptors were autoscaled.

For each set of descriptors under consideration, an adaptive least-squares discriminant development algorithm[40] was used to classify the members of the training set. Next, the variance method of feature selection[41] was used to remove descriptors that did not contribute significantly to the classification. In a typical training run, this procedure was repeated until no descriptors could be eliminated without excessive deterioration of the classification power of the discriminant.

Initially, 54 descriptors were examined. These were taken 35 at a time and checked for high multiple correlations. As highly correlated descriptors were discarded, new ones were considered until all 54 had been checked. This procedure left 18 descriptors for pattern-recognition analysis. The adaptive least-squares and variance feature selection attained best results with the nine descriptors called DS1 in Table V (Figure 3). All but nine of the training set members were correctly classified with this discriminant, a classification score of 91.6%.

Internal consistency checks were performed by excluding the misclassified compounds and forming 10 randomly generated training set/prediction set pairs from the remaining compounds. All 10 training sets were completely separated. Of the compounds in the leave-10-out sets, 87% were correctly classified.

As a more stringent test of the validity of the results, the discriminant was used to classify the 13 members of the independent prediction set, achieving a prediction score of 69%. Considering the class sizes in the prediction set, 57% correct classification would be expected by a purely random procedure. However, the predictive power is 23% less than the training set results, which required investigation.

The descriptors used to develop the discriminant were examined. A projection plot of the two largest principal components was generated, and it showed that many compounds were grouped together by simple chemical properties that, by themselves, are probably not intrinsically related to SCE activity. This two-dimensional plot showed clusters of aromatic compounds, as well as clusters of small, intermediate, and large-size molecules. These nine descriptors seem to code for size and aromaticity as well as SCE activity.

The effects of this clustering were investigated further as follows. The compounds of the training set were redistributed into two classes according to size with the 35 molecules having 10 or fewer non-hydrogen atoms in one class and the 72 larger molecules in the second class. A discriminant was trained with the nine descriptors of DS1 for the molecular size classes, and 100% correct classification resulted. A similar experiment was performed for aromaticity, yielding a classification score of 94.4%. The set of descriptors in DS1 apparently had significantly more discriminatory power in terms of molecular size and aromaticity than SCE mutagenic activity. Moreover, it is not clear whether the original training set score was a result of encoding activity or fortuitous structural differences among the classes. If structural differences exist between the biological activity classes, then pattern recognition may mistakenly focus on these differences as a way to maximize the classification results. This hinders the proper and thorough description of biological activity and may seriously hamper predictive power. It is interesting to note that of the 13 prediction compounds only two were classified as inactives. These two were propanol, the smallest compound of the group, and benzo[*a*]pyrene, the only compound whose bonds were all aromatic. Benzo[*a*]pyrene was incorrectly classified.

To investigate possible causes of these results, the individual classification power of each of the descriptors was determined. This was done for the original biological activity classes, as well as for the size and aromaticity classes, and also for a random assignment of class membership. Seven of the nine descriptors in DS1 had greater classification power in terms of size or aromaticity than for SCE activity. Having the individual classification power allows us to select those descriptors that show greater classification power for the SCE activity than for the size or aromaticity classes.

Table V. The 20 Descriptors Used To Form the Four Descriptor Sets and the Classification Results Obtained

descriptor[a]	descriptor set			
	DS1	DS2	DS3	DS4
no. of nitrogen atoms	0	0	0	1
no. of double bonds	1	1	0	0
no. of rings	1	0	0	1
total no. of interatomic paths	1	0	0	1
molecular volume (solvated)	1	0	0	0
intermediate/smallest moment of inertia	0	1	1	0
intermediate principal axis (Y)	1	0	0	0
smallest principal axis (Z)	0	1	1	0
ratio of X axis to Y axis	1	0	0	1
ratio of X axis to Z axis	0	0	1	0
ratio of Y axis to Z axis	1	0	1	1
structural symmetry index	0	1	1	0
sum of absolute values of σ charges	1	0	1	1
most negative σ atomic density	0	1	1	0
substructure count for SS 5	0	0	0	1
substructure count for SS 23	0	0	0	1
substructure count for SS 24	0	1	1	0
substructure count for SS 30	1	0	0	0
environment of SS 5	0	0	1	0
environment of SS 8	0	1	1	1

descriptor set	% correct				
	SCE	size	aromaticity	random	prediction
DS1	91.6	100.0	94.4	85.0	69
DS2	87.8	85.0	70.1	75.7	54
DS3	88.8	89.7	80.4	80.4	54
DS4	90.6	100.0	80.4	84.1	77

[a] Substructures used in descriptor development are shown in Figure 3.

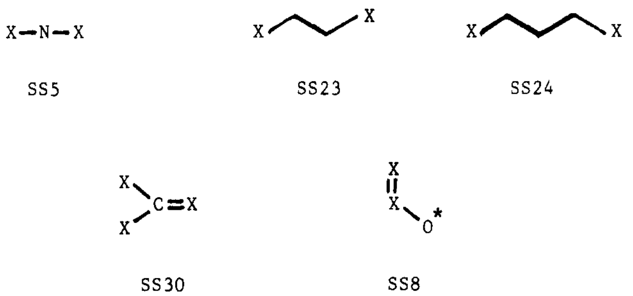


Figure 3. Substructures used for the development of environment descriptors in the SCE study.

At this stage of the study, 68 additional descriptors, mostly substructural or environmental, were generated. A preliminary pattern recognition analysis with only these descriptors was performed. The best result was 77% correct classification with eight descriptors.

To develop a discriminant from the 122 descriptors now available, only those that showed superior individual classification power for SCE activity over size and aromaticity were considered. This selection process left only 15 decriptors for the study. No descriptors had to be removed because of high multiple correlations. By use of the adaptive least-squares discriminant, these 15 descriptors could correctly classify 89.7% of the training set compounds. After variance feature selection, the best result was 87.8% with seven descriptors (DS2). This same set of descriptors gave 75.7% classification with random classes, 85.0% with the size-based classes, and 70.1% with aromatic classes. The classification score for the 13-member prediction set was 54%. A principal components plot showed no signs of clustering by activity or by any obvious structural parameter.

Although the dominance of size and aromaticity was successfully curtailed, the training and prediction scores decreased somewhat. This decrease may not be too significant as fewer descriptors were used, and the previous prediction score of 69% was apparently contingent on size and aromaticity biasing. Although the description of activity was perhaps more narrowly focused upon with this approach, no improvement in classification resulted. It is possible that the harsh criterion for choosing the initial set of descriptors discounted some descriptors that may have otherwise contributed favorably. Under this assumption, other training runs were performed that employed less strict methods of descriptor selection.

Two descriptors that showed good individual classification power, though not higher than that for size and aromaticity, were added to the descriptor set from the previous run. Training resulted in a discriminant with 10 descriptors and 88.8% classification (DS3). Scores for the random, size, and aromaticity classes were 80.4%, 89.7%, and 80.4%, respectively. The prediction score remained at 54%. Size classification had once again become higher than that for activity. Still, no improvement in activity classification was seen.

For the next test, the descriptors in the final discriminants of all four of the previous trainings were combined. After elimination of multiple occurrences and multiple correlations, 18 descriptors remained. These initial 18 descriptors correctly classified all but eight of the compounds, a score of 92.5%. Variance feature selection reduced the number of descriptors to nine, with a classification score of 90.6% (DS4). Classification with respect to the random and aromatic classes stayed low at 84.1% and 80.4%, respectively. Size classification, however, soared to 100%. The prediction score was the highest thus far at 77%. No further improvements along these lines have been obtained to date.

Evidently, size overdescription is very difficult to control without also affecting activity classification and prediction. This may point to a weakness in the data set and possibly a quirk in the randomly chosen prediction set. The compounds lacking SCE activity are nearly equally divided among small and large compounds, but the active compounds are predominantly large. However, this difference in the classes may not be chemically significant. The presence of 17 small compounds in the positive class, compounds such as acetaldehyde and dimethyl sulfate, attest that small size does not negate activity. The size difference between the classes may just be of statistical origin. There are more compounds in nature with many atoms than with few atoms. If so, it follows that there are more large than small compounds that exhibit a particular physicochemical property, and a corresponding biological activity. The problem is compounded when many of the larger molecules become active by being metabolically degraded to a few common, small species. In cases such as these, a size descriptor may have great classification power without being mechanistically relevant.

In the work thus far, considerable progress has been made in understanding the general problem. The results, however, still need to be improved. Future studies will address the heterogeneity of the data set. In addition, a study will be performed with a data set excluding compounds having marginal activity or requiring metabolic activation. Since the current prediction set has been used to test preliminary results, the final discriminant should be tested on a fresh prediction set. Ultimately, many problems in a pattern-recognition analysis can be overcome if one has a properly constituted data set. The introduction of more compounds, if and when they become available, would be a valuable contribution to this SCE study.

ACKNOWLEDGMENT

This research has been supported by the U.S. Environmental Protection Agency under Cooperative Agreement CR 807531. The contents do not necessarily reflect the views of the Agency, and no official endorsement should be inferred. The PRIME 750 computer used in these studies was purchased with partial financial support from the National Science Foundation.

REFERENCES AND NOTES

(1) Ariens, E. J. "Drug Design"; Academic Press: New York, 1971–1983; Vol. 1–10.

(2) Martin, Y. C. "Quantitative Drug Design. A Critical Introduction"; Marcel Dekker: New York, 1978.

(3) Dearden, J. C., Ed. "Quantitative Approaches to Drug Design". *Pharmacochem. Libr.* **1983**, *6*.

(4) Goldberg, Leon, Ed. "Structure-Activity Correlation as a Predictive Tool in Toxicology"; Hemispheres: Washington, DC, 1983.

(5) Topliss, J. G., Ed. "Quantitative Structure-Activity Relationships of Drugs". *Med. Chem. (Academic Press)* **1983**, *19*.

(6) Olsen, E. C.; Christoffersen, R. C., Eds. "Computer Assisted Drug Design"; American Chemical Society: Washington, DC, 1979.

(7) Unger, S. H. "Consequences of the Hansch Paradigm for the Pharmaceutical Industry". *Med. Chem. (Academic Press)* **1980**, *9*.

(8) Van Valkenburg, W. "Biological Correlations—The Hansch Approach"; American Chemical Society: Washington, DC, 1972.

(9) Humblet, C.; Marshall, G. R. "Three-Dimensional Computer Modeling as an Aid to Drug Design". *Drug Dev. Res.* **1981**, *1*, 409–434.

(10) Blaney, J. M.; Jorgensen, E. C.; Connolly, M. L.; Ferrin, T. E.; Langridge, R.; Oatley, S. J.; Burridge, J. M.; Blake, C. C. F. "Computer Graphics in Drug Design: Molecular Modeling of Thyroid Hormone–Prealbumin Interactions". *J. Med. Chem.* **1982**, *25*, 785–790.

(11) Langridge, R.; Ferrin, T. E.; Kuntz, I. D.; Connolly, M. L. "Real-time Color Graphics in Studies of Molecular Interactions". *Science (Washington, D.C.)* **1981**, *211*, 661–666.

(12) Christoffersen, R. E. "Quantum Pharmacology: Recent Progress and Current Status". In "Computer-Assisted Drug Design"; Olsen, E. C.; Christoffersen, R. E., Eds.; American Chemical Society: Washington, DC, 1979.

(13) Richards, W. G. "Quantum Pharmacology"; Butterworths: London, 1977.

(14) Bergmann, E. D.; Pullman, B., Eds. "Chemical and Biochemical Reactivity"; Israel Academy of Science and Humanities: Jerusalem, 1974.

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

(16) Kirschner, G. L.; Kowalski, B. R. "The Application of Pattern Recognition to Quantitative Drug Design". *Med. Chem. (Academic Press)* **1979**, *8*.

(17) Verloop, A. "The STERIMOL Approach: Further Development of the Method and New Applications". In "Pesticide Chemistry: Human Welfare and the Environment"; Doyle, P.; Fujita, T., Eds.; Pergamon Press: Oxford, 1983; Vol. 1.

(18) Allinger, N. L.; Yuh, Y. H. "Molecular Mechanics, Operating Instructions for MM2 and MMP2 Programs, 1977 Force Field". *QCPE* **1980**, *395*.

(19) Burkert, U.; Allinger, N. L. "Molecular Mechanics"; American Chemical Society: Washington, DC, 1982.

(20) Gund, P.; Rhodes, J. D.; Smith, G. M. "Three-dimensional Molecular Modeling and Drug Design". *Science (Washington, D.C.)* **1980**, *208*, 1425.

(21) Boyd, D. B. "Quantum Mechanics in Drug Design: Methods and Applications". *Drug Inf. J.* **1983**, *17*, 121–131.

(22) Osman, R.; Weinstein, H.; Green, J. P. "Parameters and Methods in Quantitative Structure-Activity Relationships". In "Computer Assisted Drug Design"; Olson, E. C.; Christoffersen, R. E.; Eds.; American Chemical Society: Washington, DC, 1979.

(23) Loew, G. H.; Ferrell, J.; Poulsen, M. "Mechanistic Structure-Activity Studies Using Quantum Chemical Methods: Application to Polycyclic Aromatic Hydrocarbon Carcinogens". In "Structure-Activity Correlation as a Predictive Tool in Toxicology"; Golberg, L., Ed.; Hemisphere: Washington, DC, 1983.

(24) Loew, G. H.; Poulsen, M. T.; Spangler, D.; Kirkjian, E. "Mechanistic Structure-Activity Studies of Carcinogenic Dialkylnitrosamines". *Int. J. Quantum Chem., Quantum Biol. Symp.* **1983**, *No. 10*.

(25) Alagona, G.; Ghio, C.; Kollman, P. "Bifurcated vs. Linear Hydrogen Bonds: Dimethyl Phosphate and Formate Anion Interactions with Water". *J. Am. Chem. Soc.* **1983**, *105*, 5226–5230.

(26) Ramiller, N. "Computer-Assisted Studies in Structure-Activity Relationships". *Am. Lab.* **1984**, *June*, 78.

(27) Varmuza, K. "Pattern Recognition in Chemistry"; Springer-Verlag: Berlin, 1980.

(28) Kowalski, B. R.; Bender, C. F. "The Application of Pattern Recognition to Screening Prospective Anticancer Drugs. Adenocarcinoma 755 Biological Activity Test". *J. Am. Chem. Soc.* **1974**, *96*, 916–918.

(29) Henry, D. R.; Jurs, P. C.; Denny, W. A. "Structure-Antitumor Activity Relationships of 9-Anilinoacridines Using Pattern Recognition". *J. Med. Chem.* **1982**, *25*, 899–908.

(30) Menon, G. K.; Cammarata, A. "Pattern Recognition II: Investigation of Structure-Activity Relationships". *J. Pharm. Sci.* **1977**, *66*, 304–314.

(31) Rose, S. L.; Jurs, P. C. "Computer-Assisted Studies of Structure Activity Relationships of N-Nitroso Compounds Using Pattern Recognition". *J. Med. Chem.* **1982**, *25*, 769-776.

(32) Jurs, P. C.; Ham, C. L.; Brugger, W. E. "Computer Assisted Studies of Chemical Structure and Olfactory Quality Using Pattern Recognition Techniques". *ACS Symp. Ser.* **1981**, *148*, 143-160.

(33) Norden, B.; Edlund, U.; Wold, S. "Carcinogenicity of Polycyclic Aromatic Hydrocarbons Studied by SIMCA Pattern Recognition". *Acta Chem. Scand., Ser. B* **1978**, *B32*, 602-608.

(34) Wolff, M. E. "Steroids and Other Hormones". In "Quantitative Structure-Activity Relationships of Drugs"; Topliss, J. G., Ed.; Academic Press: New York, 1983.

(35) Bodor, N.; Harget, A. J.; Phillips, E. W. "Structure-Activity Relationships in the Antiinflammatory Steroids: A Pattern-Recognition Approach". *J. Med. Chem.* **1983**, *26*, 318-328.

(36) Stouch, T. R.; Jurs, P. C. "Monte Carlo Studies of the Classifications made by Nonparametric Linear Discriminant Functions". *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 45-50.

(37) Hansch, C.; Leo, A. "Substituent Constants for Correlation Analysis in Chemistry and Biology"; Wiley: New York, 1979.

(38) Kier, L. B.; Hall, L. H. "Molecular Connectivity in Chemistry and Drug Research"; Academic Press: New York, 1976.

(39) Tou, J. T.; Gonzalez, R. C. "Pattern Recognition Principles"; Addison-Wesley: Reading, MA, 1974.

(40) Moriguchi, I.; Kopmatsu, K.; Matsushita, Y. "Adaptive Least-Squares Method Applied to Structure-Activity Correlation of Hypotensive N-Alkyl-N''-cyano-N'-pyridylguanidines". *J. Med. Chem.* **1980**, *23*, 20-26.

(41) Zander, G. S.; Stuper, A. J.; Jurs, P. C. "Nonparametric Feature Selection in Pattern Recognition Applied To Chemical Problems". *Anal. Chem.* **1975**, *47*, 1085-1093.

(42) Sporn, M. B.; Roberts, A. B.; Goodman, D. S., Eds. "The Retinoids"; Academic Press: New York, 1984.

(43) Newton, D. L.; Henderson, W. R.; Sporn, M. B. "Structure-Activity Relationships of Retinoids: Tracheal Organ Culture Assay of Activity of Retinoids"; Laboratory of Chemoprevention, Division of Cancer Cause and Prevention, National Cancer Institute: Bethesda, MD.

(44) Lindahl-Kiessling, K.; Bhatt, T. S.; Karlberg, I.; Coombs, M. M. "Frequency of Sister Chromatid Exchanges in Human Lymphocytes Cultivated with a Human Hepatoma Cell Line as an Indicator of the Carcinogenic Potency of Two Cyclopenta(*a*)phenanthrenes". *Carcinogenesis* **1984**, *1*, 11.

(45) Soper, K. A.; Stolley, P. D.; Galloway, S. M.; Smith, J. G.; Nichols, W. W.; Wolman, S. R. "Sister-Chromatid Exchange (SCE) Report on Control Subjects in a Study of Occupationally Exposed Workers". *Mutat. Res.* **1984**, *129*, 77.

(46) Latt, S. A.; et al. "Sister Chromatid Exchanges: A Report of the GENE-TOX Program". *Mutat. Res.* **1981**, *87*, 17.

Chemometrics and Distributed Software

D. L. MASSART* and P. K. HOPKE[†]

Farmaceutisch Instituut, Vrije Universiteit Brussel, Laarbeeklaan 103, B-1090 Brussels, Belgium

Received February 1, 1985

An extrapolation is made of trends in chemometrics and analytical chemistry. This leads to a picture of the analytical laboratory of 1990. The importance of software is underlined, and it is concluded that the present situation is not what it should be and that the entrance of publishers and/or learned societies could remedy that situation.

INTRODUCTION

The aims of chemometrics have been defined as follows:[1,2] "Chemometrics is the chemical discipline that uses mathematical, statistical and other methods employing formal logic (a) to design or select optimal measurement procedures and experiments, and (b) to provide maximum chemical information by analyzing chemical data."

If one eliminates words such as "mathematical", this definition could just as well be a definition of analytical chemistry itself. In fact, chemometrics from being a mere subdiscipline of analytical chemistry is evolving to be the basis of modern analytical chemistry to such a degree that many excellent analytical chemists are also excellent chemometricians, even if they do not recognize this fact. To show the all-pervading role of chemometrics in analytical chemistry, let us consider what the analytical laboratory of the future might be.

ANALYTICAL LABORATORY OF 1990

To extrapolate, one first must look at the present situation. So, let us look at the way analytical chemists proceed from the point where they have been given an analytical problem to the point where they deliver the information that they were seeking.

Figure 1 gives the main steps; namely, the development of a method, its execution, and the obtaining of information from the determination. This process is considered in some more detail in Figure 2. The very first problem facing the analytical chemist is to select a method. For example, let us suppose the analyst needs to determine a certain drug in blood. He will then need to decide first whether to use TLC, HPLC, or GLC. If, for instance, his choice is HPLC, he will then have to decide whether to try a reversed-phase or normal-phase column with UV, fluorometric, electrochemical, or post-column derivatization detection, etc. This process usually leads to the selection of an initial procedure, which is then optimized. In this step the analyst starts from initial parameter values (such as parameters describing the composition of a mobile phase, elution temperature, gradient, etc.) and by logical reasoning, which as we will show later can be formalized, arrives at the final optimal or, at least, acceptable values of the parameters.

Main step 2 of Figure 1 also can be divided into two steps (Figure 2). Usually, one needs to carry out some pretreatment of the sample, such as weighing, drying, extracting, etc., that is then followed by the actual determination. The pretreatment step is usually the more difficult one (at least in pharmaceutical or biomedical analysis). It involves the most time and cost and very often determines the quality of the final result in terms of precision and accuracy.

The third main step of Figure 1 starts with the data acquisition (i.e., the collection of data by computer). This step is followed by the transformation of the signal to chemical information. This information consists of a list of chemical identities and concentrations. Sometimes, this is the end result, but much more often this chemical information has to be translated into what could be called diagnostic or user information. Does the result of a patient's blood test indicate illness; does the result of the analysis of an air sample signify that the air that was sampled is polluted and that a certain industry has contributed to this in a significant way, etc.?

Now, let us see how a chemometrician views and contributes to this process and at the same time what tomorrow's analytical laboratory might be like. The first step is probably the most

[†] On leave from the Institute for Environmental Studies, University of Illinois, Urbana, IL.

0095-2338/85/1625-0308$01.50/0