Journal of Medicinal Chemistry

Volume 21, Number 10 October 1978

Articles

A Structure–Carcinogenicity Study of 4-Nitroquinoline 1-Oxides Using the SIMCA Method of Pattern Recognition

W. J. Dunn, III,*[1] and Svante Wold

Research Group for Chemometrics, Institute of Chemistry, Umeå University, S-901 87 Umeå, Sweden.
Received March 13, 1978

Structure–carcinogenicity data for a series of 4-nitro- and 4-hydroxyaminoquinoline 1-oxides were analyzed using the SIMCA method of pattern recognition. Using physicochemically based substituent constants to describe each compound, a principal components model was derived for the carcinogens. This model was 82% successful in predicting the carcinogenic potential of the compounds. For the 6-substituted compounds, a significant relationship between those structural parameters associated with carcinogenic potential and ability to stimulate unscheduled DNA synthesis was observed. In addition, other problems unique to the classification of carcinogens were discussed.

It is now widely recognized that many human cancers are of environmental cause.[2] The carcinogenic agents involved are often chemical compounds either naturally occurring or man-made. Hence, much human cancer might be prevented if early recognition were possible of the carcinogenic potential (CP) of chemical environmental contaminants.

Until recently the carcinogenicity of certain agents was recognized only after incidences of tumors in man had occurred as a result of exposure to these agents. Now this CP can also be assessed experimentally in a number of biological model systems, allowing extensive testing of suspected compounds. The most promising of these tests have recently been evaluated[3a] and reviewed.[3b]

In order to screen the large number of chemicals occurring in the environment within a reasonable time period, however, one would also need methods giving a fair prediction of the CP on theoretical grounds, i.e., without actually making and testing the compound. At present, predictions are based on a qualitative examination of the structure of the organic compound and the comparison of it with structures of known cancer-causing agents. Compounds suspected of being carcinogenic are then evaluated in biological models and whole animal tests. This strategy has revealed several new carcinogens but is time consuming and could be made more efficient if substances with low probabilities for causing cancer could be recognized. An extensive and thorough evaluation process, therefore, must involve some method of statistical data analysis aimed at prediction of the carcinogenic potential from the molecular structure. Such a statistical data analysis would involve a structure–activity relationship, the handling of which requires methods of pattern recognition (PaRC).

Various methods of PaRC such as the linear learning machine, linear discriminant analysis, and K nearest-neighbor approach have recently been applied to structure–activity problems.[4-9] In this report we wish to describe another such statistical structure–activity approach and illustrate its utility by analyzing a series of 33 4-nitro- and 4-hydroxyaminoquinoline 1-oxides for which carcinogenicity has been evaluated in aminal tests.

Carcinogenicity as a Classification Problem. In a recent publication our views concerning the scope of classification studies in relation to structure–activity problems were outlined[10] and applied to a study of β-adrenergic compounds. Applying the same methodology to the present example, we first describe each of the investigated compounds by structural parameters such as substituent parameter scales (Hammett, Hansch, and Taft) and steric parameters derived from molecular models and molecular mechanics. Then, regularities in these structural parameters are searched for within two classes of molecules—the class of carcinogens and the class of inactive compounds. These "regularities", the "patterns" of the classes, are determined from basic sets of substances of known classification, i.e., carcinogenic or noncarcinogenic, called the *training sets*. Objects of uncertain or unknown classification are thereafter placed in a *test set* and on the basis of the derived classification patterns, predictions are made regarding their class assignment, i.e., their carcinogenic potential. Ideally, this prediction will be in terms of statistical significance or probability.

Levels of Classification. In our earlier study[10] we identified four levels of classification that can result from pattern recognition studies. These are stated as I–IV.

I. Classification into either of a number of defined classes such as carcinogenic or noncarcinogenic.

II. Classification into either of a number of defined classes with the possibility of substances being members of neither of the classes, i.e., being outliers.

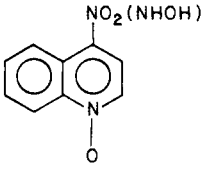
III. Classification as in II plus the measured level of activity of compounds in the classes being related to the position of the compounds in the class. In the present

0022-2623/78/1821-1001$01.00/0

investigation the desired result would be (a) prediction of a compound as being carcinogenic or not plus (b) quantification of the level of carcinogenic potency of the carcinogens in the training set.

IV. On level IV the values of *several* measured effect variables are related to the position of the objects in a particular class. For example, one may desire to relate the chemical structure of compounds to the abilities to initiate DNA repair synthesis, induce cell transformation, induce bacterial mutagenicity, and induce cancer. Studies at this level would yield information about the relevance of such processes to the carcinogenicity problem, not only about the predictive nature of such screens but also about the mechanistic significance of the screens and how these are related to induction of cancer. If there is more than one mechanism by which substances of a given type can be carcinogenic, which seems likely, the significance of each particular measured effect variable for each mechanism can possibly be assessed. In this way links between processes might be established.[11,12]

Present Study. Carcinogenicity measurements of a series of 4-nitro- and 4-hydroxyaminoquinoline 1-oxides were taken from several literature sources. The parent structure (I) for the series is given below. This series was



I

selected in part because of the structural homogeneity among the compounds in the literature. This made possible the description of the compounds in terms of tabulated physicochemically based "substituent" variables. In addition, the compounds are conformationally more or less rigid which reduces the complexity of the description problem.

Most of the substances had been tested in a consistent manner. Finally, since the data are structurally homogeneous a common mechanism of action may be operating.[13] The biological data, their sources, and the testing conditions are given in Table I.

Data. To describe the analogues of I we have chosen to parameterize each position of substitution in the quinoline nucleus with the variables π,[14] MR,[14] σ_m,[14] σ_p,[14] L, and B_4.[15] For positions 2 and 6 the additional Verloop steric constants B_1, B_2, and B_3[15] were used. Since there are 6 positions available for substitution in I, and the log P value for each analogue is included, a total of 43 variables was used. These are given in Tables I and II.

SIMCA Method. It has recently been shown that data (y) observed on compounds belonging to a single class of similar compounds, provided that certain continuity conditions regarding the data are met, can be described by a principal components (PC) model[15] (eq 1). Here y

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

are the data to be modeled, and m, b, and u are parameters that are determined to make the residuals e as small as possible. The indices in the model i and k refer to variable (structural descripter) and object (compound), respectively, while a refers to the specific component term in eq 1. This model has a simple geometric representation in an M-dimensional space (shown in Figure 1 in three-dimensional space for convenience) where each dimension corresponds

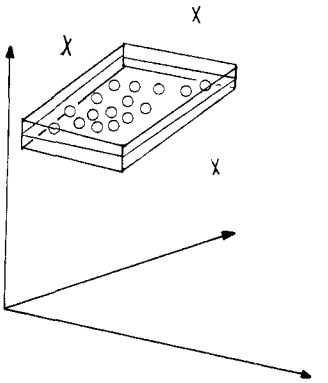


Figure 1. Principal components (PC) model for a class of objects: O = class members; X = outliers.

to one variable. The model as shown here represents the case of a PC model with two components ($A = 2$) in which the objects lie near or on a plane. The residual standard deviations (RSD) for each class can be calculated and the objects belonging to the class lie within 2 RSD values of the plane while outliers (denoted by X's) are outside the class structure. In our proposed classification scheme this represents classification at level II. SIMCA has the particular advantage in PaRC of always working at least on this level.

In a classification problem like the present one with several classes, one in principle describes each class by a separate PC model and the total analysis thus usually involves Q PC models for the Q classes (see, however, below).

There are two computational problems faced in this approach. For each class there is (1) determination of the number of components A_q required in the PC model to get an adequate description of the data and (2) evaluation of the model parameters m_i^q, b_{ia}^q, and u_{ak}^q where q denotes the qth class.

The number of components in the model is determined by a cross validation[16] procedure that minimizes the sum of the squared prediction errors, $(e_{ik}^*)^2$, where e_{ik}^* is calculated for data points, y_{ik}^*, that are deleted from the matrix before the PC analysis. A new component is added, if, after this process, the sum of prediction errors decreases. Therefore, the criterion for adding a new component to the PC model is whether the new component increases the predictive ability of the equation. This procedure has the advantage of extracting only the systematic part of the variance.

Once the number of components (A_q) is determined, the parameters m_i^q, b_{ia}^q, and u_{ak}^q are obtained. The details of these calculations have been published[16,17] and will not be further discussed here.

Classification by SIMCA. In the application of SIMCA to classification problems in which two or more classes of substances are analyzed, two situations can be recognized. The first is that in which all classes form well-defined and separated structures in the data space. This we refer to as a *symmetric* classification problem. In the other case, some of the classes have well-defined structures in the data space but the other classes contain no such structure and cannot be described by a PC model. This situation we shall refer to as an *asymmetric* classification problem. Both cases can be illustrated and are shown in Figure 2.

In the symmetric case, classification of an unknown can be made on the basis of its calculated distances from each class. These distances correspond to the RSD values for the unknown when fit to the class models. In struc-

Table I. Biological and Statistical Data for the 4-Nitroquinoline *N*-Oxides

compd	substituent	class	log P[f]	RSD[g] class 2	RSD[g] class 3	F statistic, class 2	activity: carc	activity: DNA[e]	u_i values: class 2 u_1	u_2	u_3	u_4	class 3, u_1
1	5-NO_2	1	0.95[a]	1.40		15.0	−[a]		2.19	4.05	1.96	1.19	
2	8-NO_2	1	0.76[a]	1.70		22.0	−[c]		0.82	0.41	0.04	−4.00	
3	2-*t*-Bu	1	1.07	0.80		4.80	−[b]		6.85	−7.18	2.14	1.98	
4	7-NO_2	1	0.45	1.20		11.0	−[c]		−3.49	−1.05	3.64	0.08	
5	2-Et,3-Me	1	2.56	1.00		8.10	−[b]		4.39	−4.45	1.18	1.17	
6	6-*t*-Bu (NHOH)	1	2.05	0.34	0.32	0.88	−[d]		−1.24	0.12	−3.42	1.14	−2.13
7	6-*n*-hexyl	1	3.26	0.55	0.12	2.20	−[d]	0	−1.55	0.14	−4.31	1.51	1.62
8	6-*n*-hexyl (NHOH)	1	3.35	0.55	0.11	2.20	−[d]		−1.55	0.14	−4.31	1.51	1.66
9	6-cyclohexyl (NHOH)	1	2.89	0.39	0.27	1.20	−[d]		−1.41	0.14	−3.90	1.33	−0.59
10	6-cyclohexyl	1	2.80	0.39	0.28	1.20	−[d]		−1.41	0.14	−3.90	1.33	−0.56
11	H	2	1.01[a]	0.43		1.40	+[c]	101	0.39	0.15	0.01	−0.43	
12	2-Me	2	1.25[a]	0.28		0.58	+[c]	88	3.60	−3.49	1.07	0.77	
13	5-Me	2	1.25[a]	0.44		1.50	+[c]	44	2.16	3.97	1.92	1.13	
14	6-Me	2	1.24[a]	0.30		0.69	+[c]	63	−0.14	0.14	−1.08	0.07	
15	7-Me	2	1.49[a]	0.53		2.10	+[c]	41	−2.53	−0.76	2.87	−0.12	
16	8-Me	2	1.48[a]	0.35		0.93	+[c]	5	1.05	0.55	0.06	−6.62	
17	6-*n*-Bu	2	3.11	0.39		1.20	+[d]	52	−0.81	0.15	−2.70	0.78	
18	6-*t*-Bu	2	3.06	0.40		1.20	++[a]	14	−1.24	0.12	−3.42	1.14	
19	6-NO_2	2	0.90[a]	0.25		0.49	++[a]		−0.37	0.14	−1.63	0.30	
20	6-Cl	2	1.40[a]	0.18		0.25	+[c]	42	−0.34	0.14	−1.50	0.26	
21	8-F	2	1.00[a]	0.61		2.80	++[a]		0.61	0.28	0.03	−2.27	
22	2-Et	2	1.76	0.26		0.51	+[c]		4.67	−4.69	1.44	1.16	
23	6,7-Cl_2	2	2.49	0.35		0.92	+[c]		−4.47	−1.15	2.41	0.77	
24	7-Cl	2	1.75	0.21		0.34	+[c]		−3.74	−1.13	3.92	0.07	
25	6-COOH	2	0.42	0.26		0.49	±[b]	2	−0.34	0.15	−1.58	0.28	
26	5-Cl	2	1.75	0.38		1.10	+[c]		2.69	5.12	2.50	1.62	
27	6-NO_2 (NHOH)	2	−0.28	0.25	0.49		+[c]		−0.37	0.14	−1.63	0.30	
28	6-*n*-Bu (NHOH)	2	2.10	0.39	1.20		+[d]		−0.81	0.15	−2.70	0.78	
29	3-Me	0	1.24[a]				+[a]	2	0.18	0.39	−0.06	−0.50	
30	3-Cl	0	1.33[a]				+[a]		0.17	0.40	−0.06	−0.50	
31	3-F	0	1.25				−[b]	4	0.39	0.15	0.01	−0.43	
32	3-Br	0	1.99				+[c]		0.37	0.54	−0.10	−0.55	
33	3-OMe	0	0.43				−[c]		0.08	0.49	−0.09	−0.53	
34	6-CF_3	0		0.23	0.41				−0.48	0.18	−2.30	0.54	
35	7-CF_3	0		0.56	2.30				−4.55	−1.38	4.68	0.18	

[a] See ref 22. The compounds from this source were tested using a variety of methods, strains of animals, and dosages. No quantitative ranking of the derivatives is possible. [b] See ref 26. The compounds from this source were tested using a variety of methods, strains of animals, and dosages. No quantitative ranking of the derivatives is possible. [c] See ref 27. The compounds reported from this source were assayed by dissolving or suspending them in propylene glycol (5 mg/mL) and injecting subcutaneously into the left groin of the mouse (normal ddN strain, 20 per test) in doses of 0.1 mL, the injections being repeated at the same site six times at intervals of 10 days. In the case of compounds that produced too severe local reactions, the dosage had to be cut down. Mice that developed tumors were recorded and at death an autopsy was performed. The compounds that produced tumors at the injection site were considered carcinogenic while those that did not after 300 days were considered noncarcinogenic. Most tumors formed were fibrosarcomas. [d] See ref 28. Testing was done as in ref 27. [e] See ref 29. [f] Unless noted estimated according to ref 24. [g] RSD = the residual standard deviation for the descriptors for that molecule when it is fit to the principal component model for the carcinogens (class 2). The residual standard deviation for the residuals of the descriptors for all members of the class when fit to class 2 = 0.36. The residual standard deviation for all members of class 3 when fit to the model for that class = 0.24.

Table II. Physicochemical Substituent Parameters

substituent	π^a	MR[a]	$\sigma_m{}^b$	$\sigma_p{}^b$	L^c	$B_1{}^c$	$B_2{}^c$	$B_3{}^c$	$B_4{}^c$
H	0.00	0.10	0.00	0.00	2.00	1.00	1.00	1.00	1.00
NO_2	−0.28	0.74	0.71	0.81	3.44	1.70	1.70	2.44	2.44
Cl	0.71	0.60	0.37	0.24	3152	1.80	1.80	1.80	1.80
t-Bu	1.98	1.96	−0.09	−0.15	4.11	2.59	2.86	2.86	2.97
Me	0.56	0.58	−0.06	−0.14	3.00	1.52	1.90	1.90	2.04
Et	1.02	1.03	−0.08	−0.13	4.11	1.52	1.90	1.90	2.97
F	0.14	0.09	0.33	0.15	2165	1.35	1.35	1.35	1.35
OMe	−0.02	0.79	0.10	−0.12	3.98	1.35	1.90	1.90	2.87
n-hexyl	3.32	2.89	−0.08	−0.15	8.22	1.52	1.90	1.90	5.87
cyclohexyl	2.51	2.67	−0.15	−0.22	6.17	2.04	3.16	3.16	3.49
n-Bu	2.13	1.96	−0.08	−0.16	6.17	1.52	1.90	1.90	4.42
Br	0.86	0.89	0.37	0.26	3.83	1.95	1.95	1.95	1.95
COOH	−0.32	0.69	0.36	0.44	3.91	1.60	1.60	2.36	2.66
CF_3	0.88	0.52	0.46	0.54	3.30	1.98	2.44	2.44	2.61

[a] Pomona College Medicinal Chemistry Data Bank. [b] M. Sjostrom and S. Wold, *Chem. Scr.*, 9, 200 (1976). [c] See ref 15.

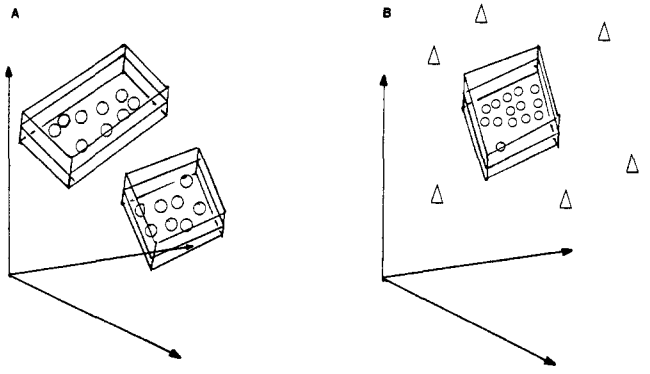


Figure 2. (A) Class structure for symmetric classification. (B) Class structure for asymmetric classification.

ture–activity studies such cases are common and may be encountered in attempts to separate and classify enzyme substrates from inhibitors or receptors agonists from antagonists,[10] for example.

The asymmetric case may also be frequently encountered in structure–activity studies. It results from the testing strategy and is the natural consequence of results that are reported as binary. The present case, where a substance is either carcinogenic or not, is an example where an asymmetric classification problem may be encountered. It is logical to assume that activity in this case may be very structurally specific whereas inactivity may result in any number of ways.

This corresponds, mathematically, to a well-defined structure of the active class, well described by a PC model, while the inactive class is spread "randomly" in the data space. This asymmetric classification has been used in the present application. In such cases a different strategy must be used for classification. Since one of the classes cannot be described by a PC model, a scheme must be constructed in which the classification can be based on the class which can be described by a PC model. Thus, a PC description of the active class is obtained on the basis of the training set of active compounds. A compound of unknown class assignment is then fit to the PC model. If its RSD is inside a confidence interval for the model as calculated from the training set of active compounds it is considered to be a member of that class. If it is not inside this confidence interval of the model it is considered not to be a member of that class. This is therefore an hierarchical scheme. It must be used with caution and can only be applied to classification of those compounds which are analogous to the training set.

Validation of Classification Results. One of the most important aspects of any classification study, irrespective of the method used, is a validation of the results. It has recently been reported[18] that, using the linear learning machine to classify a series of fire hazard compounds, a validation of the training set classification gave no better results than 68% even though the training set as a whole initially was classified 100% "correctly". Validation was done using the "leave one out" or jacknife method.

In this report classification is validated by leaving out a quarter of the training set and then classifying these compounds on the basis of the PC model derived from the remaining three quarters of the training set. The leaving out was then rotated to another quarter of the training set, etc., until each compound had been left out once and only once. A high prediction rate of the compounds left out in this case indicates stable class structure.

Elimination of Irrelevant Variables. In the present study each compound is described by 43 variables. To get an efficient classification a means must be found to eliminate those variables which do not contribute in defining class structure or do not participate in differentiating the classes. Various methods, so-called feature ranking and reduction[19] and prior feature reduction,[20] have been described and used in structure–activity studies. They are, however, usually aimed at the maximization of class separation and therefore cannot be used in the present situation where the initial number of variables exceeds the number of compounds. This would lead to an overoptimization of class separation. We have used instead information from the residuals as a basis for determining variable significance. The ability of a variable, i, to determine class structure we call modeling power, ψ_i, defined in eq 2. S_i is the residual standard deviation of the variable

$$\psi_i = 1 - S_i/S_{i,y} \qquad (2)$$

i, and $S_{i,y}$ that for y_i over all classes or the well-structured classes in the asymmetric case. If S_i becomes small compared to $S_{i,y}$, the PC model for the class predicts the variable's value well and, thus, a value of ψ_i near 1 implies good modeling power while a value near 0 implies low modeling power.

Variables were eliminated in the classification on the basis of low ψ_i. Such a basis for variable deletion in classification studies prevents a gross exaggeration of differences between classes since the variable deletion is not made in order to increase class separation but rather to enhance class description.

Results

Of the 33 compounds in the data set 18 had consistently given positive results in animal tests and ten had consistently given negative results. The mono-3-substituted quinolines (**29–33**) had given inconsistent results depending on the source. From examination of Table I it can be seen

Table III. b_{ia} Values[a]

	π_2	MR_2	σ_{m2}	σ_{p2}	L_2	B_{12}	B_{22}	B_{32}	B_{42}	MR_3	π_5	MR_5	σ_{m5}
m_i	−0.17	−0.17	0.12	0.09	−0.13	−0.16	−0.14	−0.14	−0.13	−0.20	0.20	−0.01	−0.13
b_{i1}	0.17	0.15	−0.21	−0.24	0.21	0.14	0.18	0.18	0.21	−0.04	0.19	0.15	0.06
b_{i2}	−0.18	−0.18	0.25	0.28	−0.24	−0.16	−0.21	−0.21	−0.25	0.05	0.42	0.33	0.14
b_{i3}	0.05	0.05	−0.07	−0.08	0.07	0.04	0.06	0.06	0.07	−0.01	0.21	0.16	0.07
b_{i4}	0.06	0.06	−0.08	−0.09	0.08	0.05	0.07	0.07	0.08	−0.01	0.17	0.13	0.06

	σ_{p5}	L_5	B_{45}	π_6	MR_6	L_6	B_{16}	B_{26}	B_{36}	B_{46}	π_7	MR_7	σ_{m7}
m_i	−0.17	0.00	−0.04	−0.29	−0.26	−0.23	−0.13	−0.23	−0.16	−0.22	0.22	0.03	−0.07
b_{i1}	0.02	0.16	0.14	−0.08	−0.10	−0.10	−0.16	−0.12	−0.13	−0.10	−0.36	−0.30	−0.19
b_{i2}	0.05	0.35	0.29	0.00	0.00	0.00	−0.01	−0.00	0.00	0.00	−0.11	−0.09	−0.06
b_{i3}	0.03	0.18	0.15	−0.18	−0.24	−0.22	−0.28	−0.24	−0.29	−0.26	0.34	0.30	0.16
b_{i4}	0.02	0.14	0.12	0.09	0.10	0.10	0.14	0.11	0.12	0.11	0.04	0.03	0.03

	σ_{p7}	L_7	B_{47}	π_8	MR_8	σ_{m8}	σ_{p8}	L_8	B_{48}
m_i	−0.13	0.06	0.00	0.20	−0.09	−0.13	−0.18	−0.05	−0.07
b_{i1}	−0.10	−0.33	−0.26	0.06	0.04	0.01	0.00	0.05	0.04
b_{i2}	−0.03	−0.10	−0.08	0.04	0.02	0.00	0.00	0.03	0.02
b_{i3}	0.08	0.31	0.26	0.00	0.00	0.00	0.00	0.00	0.00
b_{i4}	0.02	0.04	0.03	−0.56	−0.36	−0.04	0.06	−0.40	−0.39

[a] Subscript i refers to position of substitution.

Table IV. b_{ia} Values for PC Model for Class 3

	log P	π	MR	σ_m	L	B_4	σ_p	B_1	B_2	B_3
m_i	1.34	1.84	1.87	−0.73	1.73	1.61	−0.87	1.49	1.88	1.69
b_{i1}	0.34	0.32	0.21	0.06	0.54	0.57	0.08	−0.28	−0.14	−0.13

that they are either inactive or weakly active and could possibly constitute a class of their own. These compounds were placed in a test set and were not included in the analysis.

In the initial stages of the analysis, it was found that the inactive compounds did not form a homogeneous class with definite structure. The class of active substances, however, was shown to have substantial structure and from the principal components analysis a four-component model (A = 4) was found to best compounds the data. On the basis of low modeling power, eight variables were deleted and on the basis of the remaining 35 variables classification was carried out. The statistical data and model data are given in Table I and the variables used and the variable parameters are given in Table III.

By considering the classification problem to be an asymmetric one as discussed earlier, a hierarchical scheme was used in classifying the compounds in the two classes (Figure 3). F statistics were calculated for classification as a carcinogen or noncarcinogen based on the hierarchical scheme; for being in class 2, $F < (F_{22\,403;\alpha=0.01}) = 2.03$ for compounds included in the calculation of the class model and $F < (F_{31\,403;\alpha=0.01}) = 1.86$ for the other compounds (class 1 and test set). Comparison with the F statistics for each compound in Table I shows that of the noncarcinogens, compound **6** is classified as a carcinogen and compounds **9** and **10** are just inside class 2. Compounds **7** and **8**, the two 6-*n*-hexyl compounds, (F = 2.20 compared to F = 2.03 for their inclusion in class 2) are just outside class 2. The other inactive compounds are correctly predicted to be far outside of the active class.

Of the carcinogens (class 2 compounds) the 7-Me and 8-F compounds are predicted to be outside of their class. Therefore, based on this scheme $^{16}/_{18}$ or 89% of the active and $^{7}/_{10}$ or 70% of the inactive compounds are correctly predicted by the model. The classification was validated as previously described. Excluding the 6-substituted compounds, the verification resulted in 8-Me, 5-Me, 8-F, 7-Me, and 5-Cl being incorrectly classified. Based on χ^2 this result of $^{18}/_{23}$ being correct is significantly ($p \leq 0.01$) better than chance.

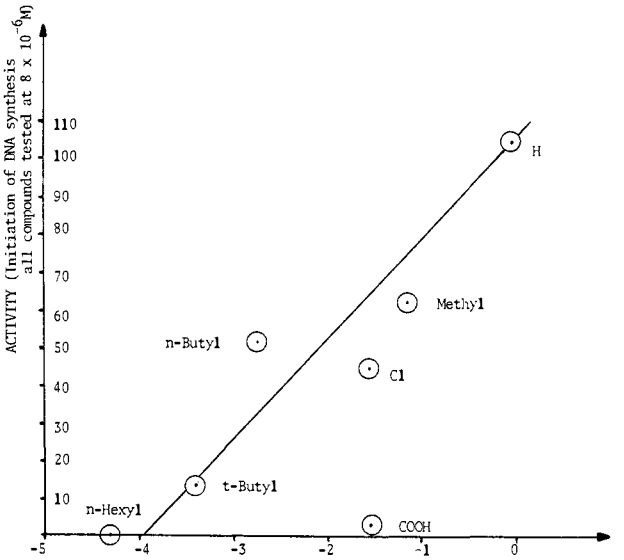


Figure 3. Plot of u_3 vs. relative potency for initiating DNA synthesis of 6-substituted compounds.

From our classification (Table I) it can be seen from the F statistics that the inactive 6-substituted compounds (**6–10**) are close to the active class (class 2). The two *n*-hexyl compounds are placed correctly just outside of class 2 while the 6-*tert*-butyl and the two 6-cyclohexyl compounds are classified as active. Hence, the PC model for the active class does not give reliable information about the inactive 6-substituted compounds. An inspection of the data suggests that a large substituent in the 6 position may make the compound inactive, even if it has the "right" pattern of substitution in other positions—these 6-substituted compounds form a special class. Using only variables specific for position 6, a one-component model described the five inactive 6-substituted compounds well, while another PC model well described the active 6-substituted compounds. All of the 6-substituted compounds were correctly classified using these two models. The classification was validated as discussed above with

the result that again all were correctly classified. The statistical data are given in Table IV.

Graphic Analysis of Model Parameters. The u_k values for the compounds derived from the PC model for a class indicate the position of the compounds in the space which describes the class. Within the class structure it is reasonable to assume that similar substances, e.g., strong carcinogens, will cluster. To relate the position of carcinogens in their class structure to their potency is difficult because of the nature of carcinogen data. Carcinogenicity is difficult to quantitate in the sense that a parameter such as an ED_{50} cannot be defined for the process. We have attempted instead to relate the position of the carcinogens as determined from the PC model for class 2 to their ability to initiate unscheduled DNA synthesis.[21]

For a number of compounds in this study this activity has been determined.[21] This test is a measure of the ability of a compound to damage DNA. This is considered by some to be the possible rate-limiting step in the mutagen/carcinogen process[3] and the existence of a relationship between the structural parameters that leads to a carcinogenic response and the initiation of DNA synthesis could be significant.

For the 6-substituted compounds for which this activity was available there is a significant relationship shown in Figure 3. For this subset of seven compounds the parameter u_3 from the carcinogen PC model is strongly correlated with the ability to initiate DNA synthesis except for the COOH analogue which deviates from the relationship. This may be due to our description of the substituent with parameters for the neutral form. There was no complete set of constants for the anionic form so this could not be tested. The other compounds for which this activity was available could not be included in the analysis, which indicates that in fact class 2 might be divided into subclasses.

It has been reported that a steric effect is responsible for the inactivity of the 6-alkyl compounds.[28] Examination of the b_3 values (b_{3i} for π_6, MR_6, etc., from Table III) for the compounds in Figure 3 shows that they are of the same sign and of similar magnitude. This shows that π, MR, and the Verloop steric constants contribute to the third component in the PC model for class 2 to about the same extent. The effect cannot be said to be solely a steric one but a complex combination of the steric and hydrophobic effects of the 6-substituent.

Summary of the SIMCA Analysis. Ten noncarcinogenic compounds (class 1) and 18 carcinogenic compounds (class 2) of the basic structure I were each described by 43 variables of physicochemical nature. A four-components PC model described 54% of the variation of 35 variables in the data of class 2. Eight variables contained mainly noise. No good PC model was obtained for class 1. A subset of class 1, inactive 6-substituted compounds, was well described by a one-component PC model in ten variables. A subset of class 2, active 6-substituted compounds, was well described by another one-component PC model in the same ten variables. The classification scheme used is shown in Figure 4. Validation of the scheme showed that this classification is statistically highly significant ($p < 0.01$). Correlations between the level of initiation of unscheduled DNA synthesis and the position of the 6-substituted compounds in the PC model for class 2 (35 variables, $A = 4$) were obtained.

Discussion

There have been few attempts to predict the CP of compounds using discriminant or classification methods. Okano et al.[22] have used a graphically determined discriminant function of ultraviolet absorption data and partition coefficient data in an attempt to distinguish carcinogenic from noncarcinogenic compounds similar to those in this study. Using a similar approach Morgan and co-workers[23] attempted to classify polycyclic aromatic hydrocarbons which were carcinogenic and noncarcinogenic. In both of these studies measured variables were used and the results were not validated.

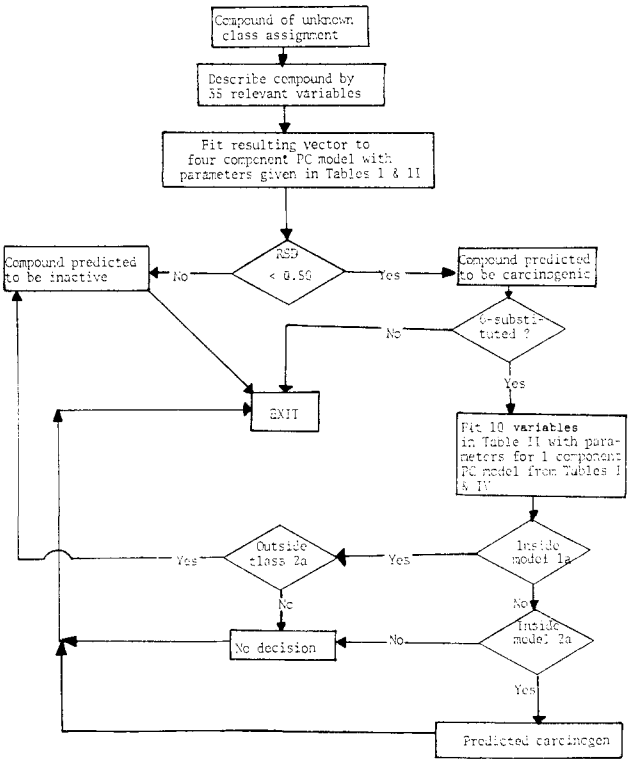


Figure 4. Hierarchical scheme for the classification of 4-nitroquinoline 1-oxides and 4-hydroxyaminoquinoline 1-oxides.

Our results indicate that PaRC methods can be used to assess the CP of chemical compounds on the basis of nonmeasured variables. The differentiation between carcinogens and noncarcinogens can be made on the basis of structural variables obtained from tabulated physicochemically based substituent constants. The experimental log P was included for a number of compounds in this study because they were available but this variable can be estimated from hydrophobic fragment constants.[24]

Used with the newly developed screens, such as initiation of DNA synthesis[21] or the bacterial mutagen tests,[25] the SIMCA method of PaRC can enhance the efficiency of detecting potential carcinogens. On the basis of training sets such as those derived in this study, the CP of an untested compound can be estimated and this information can be used in selecting compounds for more extensive testing. To illustrate this we have used the 6-CF_3 and 7-CF_3 compounds, neither of which to our knowledge has been tested. From Table I it can be seen that the 6-CF_3 compound is clearly predicted to be a member of class 2. From its u_3 value it is predicted to have moderate but significant activity similar to the 6-*n*-butyl or 6-Cl analogues in the DNA screen. The 7-CF_3 compound, on the other hand, is predicted to be outside of class 2 and therefore inactive. The 6-CF_3 analogue would be given priority over the 7-CF_3 compound for more extensive testing.

Such analyses as this, being based on similarity and analogy, are heavily dependent upon an adequate data base, i.e., experimental data of series of compounds with closely similar structure. For most classes of carcinogens

adequate data of this type are not available for analogy studies. We therefore feel that experiments should be performed to create such data.

Acknowledgment. We are grateful for support from the Swedish Natural Science Research Council and the Institute of Applied Mathematics, Stockholm.

References and Notes

(1) On leave from the Department of Medicinal Chemistry, College of Pharmacy, University of Illinois/Medical Center, Chicago, Ill.
(2) J. Higginson, Canadian Cancer Conference, Pergamon Press, Oxford, 1968, pp 40–75.
(3) (a) 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). (b) B. A. Bridges, *Nature (London)*, **261**, 195 (1976).
(4) K. H. Ting, R. C. T. Lee, G. W. A. Milne, H. Shapiro, and A. M. Gaurino, *Science*, **180**, 417 (1973).
(5) B. R. Kowalski and C. F. Bender, *J. Am. Chem. Soc.*, **96**, 916 (1974).
(6) A. J. Stuper and P. C. Jurs, *J. Am. Chem. Soc.*, **97**, 182 (1975).
(7) K. C. Chu, R. J. Feldman, M. B. Shapiro, G. F. Hazard, and R. I. Geran, *J. Med. Chem.*, **18**, 539 (1975).
(8) L. J. Soltzberg and C. L. Wilkens, *J. Am. Chem. Soc.*, **99**, 439 (1977).
(9) Y. C. Martin, J. B. Holland, C. H. Jarboe, and N. Plotnikov, *J. Med. Chem.*, **17**, 409 (1974).
(10) W. J. Dunn, III, S. Wold, and Y. C. Martin, *J. Med. Chem.*, **21**, 922 (1978).
(11) M. L. Weiner and P. H. Weiner, *J. Med. Chem.*, **16**, 655 (1973).
(12) H. Wold, "On the Transition from Pattern Recognition to Model Building in Mathematical Economics and Game Theory. Essays in Honor of Oskar Morgenstern", R. Henn and O. Moeschlin, Ed., Springer-Verlag, Berlin, 1977.
(13) S. Kondo, *Br. J. Cancer*, **35**, 595 (1977).
(14) C. Hansch, A. Leo, S. H. Unger, K. H. Kim, D. Nikiatani, and E. J. Lien, *J. Med. Chem.*, **16**, 1207 (1973).
(15) A. Verloop, W. Hoogenstraaten, and J. Tipker in "Drug Design", Vol. V, E. J. Ariens, Ed., Academic Press, New York, N.Y., 1977.
(16) S. Wold, *Pattern Recognition*, 8, 127 (1976).
(17) S. Wold and M. Sjostrom in "Chemometrics, Theory and Practice", ACS Symposium Series No. 52, B. R. Kowalski, Ed., American Chemical Society, Washington, D.C., 1977.
(18) C. P. Weisel and J. L. Fasching, *Anal. Chem.*, **49**, 2114 (1977).
(19) K. C. Chu, *Anal. Chem.*, **46**, 1181 (1974).
(20) A. J. Stuper, W. E. Brugger, and P. C. Jurs in ref 17.
(21) R. H. C. San and H. F. Stitch, *Int. J. Cancer*, **16**, 284 (1975).
(22) T. Okano, J. Maenosono, T. Kano, and I. Onado, *Gann*, **64**, 227 (1973).
(23) D. D. Morgan, D. Warshowsky, and T. Atkinson, *Photochem. Photobiol.*, **25**, 31 (1977).
(24) G. G. Nys and R. F. Rekker, *Eur. J. Med. Chem.*, **9**, 361 (1974).
(25) J. McCann, E. Choi, E. Yanasaki, and B. N. Ames, *Proc. Natl. Acad. Sci. U.S.A.*, **72**, 5135 (1976).
(26) J. S. Paul, P. Montgomery, and J. B. Louis, *Cancer Res.*, **31**, 413 (1971).
(27) Y. Kowazoe, M. Tachibano, M. Araki, and W. Nakahara, *Biochem. Pharmacol.*, **16**, 631 (1967).
(28) Y. Kowazoe, M. Araki, and W. Nakahara, *Chem. Pharmacol. Bull.*, **17**, 544 (1969).
(29) H. F. Stitch, R. C. H. San, and Y. Kowazoe, *Nature (London)*, **229**, 416 (1971).

Synthesis and Biochemical Evaluation of Inhibitors of Estrogen Biosynthesis[1]

Robert W. Brueggemeier,[2a] E. Elizabeth Floyd, and R. E. Counsell*[2b]

The Interdepartmental Program in Medicinal Chemistry and the Department of Pharmacology, The University of Michigan, Ann Arbor, Michigan 48109. *Received November 18, 1977*

The synthesis and biochemical evaluation of various C_{19}-steroidal derivatives as inhibitors of estrogen biosynthesis are described. Steroids with substitutions on the A or B ring were synthesized by Michael addition of various thiol reagents to appropriate dienone intermediates. An in vitro assay employing the microsomal fraction isolated from human term placenta was used to evaluate aromatase inhibitory properties. Agents exhibiting high inhibitory activity were further evaluated in initial velocity studies (low product formation) to determine apparent K_i values. Several 7α-substituted androst-4-ene-3,17-diones were effective competitive inhibitors and have apparent K_i values equal to or less than the apparent K_m of 0.063 μM for the substrate androstenedione.

Inhibitors of estrogen biosynthesis have potential use as pharmacological tools and therapeutic agents. Such compounds can aid in the evaluation of the structural requirements of the enzymatic site, the purification of the aromatase enzyme, and the determination of estrogen function in biochemical processes. Therapeutically, aromatase inhibitors have potential use in the control of reproduction since a decrease in estrogen levels would result in insufficient uterine development. A more immediate use of inhibitors of estrogen biosynthesis would be in the treatment of disease states. A potent aromatase inhibitor would be a possible alternative to endocrine ablation in the treatment of advanced estrogen-dependent mammary carcinoma.

Four reports have appeared in the literature concerning inhibitors of estrogen biosynthesis. Schwarzel et al.,[3] Schubert et al.,[4] and Siiteri and Thompson[5] studied various available steroids for their ability to block the aromatization of androstenedione. The fourth report by Bellino et al.[6] examined bromoandrogens for their ability to inactivate the enzymatic site.

The objective of this research was to develop new agents as inhibitors of estrogen biosynthesis. At the outset, the research problem was considered to involve four steps: (1) to synthesize "lead" compounds as potential inhibitors; (2) to develop a screening assay for inhibitors and evaluate the feasibility of the study; (3) to synthesize additional inhibitors based on the screening results; and (4) to perform follow-up kinetic analysis on the more effective inhibitors.

Previous studies[3,7,8] indicated that C_{19} steroids resembling the substrate androstenedione most effectively interact with the active site of aromatase. On this basis, it was reasoned that effective inhibitors should retain the C_{19}-steroid nucleus as well as ketonic functions at the 3 and 17 positions. The introduction of various substituents

0022-2623/78/1821-1007$01.00/0