1 H, C*H*O), 2.77 (m, 1 H, C*H*—CHO), 2.33 (d of d of d, 1 H, C*H*—CH=CH—CHO, J_1 = 9.7, J_2 = 7.7, J_3 = 2.5 Hz), 2.03 (d, 1 H, O*H*, *J* = 8.3 Hz), 1.77 (d, 1 H, O*H*, *J* = 4.5 Hz); UV (MeOH) λ_{max} (ϵ) 272 (1190), 265 (920), 259 (570), 251 (412).

Anal. Calcd for $C_{20}H_{18}O_2$: C, 82.73; H, 6.25. Found: C, 81.90; H, 7.24.

1: mp (melts at 214 °C dec); IR (KBr) 3027, 2942, 1625, 1600, 1470, 1492 cm^{-1}; NMR ($CDCl_3$) 7.09 (m, 4 H, Ar*H*), 7.03 (m, 4 H, Ar*H*), 5.91 (m, 4 H, C*H*=C*H*), 4.25 (d, 2 H, ArC*H*Ar, *J* = 11.5 Hz), 3.52 (m, 2 H, CH=CH—CH—CH=CH); UV (MeOH) λ_{max} (ϵ) 284.5 (3850), 276 (2140), 268 (1080), 256 (1150), 219 (24 100).

Mass spectrum and elemental analysis were not attempted because of the low thermal stability of **1**.

2: mp (melts at 214 °C dec); IR (CCl_4) 3070, 3035, 2940, 2890, 1468, 1459 cm^{-1}; NMR ($CDCl_3$) 7.35–7.06 (m, 8 H, Ar*H*), 5.42–5.32 (m, 4 H, C*H*=C*H*—C*H*=C*H*), 4.21 (s, 2 H, ArC*H*Ar), 3.06 (s, 2 H, CH=CH—C*H*—C*H*—CH=CH); UV (MeOH) λ_{max} (ϵ) 291 (1150), 279 (2310), 272 (3610), 266 (3230), 260 (2570).

Mass spectrum and elemental analysis were not attempted because of the low thermal stability of **2**.

10: yellow oil; IR (KBr) 3020, 2920, 1680, 1470, 1460, 1390, 765, 740 cm^{-1}; NMR ($CDCl_3$) 7.30–3.00 (m, 8 H, Ar*H*), 6.62 (d of d, 1 H, C*H*=CH—CO, J_1 = 10.5, J_2 = 3.6 Hz), 5.71 (d of d, 1 H, CH=C-*H*—CO, J_1 ; 9.6 Hz, J_2 = 2.1 Hz), 4.26 (d, 1 H, ArC*H*Ar, *J* = 2.0 Hz), 4.06 (br s, 1 H, ArC*H*Ar), 2.88 (m, 1 H, C*H*—CH=CH—CO), 2.61 (m, 2 H, CH_2—CO), 2.17 (d of d, C*H*—CH_2—CO, J_1 = 14.0, J_2 = 3.6 Hz); UV (MeOH) λ_{max} (ϵ) 272 (1310), 265 (1442), 219 (9960).

Thermolyses of 1 and 2. Thermal decomposition of **1** (1.88 × 10^{-5} to 4.69 × 10^{-5} M) and **2** (1.34 × 10^{-5} M) in *n*-octane was followed by appearance of anthracene absorption at 375 nm. Measurements were made in 1-cm path length quartz cuvettes by using a Cary 219 UV/VIS spectrophotometer equipped with a thermostatic cell compartment. Temperature was measured simultaneously with a copper–constantan thermocouple placed in an identical cell. Kinetic plots were analyzed by the Guggenheim method[29] with time intervals optimized for linearlity, giving first-order rate constants. Arrhenius plots were linear, giving experimental activation energies which were converted to activation parameters with conventional transition-state theory.

Quantum Yield for Total Anthracene Formation. A ferrioxalate actinometer was made according to the procedure by Murov;[11] the procedure was tested until consistent results could be obtained. Stock solutions of **1** and **2** were adjusted to optical densities of 0.1 at 284 and 291 nm, respectively. Aliquots (0.5 mL) were deaerated and irradiated in 1-mL volume, 1-cm path length, narrow cells to minimize uneven stirring effects. The length of irradiation was 0.5–5.0 min. All sample trials were done by using the Perkin-Elmer MPF-4 spectrometer as the light source; the excitation slit was set at 10.0-nm band-pass. Samples were run at 0 ± 0.5 °C. Exposure of the actinometer solutions was done under identical conditions at 23.0 ± 0.5 °C for 15–40 min.

Quantum Yield for Excited Anthracene Formation. Stock solutions of **1** and **2** in spectrophotometric-grade methanol were adjusted to optical densities of 0.1 at 284 and 291 nm, respectively. Solutions were degassed by multiple freeze–thaw cycles at 0.01 torr. Emission and excitation spectra were taken on a Perkin-Elmer MPF-4 spectrofluorimeter with corrected spectra unit at 0 ± 0.5 °C from 220 to 600 nm. The excitation slit was set at 1.0-nm band-pass; the emission slit was 1.0 nm for **1** and 2.0 nm for **2**. Each sample was used only once to prevent interference from anthracene produced during the fluorescence measurement. Standards[30] of 9,10-diphenylanthracene in ethanol and quinine sulfate in 1 N H_2SO_4 were used to calibrate the emission quantum yield.

Acknowledgment. We thank the National Science Foundation for the supprt of this work, Dr. Tim Kwok Mak for developing the synthesis of diene **4** used in this investigation, Professor David Evans for a valuable suggestion in the synthesis of these compounds, and the National Cancer Institute for a grant to the University of Chicago Cancer Research Center to purchase the NMR spectrometer used in this work. One of us (P. C.) thanks the Hedwig Loeb Fellowship Fund of the University of Chicago for an undergraduate fellowship for carrying out a part of this study.

Registry No. 1, 80409-72-9; **2**, 80409-73-0; **3**, 42846-36-6; **4**, 80446-29-3; **5a**, 92241-42-4; **5b**, 92241-43-5; **5c**, 92241-44-6; **6a**, 80409-74-1; **6b**, 92241-45-7; **6c**, 80409-75-2; **7a**, 92343-32-3; **7b**, 80409-76-3; **7c**, 80409-77-4; **8a** (isomer 1), 80409-79-6; **8a** (isomer 2), 80446-28-2; **8b**, 80409-78-5; **9**, 92241-46-8; **10**, 92241-47-9; $CH_3C(OMe)_2CH_3$, 77-76-9; PhCH(OMe)$_2$, 1125-88-8; 1,4-cyclohexadiene, 628-41-1; anthracene, 120-12-7.

[29] Guggenheim, E. A. *Philos. Mag.* **1926**, *7*, 538–543.

[30] (a) Ware, W. R.; Baldwin, B. A. *J. Chem. Phys.* **1964**, *40*, 1703–1705. (b) Birks, J. B.; Dyson, D. J. *Proc. R. Soc. London, Ser. A* **1963**, *275*, 135–148.

Artificial Intelligence Approach to Structure–Activity Studies. Computer Automated Structure Evaluation of Biological Activity of Organic Molecules

Gilles Klopman

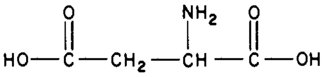
Contribution from the Chemistry Department, Case Western Reserve University, Cleveland, Ohio 44106. *Received March 9, 1984*

Abstract: A new program is introduced to study the relationship between structure and biological activity of organic molecules. The computer-automated structure evaluation program automatically recognizes molecular structures from the KLN code, a molecular linear coding routine, and proceeds automatically to identify, tabulate, and statistically analyze biophores, i.e., substructures believed to be responsible for known or anticipated biological activity of groups of molecules. The method is applied to the study of the carcinogenicity of polycyclic aromatic hydrocarbons, the carcinogenicity of *N*-nitrosamines in rats, and the pesticidal activity of some ketoxime carbamates.

The evaluation and prediction of the biological effect of chemicals has been at the center of preoccupation of chemists involved in drug development as well as of those concerned about the effect that chemicals may have on the environment.

In the last decade, several methods have emerged that have the potential of helping to solve the problem. They all, in one form or another, revolve around the basic concept that some correlation exists between structure and activity. Among those methods that

Table I. Subunits of Size 3 to 5, Generated by Aspartic Acid[a]



size 3	size 4	size 5
C″—CH$_2$—CH—	C″—CH$_2$—CH—C=	OH—C″—CH$_2$—AH—C=
C″—AH—CH$_2$—	C″—AH—CH$_2$—C=	O=C—CH$_2$—AH—C=
NH$_2$—CH—CH$_2$—	NH$_2$—CH—CH$_2$—C=	OH—C″—AH—CH$_2$—C=
NH$_2$—CH—C=	OH—C″—AH—CH$_2$—	O=C—AH—CH$_2$—C=
O=C—OH	C=C—AH—CH$_2$—	OH—K″—CH—CH$_2$—C=
OH—K″—CH—	OH—K″—CH—CH$_2$—	OH—K″—CH$_2$CH—C=
OH—K″—CH$_2$—	OH—K″—CH$_2$—CH—	O=K—CH—CH$_2$C=
O=K—CH—	O=K—CH—CH$_2$—	O=K—CH$_2$—CH—C=
O=K—CH$_2$—	O=K—CH$_2$—CH—	OH—K″—CH$_2$—CH—NH$_2$
	OH—K″—CH—NH$_2$	O=K—CH$_2$—CH—NH$_2$
	O=K—CH—NH$_2$	

[a] A indicates a carbon atom to which an amine group is attached, K indicates a carbon to which a side oxygen atom is attached, double prime indicates that the corresponding atom is unsaturated.

have shown greatest promise are the quantitative structure activity relationships (QSAR),[1] the pattern recognition methods (PR),[2,3] and discriminant analysis (DA).[4-6]

The QSAR methods are usually implemented as multivariate regression analysis programs where a linear correlation is sought between the observed biological activity of a series of congeneric substances and some more or less arbitrary descriptors. These include partition coefficients,[7] geometrical characteristics, often obtained from molecular mechanics programs,[8] and some reactivity characteristics, sometimes obtained from substituent tables,[1,9] but more often determined by quantum mechanical calculations.[10,11] These techniques have been widely applied to a multitude of biological test cases and often lead to correlations that are good enough to be used to develop new, and hopefully optimized, drugs.

One of the major difficulties in QSAR, as well as in PR work, is the selection of relevant properties to be used as descriptors. Indeed, to a large degree, the success of an application depends on the outcome of an *intelligent* and *discriminative* analysis of the factors that may be important in the mechanism of action of the molecules under study. Human intelligence is required to define what the relation might be between the structure and activity, and it is only after a qualitative understanding has been reached that meaningful descriptors can be defined and calculated.

The problem is particularly difficult to solve in the more heuristic approaches because of the need to find relevant descriptors that are general enough to exist in several molecules, and accessible enough to permit computations.[3,12]

In order to circumvent these difficulties and avoid the painstaking search of appropriate descriptors, we considered the possibility of writing a program that would be capable of evaluating automatically potential descriptors and, through discriminative analysis, select those that seem to be responsible for the observed property. Such a program would have *intelligence* in that it would have the ability to learn from the data and, hopefully, use this knowledge to establish causal relationships. In practice, it would be able to find appropriate descriptors and eliminate the need for the investigators to find a handle on the problem, prior to investigating it.

Before this can be achieved though, it is essential to define a suitable set of descriptors through which the computer can interact with the problem. This set should be general enough to be useful for a large number of problems. In the current version of our program, it consists of all possible substructural fragments that may be observed in organic molecules.

The use of substructural units as descriptors makes considerable sense to chemists who are used to relating chemical properties to functionalities consisting of one to four atoms, e.g., COOH, NH_2, etc., and it is not surprising to find that much interest has already centered around the best ways of selecting appropriate substructural units.[3,5,6,13] In most published methods, a limited number of such topological descriptors, e.g., keys, are preselected[5,14] while in a few others an open-ended approach is followed.[3,13] Of particular relevance to us is the work of Hodes[12] in which each type of structural feature, consisting of augmented atoms AA, is assigned an activity weight, based on the statistical significance of its frequency of occurrence in the "active" training set. The likelihood that a new compound has activity is then computed on the basis of the weights of all its fragments.

Unfortunately, the chemical analogy was often lost in previous methods because the activity, or lack of it, was related in a complex manner to the presence of a large number of keys. *Causal relationships* were not established, probably because the selected descriptors were rarely adequate, being either too small or too restricted to be associated with the possible complex entities that give rise to *biological functionality*.

In this paper, we present the results of our initial efforts, dealing mostly with the identification of relevant descriptors. In a subsequent communication, we will show the quantitative implementation of this concept and its applications in the field of drug development.

Method

The computer automated structure evaluation technique consists in tabulating, for each molecule of a training set, the type of fragments that can be formed by breaking up the molecule into linear subunits containing between 3 and 12 interconnected heavy atoms (i.e., linear chains of interconnected atoms other than H), together with the hydrogen atoms attached to them,[15] and two sets of labels. One of the sets of labels

[1] Osman, R.; Weinstein, H.; Green, J. P. *ACS Symp. Ser.* **1979**, *112*, 21.

[2] (a) Kirschner, G. L.; Kowalski, B. R. *Drug Des.* **1978**, *8*. (b) Yuta, K.; Jurs, P. C. *J. Med. Chem.* **1981**, *24*, 241. (c) Henri, D. R.; Jurs, P. C.; Denny, W. A. *J. Med. Chem.* **1982**, *25*, 899.

[3] Chu, K. C.; Feldmann, R. J.; Shapiro, M. B.; Hazard, G. F., Jr.; Geran, R. I. *J. Med. Chem.* **1975**, *18*, 539.

[4] Dove, S.; Fronke, K.; Monshjan, O. L.; Schkuljev, W. A.; Chashajan, L. W. *J. Med. Chem.* **1979**, *22*, 90.

[5] Enslein, K.; Craig, P. N. *J. Toxicol. Environ. Health* **1982**, *10*, 521–530.

[6] (a) Hodes, L. *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 128–132. (b) Hansch, C. In "Drug Design"; Ariens, E. J., Ed.; Academic Press: New York, 1971, Vol. 1. (c) Hansch, C.; Dunn, W. J., III *J. Pharm. Sci.* **1972**, *61*, 1.

[7] (a) Hansch, C. In "Drug Design"; E. J. Ariens., Ed., Academic Press: New York, 1971, Vol. 1. (b) Hansch, C.; Dunn, W. J., III *J. Pharm. Sci.* **1971**, *61*, 1.

[8] (a) Duchamp, D. J. *ACS Symp. Ser.* **1979**, *112*, 79. (b) Weintraub, H. J. R. *Ibid.* **1979**, *112*, 353.

[9] Martin, Y. C. In "Quantitative Drug Design"; Grunewald, G. L.; Ed.; Marcel Dekker, Inc.: New York, 1978; Medicinal Research Series, Vol. 8.

[10] (a) Weinstein, H. *Int. J. Quantum Chem.* **1975**, *QBS2*, 59. (b) Grunewald, G. L.; Creese, M. W.; Walters, D. E. *ACS Symp. Ser.* **1979**, *112*, 439.

[11] Petit, B.; Potenzone, R.; Hopfinger, A. J.; Klopman, G.; Shapiro, M. *ASC Symp. Ser.* **1979**, *112*, 552.

[12] Hodes, L.; Hazard, G. F.; Geran, R. I.; Richman, S. *J. Med. Chem.* **1977**, *20*, 496.

[13] Hodes, L. *ACS Symp. Ser.* **1979**, *112*, 583, 14.

[14] Jurs, P. C.; Chou, J. T.; Yuan, M. *ACS Symp. Ser.* **1979**, *112*, 103.

[15] See also the BASIC (Basel Information Center for Chemistry) method for keys including linear sequences 4 to 6.

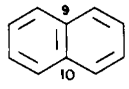
Table II. Sample Fragments for Analysis of Acidity of Organic Acids[a]

fragment	total no.	no. present in inactives	no. present in actives
CH_3—CH_2—CH_2—	25	16	9
CH_2—CH_2—CH_2—	63	35	28
CH=CH—CH_2—	11	7	4
C″—CH_2—CH—	14	7	7
O=C—CH_2—	55	25	30 +
O=C—OH	42	4	38 +++
CH=C—OH	8	2	6 +
CH_2—CH—Cl	10	6	4
NH_2—CH—CH_2—	7	6	1 -
NH_2—CH—C=	3	3	0 -

[a] Total number of molecules = 100. Number of actives = 40.

indicates the multiplicity of the bonds that join the atoms of the sequence, and the other indicates the presence of a side chain consisting of a terminal functionality such as a halogen, NH_2, COOH, etc. These fragments will become the descriptors of the correlation, and their selection is entirely driven by the nature of the compounds of the data base. The procedure is open ended and completely automatic. For example, aspartic acid would generate the fragments indicated in Table I. (No attempt is made at this time to remove redundancies; this is done at a later stage of the calculations.)

In the case of fused rings, though, special care was taken to include only the fragments formed by the envelope. Thus, in naphthalene, no subunit contains the sequence of atoms C_9–C_{10} since the bond they form is common to two rings. This kind of bond is chemically inert except in very rare cases. (Larger bridges, as in norbornane for example, are unaffected by this procedure and do appear in the corresponding fragments.)



All subunits belonging to an active molecule are labeled active while those belonging to an inactive molecule are labeled inactive. Once all molecules that constitute the training set have been entered, a statistical analysis of the fragment distribution is made. A binomial distribution[16] is assumed, and each type of fragment is considered irrelevant if its distribution among actives and inactives is the same as that of the total sample of molecules. Any significant discrepancy from a random distribution of subunits between the active and inactive pool is then taken as an indication that the subunit is relevant to the property being examined. Thus a fragment is considered relevant if it is found in a binomial distribution that would have had at most a 5% chance of being observed if its occurrence was random. It is labeled as activating if its distribution is skewed toward active molecules and inactivating otherwise.

Let us consider as an example the hypothetical problem of determining why some organic molecules change the color of litmus paper to red. Let us suppose that we do not know which structural feature is responsible for the observed acidity and that 100 diverse compounds are used in the training set. Of these, 40 compounds are acids and 60 are not. When entered in the computer, each molecule is broken into its fragments. The fragments are catalogued, redundancies are eliminated, and, when all molecules have been entered, the statistical analysis of their distribution is performed. A sample of such a distribution is shown in Table II.

From this, it can be seen that the fragment CH_2-CH_2-CH_2, for example, was found in 63 molecules, of which 28 were acids and 35 were not. Such a distribution is about what is to be expected if it occurred randomly, and the fragment is thus not perceived as being relevant to acidity. A different situation exists for the fragment O=C—OH which is found in 38 acids and 4 nonacids. This indicates a strong relationship, and O=C—OH is therefore selected as possibly responsible for the observed property. Discrepancies, such as the fact that 4 of the fragments were found in nonacids, are due either to experimental error or to the fact that a *deactivating* fragment such as NH_2 was also present in the molecule.

The computer automated structure evaluation technique clearly borrows some of its mechanics from pattern recognition techniques[3] and discriminant analysis[12] and shares some of their shortcomings as well. For example, the program cannot currently assess the potency of drugs and does not provide a quantitative correlation between activity and descriptors (a quantitative implementation of the program is, however, being prepared and will be presented shortly). Yet it does provide intuitive background and general guidelines for further, more quantitative analysis. This has been achieved by allowing the size of the keys to be large enough to encompass "biological functionalities". As will be shown below, it is not unusual to find that relevant keys extend over 8 to 10 atoms, as compared to 2 or 3 commonly found to represent "chemical functionalities".

Table III. Character Codes[a]

atom		+ 1-H	+ 2-H	+ 3-H	+ 1=O	+ 2=O
H	-H					
F	-F					
G	-CL					
B	-BR					
I	-I					
C	-C	D -CH	R -CH2	M -CH3	T -CO	
N	-N(3)	E -NH	A -NH2		U -NO	X -NO2
Q	-N+					
O	-O	K -OH				
P	-P(3)				W -PO	
Z	-P(5)				L -PO	
S	-S	J -SH			V -SO	Y -SO2

[a] * aromatic ring. (aliphatic ring. = chain of CH_2's. (and = are followed by a number indicating the number of carbon atoms.) close last ring. / close first ring.

Another shortcoming, which the method currently shares with some of the methods that preceded it, is related to the fact that, so far, the descriptors are only topological in nature and, for example, do not include such traditional descriptors as partition coefficients. As a result, the range of applications of the method must necessarily be restricted to systems that are not overly sensitive to those potentially important descriptors.

The computer package that implements this technique consists of two complementary computer programs: the CASE[17] program proper and the SPLOT program. The input to either program consists of the coded name of a molecule, or a file containing the coded name of several molecules, and, for each molecule, an integer, 0 or 1, indicating its activity in the biological assay under consideration. The coded name is obtained from the KLN code (Table III) described in a previous paper.[18] For example, aspartic acid would be coded as **KTRDATK** and styrene, C_6H_5CH=CH_2 as **D2DD2DD2C/D2R** or ***D2R**.

The SPLOT program is merely used to check the adequacy of the KLN code: it reads and decodes the KLN code and generates a reasonable tri-dimensional geometry for each molecule. The resulting geometry is then plotted on the terminal or printed.

The program CASE performs automatically all operations related to the structure–activity analysis; it reads and decodes the KLN code and then proceeds to generate the connectivity matrix and determines the type and number of occurrences of each possible subunit that exists in each molecule. After redundancies are eliminated, the program performs the statistical analysis and displays the "relevant" fragments as **Biophores** if they are found to activate a molecule or **Biophobes** if, on the contrary, they prevent activity. Additional data can be introduced at any time and result in an immediate updating of the internally generated fragment descriptors.

The program can also be used in a predictive mode by entering a question mark in reply to the query relevant to the activity of the molecule. In this case, the program compares the fragments generated by the new molecule to those that are held in memory and projects the probability that the new molecule is active or not.

The programs are currently implemented on a DIGITAL EQUIPMENT VAX 11/750 computer outfitted with a 470 Mbyte Winchester Disk and a VT100 terminal upgraded with a TEKTRONIX 4010 compatible retrographics board.

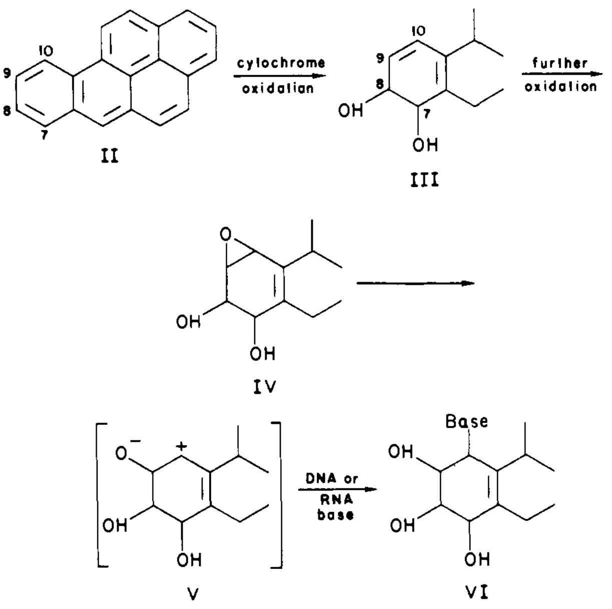
We illustrate, below, the use of the program in three types of structure–activity applications: the carcinogenicity of aromatic hydrocarbons, the carcinogenicity of *N*-nitrosamines in rats, and the pesticidal activity of a series of ketoxime carbamates. The first two systems have been extensively studied in the literature and their mechanism of action is fairly well documented. We merely applied our technique to these cases to test its ability to reproduce known trends and identify the known relevant descriptors.

[16] Bevington, P. R. *Data Reduct. Error Anal. Phys. Sci.* **1969**, *27*.

[17] Not to be confused with the CASE program (Computer Assisted Structure Elucidation) developed by Munk, M., Arizona State University.

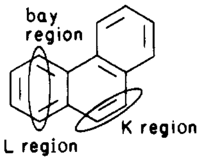
[18] Klopman, G.; McGonigal, M. *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 48.

Scheme I



Carcinogenicity of Polycyclic Aromatic Hydrocarbons (PAH)

The potential carcinogenic activity of polycyclic aromatic hydrocarbons has been known ever since it was suggested, at the turn of the century, that they might be implicated in the high incidence of cancer among chimneysweeps. The chemical properties of PAH's are mainly due to the distribution of their π electrons, and this renders them easily amenable to simple quantum mechanical calculations such as the Huckel method. In one of the earliest attempts to relate structures and carcinogenicity, B. and A. Pullman proposed, in 1955,[19] that the carcinogenicity of the PAH resulted from the existence of an active K region and an inactive L region.



I

The proposal met with considerable controversy but remained valid until the midseventies when several authors[20] showed that, in many cases, the metabolism of PAH proceeded via a totally different route involving a terminal benzene ring.

For example, in benzo[*a*]pyrene (II), after initial oxidation at the 7–8 position (III), the molecule is transformed to the diol epoxide (IV). It is this epoxide that, upon ring opening, is believed to react with one of the components of DNA or RNA to give rise to an event conducive to cancer (Scheme I).

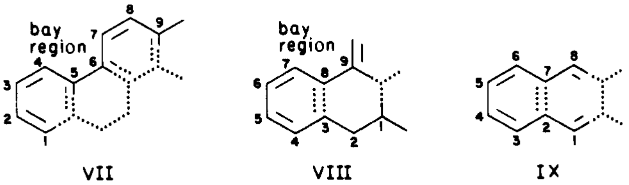
Once the active site had been identified, it became clear that the ability to form the cation V is important in determining the propensity of the parent hydrocarbon to be carcinogenic. This was shown to be the case by Jerina,[21] who found that the stabilization of the "bay region" cation is directly proportional to the carcinogenicity of the hydrocarbon.

The way we approached the problem was to code all but five randomly chosen unsubstituted PAH's listed by A. Dipple in "Chemical Carcinogens"[22] and label them as inactive if they are so described in the reference or active if they are either moderately or very active. Thus 23 of these 38 PAH's were listed as inactive, and 15 had various degrees of activity. Only fragments found in binomial distributions that would be observed with less than 2.5% probability if their occurrence was random, were considered relevant. Of the 2876 fragments identified by the program, only two active (VII and VIII) fragments and one inactive (IX) fragment qualified under these stringent constraints.

Table IV. Carcinogenicity of Polycyclic Aromatic Hydrocarbons

	presence of fragments			carcinogenicity		
molecule	VII	VIII	IX	calcd[a]	calcd[b]	obsd
benzene				–	*	–
naphthalene				–	24	–
anthracene				–	10	–
pyrene				–	0	–
naphtacene				–	21	–
triphenylene				–	*	–
chrysene				–	100	+
benz[*a*]anthracene			×	–	21	+
benz[*c*]phenanthrene				–	100	++
benz[*a*]pyrene	×	×		+	100	+++
benz[*e*]pyrene				–	0	–
pentacene			×	–	21	–
benz[*a*]naphtacene			×	–	21	–
benz[*b*]chrysene	×		×	–	10	–
benz[*c*]chrysene	×			+	100	++
benz[*g*]chrysene	×			+	100	++
picene	×			+	98	–
dibenz[*ac*]anthracene			×	–	24	+
dibenz[*ah*]anthracene				–	*	++
dibenz[*cg*]phenanthrene				–	*	–
dibenz[*bg*]phenanthrene			×	–	10	–
pentaphene			×	–	10	–
anthanthrene				–	*	–
benz[*ghi*]perylene				–	*	++
dibenz[*ae*]pyrene	×	×		+	100	+++
dibenz[*al*]pyrene		×		+	100	+++
dibenz[*ah*]pyrene	×	×		+	100	+++
dibenz[*el*]pyrene				–	*	–
naphtho[2,3-*a*]pyrene			×	–	100	++
naphtho[2,3-*e*]pyrene			×	–	0	–
dibenz[*bk*]chrysene			×	–	10	–
dibenz[*aj*]naphthacene				–	*	–
dibenz[*ac*]naphthacene			×	–	21	++
anth[1,2-*a*]anthracene			×	–	10	–
benz[*c*]pentaphene			×	–	10	–
naphtho[1,2-*a*]triphenylene				–	*	–
hexacene			×	–	21	–
tribenz[*aei*]pyrene	×	×		+	100	++

[a] Calculated with the 3 fragments that have 99+% chance of being related to activity. [b] Probability that the molecule is active, calculated with fragments that have an 85+% chance of being related to activity. An asterisk indicates that neither active nor inactive fragments could be found. The molecule is presumed to be inactive.



VII VIII IX

activating fragments deactivating fragment

Thus, our correlation is really based on 3 parameters, i.e., the presence, or not, of substructures VII, VIII, and IX, containing 9, 9, and 8 heavy atoms, respectively. No smaller fragment appeared "relevant", supporting our contention that, to be of biological significance, the subunits used as descriptors may have to be larger than previously thought.

Biophore VII existed in 6 active molecules and 2 inactive molecules; VIII was present in 5 molecules, all actives, and biophobe IX was found in 10 inactive molecules and 4 active ones (Table IV).

[19] Pullman, A.; Pullman, B. *Adv. Cancer Res.* **1955**, *3*, 117.

[20] (a) Borgen, A.; Darvey, H.; Castagnoli, N.; Crocker, T. T.; Rassummsen, R. E.; Wang, I. Y. *J. Med. Chem.* **1973**, *16*, 502. (b) Depierre, J. W.; Ernster, L. *Biochem. Biophys. Acta* **1977**, *473*, 149.

[21] Jerina, D. M.; Lehr, R. E. In Microsomes and Drug Oxidation"; Ulrich, V., Roots, I., Hildebrant, A. G., Eatabrook, R. W., Conney, A. H., Eds.; Pergamon Press, Inc.: Oxford, 1977; p 709.

[22] Dipple, A. *ACS Monograph* **1976**, *173*, 245.

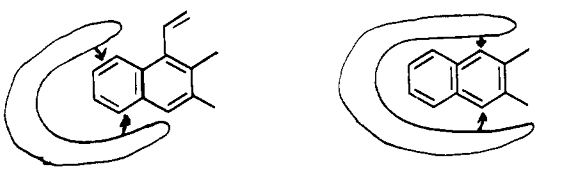
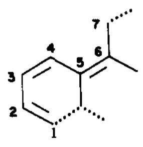


Figure 1. Postulated mechanism of oxidative metabolism of PAH's.

Structure VIII seems particularly important as it appeared in every PAH that showed high activity. It is interesting to note that both biophores VII and VIII define a "bay region" while the inactivating biophobe IX indicates, at best, the existence of an active "L" region. Thus, our results would have indicated the importance of the bay regions even without the current knowledge that they are, indeed, implicated in the carcinogenic process.

It is also interesting to note that a minimal bay region fragment, such as X, does not appear to be important, and that a CH group at position 8 of structure VII and 2 of structure VIII seems to be necessary for activity. Similarly, the absence of a hydrogen on carbons 9 of VII and 1 of VIII seems to be important as well. For example, while fragment VIII appears in 5 molecules, all active, the corresponding fragment where group 1 (a substituted C in VIII) remains undefined appears in 22 molecules, of which 9 are inactives and 11 are actives. Clearly, group 1 needs to be a substituted carbon atom. Similarly, if the H on the CH of group 2 of VIII is replaced by another carbon atom, the corresponding fragment appears in 10 molecules of which half are active and half inactive. Thus it is found that the constraints represented by the structure of the relevant fragments are very strict. However, whether these structural requirements are relevant to the mechanism of the carcinogenic event or whether the observation is linked to the existence of geometrical constraints necessary to enhance the activity of the corresponding carbocation remains to be elucidated. What the data seem to show, though, is that the metabolism of these species takes place, at least initially, around a somewhat unsubstituted ring. Detoxification will then take place if either the L region 1-8 (IX) or the 1-2 (VIII) or 8-9 (VII) bonds are available for further oxidation. This is illustrated in Figure 1, where the action of the oxidizing enzyme is shown as a function of the geometrical constraints of the relevant substructural fragments.



X

Using the existence of an active fragment as an indication of activity and the absence of an active fragment or presence of an inactivating fragment as an indication of inactivity, we were able to account for the observed activity of 29 out of the 38 compounds (Table IV). When the system was expanded to include all fragments that show at least an 85% chance of being related to carcinogenicity, we were able to reproduce the reported activity of 32 out of the 38 compounds.

Finally, we used the program to predict the activity of the five compounds that were left out of the training set. The results are shown in Table V. As can be seen, the program correctly identified three out of the five compounds and was unable to come up with a prediction in the other two cases.

Carcinogenicity of *N*-Nitrosamines in Rats

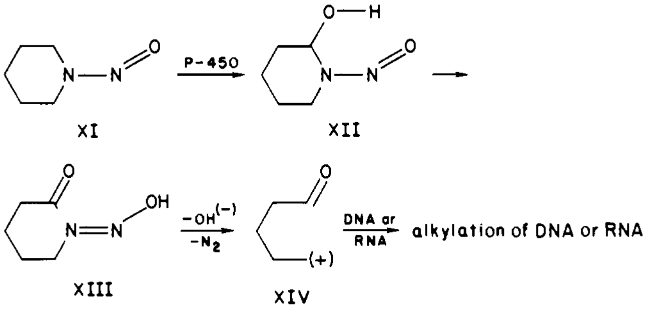
Some *N*-nitrosamines produce oesophagus and other cancers in rats. It is generally believed that the *N*-nitrosamines are not direct acting carcinogens but need metabolitic activation to express their activity.[23] The activation is possibly provided by cytochrome P-450 leading to the possible mechanism of Scheme II.

Table V. Predicted Carcinogenicity of Polycyclic Aromatic Hydrocarbons

molecule	carcinogenicity calcd[a]	carcinogenicity obsd
phenanthrene	0	inactive
perylene	*	inactive
dibenz[*aj*]anthracene	*	active
dibenz[*ai*]pyrene	100	active
benz[*b*]pentaphene	10	inactive

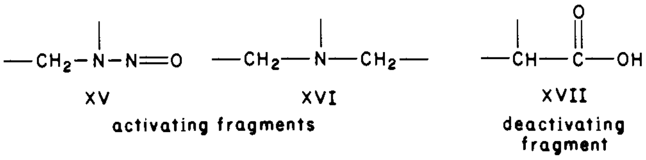
[a] Probability that the molecule is active, calculated with fragments that have an 85+% chance of being related to activity. An asterisk indicates that no predictions could be made because neither an active nor an inactive fragment has been found. The molecule is presumed to be inactive.

Scheme II



The training set here consists of 39 cyclic *N*-nitrosamines tested by Lijinsky[24] on rats. It includes 27 active carcinogens (i.e., a daily dose of 5% of the total food intake led to 50% of the animals dying from cancer in less than 100 weeks) and 12 inactive compounds (i.e., more than 50% of the animals still alive after 100 weeks). The data base is amenable to QSAR studies, and we have previously obtained good quantitative correlations when the data were analyzed by a nonlinear QSAR method.[11] We were curious, though, to see if the CASE technique would reproduce the features obtained previously.

Only three relevant, and quite trivial, fragments were found by the program to have a better than 98% chance of being related to activity: two activating biophores XV and XVI, and one inactivating biophobe XVII.



Note that the "relevant" fragments are considerably smaller in size here. Thus it is probable that, in this case, positive results could also have been obtained with methods based on smaller subunits such as "augmented atoms".[12] Actually, careful visual inspection of the data might even be sufficient in this case.

The presence of an activating fragment coupled with the absence of inactivating ones led to the correct identification of all 27 active and marginally active compounds (Table VI). Ten compounds which lack an activating fragment or include an inactivating one are correctly identified as inactive molecules.

The most interesting aspect of this investigation is the finding that the active subunit XV needs two hydrogen atoms on the carbon atom bonded to the NNO group. This, indeed, would be required if Scheme II is the correct metabolic path, since the condition for generating the ultimate carbocation is the necessity

[23] (a) Hecht, S. S.; McCoy, G. D.; Chen, C. B.; Hoffmann, D. *ACS Symp. Ser.* **1981**, *174*, 49. (b) Wishnok, J. S. *ACS Symp. Ser.* **1981**, *174*. 77.

[24] (a) Lijinski, W., the data base was provided by Dr. W. Lijinski, Chemical Carcinogesis Program, Frederick Cancer Research Center, Frederick, MD 21701, 1978. (b) Lijinski, W.; Taylor, H. W. *Cancer Res.* **1976**, *36*, 1988.

Table VI. Carcinogenicity of Cyclic *N*-Nitrosamines

		carcinogenicity	
	molecule	calcd[a]	obsd
1	nitrosopiperidine	100	+++
2	2-methylnitrosopiperidine	100	++
3	4-methylnitrosopiperidine	90	+++
4	2,6-dimethylnitrosopiperidine	*	–
5	3,5-dimethylnitrosopiperidine	100	+
6	2,2,6,6-tetramethylnitrosopiperidine	*	–
7	4-phenylnitrosopiperidine	90	++
8	4-*tert*-butylnitrosopiperidine	90	–
9	3-hydroxynitrosopiperidine	95	+++
10	4-hydroxynitrosopiperidine	90	+++
11	4-ketonitrosopiperidine	87	+++
12	2-carboxynitrosopiperidine	0	–
13	4-carboxynitrosopiperidine	0	–
14	4-chloronitrosopiperidine	100	++++
15	3,4-dichloronitrosopiperidine	100	++++
16	3,4-dibromonitrosopiperidine	100	++++
17	nitroso-1,2,3,6-tetrahydropyridine	83	++++
18	nitrosomethylphenidate	0	–
19	nitrosopyrrolidine	87	++
20	2,5-dimethylnitrosopyrrolidine	*	–
21	3,4-dichloronitrosopyrrolidine	100	++++
22	2-carboxynitrosopyrrolidine	0	–
23	2-carboxy-4-hydroxynitrosopyrrolidine	0	–
24	nitroso-3-pyrroline	86	++
25	nitrosomorpholine	87	+++
26	2,6-dimethylmorpholine	100	++++
27	nitrosothiomorpholine	87	++
28	nitrosophenmetrazine	*	–
29	dinitrosopiperazine	87	+++
30	2,5-dimethyldinitrosopiperazine	100	+++
31	2,6-dimethyldinitrosopiperazine	100	++++
32	2,3,5,6-tetramethyldinitrosopiperazine	*	–
33	dinitrosohomopiperazine	100	++++
34	nitrosopiperazine	83	–
35	nitrosoazetidine	83	+++
36	nitrosohexamethyleneimine	100	++++
37	nitrosoheptamethyleneimine	100	++++
38	nitrosooctamethyleneimine	100	+++
39	nitrosododecamethyleneimine	100	++

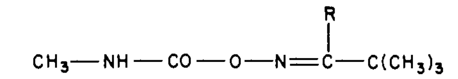
[a] Probability that the molecule is active, calculated with fragments that have an 85+% chance of being related to activity. An asterisk indicates that no predictions could be made because neither an active nor an inactive fragment has been found. The molecule is presumed to be inactive.

Table VII. Prediction of Carcinogenicity of Cyclic *N*-Nitrosamines

molecule	calcd[a]	obsd
3-methylnitrosopiperidine	100	active
nitrosoguvacoline	0	inactive
2-methyldinitrosopiperazine	100	active
4-methylnitrosopiperazine	87	active
N-nitroso-L-thioproline	0	?
N-nitroso-L-methylthioproline	0	?

[a] Probability that the molecule is active, calculated with fragments that have an 85+% chance of being related to activity.

Table VIII. Prediction of Insecticidal Properties of Certain Ketoxime Carbamates



R	calcd[a]	obsd
$-CH_2-S-CH_3$	100	active
$-CH_2-S-CH_2-CH_3$	100	active
$-CH_2-O-CH_3$	0	inactive
$-CH_3$	*	active

[a] Probability that the molecule is active, calculated with fragments that have an 85+% chance of being related to activity. An asterisk indicates that no predictions could be made because neither an active nor an inactive fragment has been found. The molecule is presumed to be inactive.

of oxidizing the carbon vicinal to the NNO group, and this can only happen if a carbon hydrogen bond can be transformed into a carbon–hydroxyl bond.

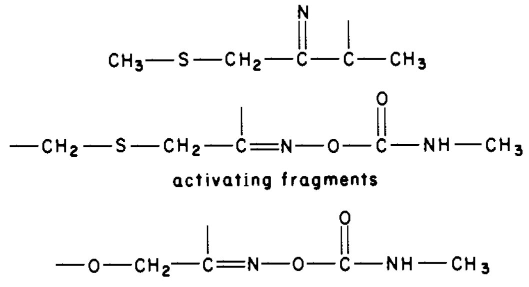
There does not seem to be any structural relationship between the positions of the deactivating group and the activating ones. This leads us to believe that the deactivating fragments do not act by modifying the reactivity of the activating fragment. Rather they provide an alternative metabolic pathway, either by reacting, in priority, with another enzyme, or by letting the molecules be transported out of range of the target. The latter could be due to a large change in partition coefficient generated by the presence of the carbonyl or hydroxyl groups. Whatever the reason, three "automatically determined" parameters allow us to classify correctly 37 out of the 39 compounds that constitute the data base.

Here again, we tested the predictive power of the program by applying it to four compounds that were withheld from the training set. The results are shown in Table VII. As can be seen, we obtained correct results for all four compounds. We also predict the two *N*-nitroso-L-prolines to be inactives. This result is of interest as it was reported that they were found to exist in human urine.[25]

Insecticidal Activity of Ketoxime Carbamates

Certain 2-butanone *O*-(methylaminocarbonyl)oximes were synthesized and screened for insecticidal and acaricidal activity by Magee and Limpel.[26] The lethal concentration required to kill 50% of the population (LC_{50}) was reported for several insect species. We selected the activity against the two spotted spider mites for our study because of the greater variability of the observed results. We randomly selected 39 of the compounds and called them inactive if their LC_{50} was larger than 500 ppm, marginal if their LC_{50} was larger than 300 ppm, and active otherwise. Thus 23 were labeled marginal or active and 16 were labeled inactive.

The CASE program identified 4155 fragments of which 1573 belonged to inactive molecules, 395 to marginal ones, and 2187 to active molecules. Of these, only 10 appeared relevant to the insecticidal property. The three statistically most significant ones are shown below:



With these, and the other relevant fragments, the program identified correctly all of the inactive molecules. Of the 4 marginal compounds, 2 were identified as active and 2 as inactive. Of the remaining 19 compounds, all but 4 were correctly identified as active. Encouraged by these results, we examined, in the predictive mode, the 4 remaining compounds. The results appear in Table VIII. The inactive and 2 of the 3 active compounds were correctly identified.

Conclusion

The computer automated structure evaluation technique provides a simple method for the analysis of structure–biological activity. While at this stage of development the program cannot address the question of potency, it has the advantage of not being restricted to a limited number of preselected descriptors. Indeed, the program automatically selects from the thousands of possible

[25] Tsuda, M.; Hirayama, T.; Sugimura, T. *Gann* **1983**, *74*, 331.

[26] Magee, T. A.; Limpel, L. E. *Agric. Food Chem.* **1977**, *25*, 1376.

structural descriptors those that best fit the data. The result is that only a few, well-suited descriptors are sufficient to catalogue and/or predict the activity of many molecules.

Acknowledgment. The author thanks Professor H. Rosenkranz for stimulating this work, for helping with numerous suggestions, and for providing much of the needed biological data. The author is also grateful to Professor Angela Sharkey for the help she provided him with the statistical analysis of the data and to the Environmental Protection Agency for their financial support for the study of structure–activity relationships in potential carcinogens (Grant No. R810063).

Orbital Interactions in Some Polycycloalkyl Halides: A Photoelectron Spectroscopic Study

Ramyani S. Abeywickrema,† Ernest W. Della,† Paul E. Pigou,† Michelle K. Livett,‡ and J. Barrie Peel*‡

Contribution from the School of Physical Sciences, The Flinders University of South Australia, Bedford Park, South Australia 5042, and the Department of Physical Chemistry and Research Centre for Electron Spectroscopy, La Trobe University, Bundoora, Victoria, Australia 3083.

Received July 10, 1984

Abstract: The He I photoelectron spectra of tricyclo[3.1.1.0^{3,6}]heptane and its 6-bromo and 6-iodo derivatives and of 1-bromo- and 1-iodo- cubane and adamantane have been measured. The low-ionization energy region of each alkyl halide spectrum shows considerable variations from the simple patterns observed for acyclic alkyl halides. The competition between spin–orbit coupling and conjugative effects which characterizes ionizations involving the halogen lone-pair orbitals is examined by reference to He II spectral measurements. Assignment of the spectra is facilitated by ab initio molecular orbital calculations based on a valence-electron model-potential method. In 1-bromotricyclo[3.1.1.0^{3,6}]heptane and bromocubane the lone-pair bromine orbitals, n_{Br}, appear in two ionization bands involving doubly degenerate e orbitals with respectively antisymmetric and symmetric admixtures of alkane orbitals. In 1-bromoadamantane and 1-iodoadamantane the σ_{CX} bonding character is shared between two orbitals of a_1 symmetry. In each of the iodo compounds the first ionization energy is associated predominantly with iodine character, whereas in the bromo compounds, the first photoelectron band is of varying bromine character. In 6-bromotricyclo[3.1.1.0.^{3,6}]heptane the first band is of alkane character.

The study of alkyl halides by He I photoelectron (PE) spectroscopy has been confined mainly to the smaller acyclic systems up to and including the isomers of the butyl halides.[1-7] Among the monocyclic alkyl halides, cyclopropyl and cyclobutyl bromides have been included in a study of the competition between spin–orbit coupling and conjugation effects.[1] A recent study on some bromo and iodo bicycloalkanes examined the same effects.[8] However, among the polycyclic alkyl halides, only 1-bromoadamantane has been investigated.[9]

The He I PE spectra of alkyl halides (excluding alkyl fluorides) are characterized by two main features. First, the low ionization energy (IE) bands, attributed to the photoejection of electrons from molecular orbitals (MOs) localized on the halogen atom, normally feature an intense doublet caused by spin–orbit (SO) coupling effects in the molecular ions. The magnitude of the splitting and the variation in sharpness of these bands provide a measure of the interaction between halogen and alkane orbitals. The admixture of alkane character in these orbitals can be shown by a He I/He II relative intensity analysis when He II PE data are available.

Second, the comparison of the PE spectrum of an alkyl halide with that of its parent alkane normally shows the simple stabilizing effect of halogen substitution in increasing the alkane IEs, with, for example, bromine producing larger IE shifts than iodine owing to its higher electronegativity.

This basic description, which is generally accurate for the acyclic bromo and iodo alkanes, is varied in the case of many cyclic alkyl halides. In cyclopropyl bromide, the lone-pair band at lower IE is considerably broadened through conjugative interaction with cyclopropyl orbitals, and its separation from the sharp second band is somewhat greater than that expected by SO coupling effects alone. In 1-bromoadamantane,[9] the characteristic SO doublet is lost, caused by a strong interaction of halogen and alkane orbitals which are proximate in energy. By comparison the bromobicycloalkanes[8] show the SO split peaks with the magnitude of the splitting decreasing in going from 1-bromobicyclo[2.1.1]hexane to 1-bromobicyclo[2.2.1]heptane to 1-bromobicyclo[2.2.2]octane. Together with increases in the breadth of these bands through the series, this indicates increases in conjugative interaction of alkane orbitals with the nonbonding bromine orbitals.

So in large polycyclic alkanes, particularly where ring strain leads to low IE bands in the PE spectra, such interactions are likely to be common. In certain polycycloalkyl halides nonbonding halogen orbitals may not only lose their localized identity, but may not even be associated with the first ionization band of the molecule.

The bromo and iodo polycycloalkanes included in this study are 6-bromotricyclo[3.1.1.0^{3,6}]heptane (**1b**), 6-iodotricyclo[3.1.1.0^{3,6}]heptane (**1c**), bromocubane (**2b**), iodocubane (**2c**),

† The Flinders University of South Australia.
‡ La Trobe University.

(1) F. Brogli and E. Heilbronner, *Helv. Chim. Acta*, **54**, 1423 (1971).
(2) D. W. Turner, C. Baker, A. D. Baker, and C. R. Brundle in "Molecular Photoelectron Spectroscopy", Wiley-Interscience, London, 1970.
(3) J. L. Ragle, I. A. Stenhouse, D. C. Frost, and C. A. McDowell, *J. Chem. Phys.* **53**, 178 (1970).
(4) J. A. Hashmall and E. Heilbronner, *Angew. Chem. Int. Ed. Engl.*, **9**, 305 (1970).
(5) A. D. Baker, D. Betteridge, N. R. Kemp, and R. E. Kirby, *Anal. Chem.*, **43**, 375 (1971).
(6) K. Kimura, S. Katsumata, Y. Achiba, H. Matsumoto, and S. Nagakura, *Bull. Chem. Soc. Jpn.* **46**, 373 (1973).
(7) R. G. Dromey and J. B. Peel, *J. Mol. Struct.* **23**, 53 (1974).
(8) E. W. Della, R. S. Abeywickrema, M. K. Livett, and J. B. Peel, *J. Chem. Soc., Perkin Trans.*, in press.
(9) S. D. Worley, G. D. Mateescu, C. W. McFarland, R. C. Fort, and C. F. Sheley, *J. Am. Chem. Soc.*, **95**, 7580 (1973).

0002-7863/84/1506-7321$01.50/0