Acta Chemica Scandinavica B 32 (1978) 602—608

Carcinogenicity of Polycyclic Aromatic Hydrocarbons Studied by SIMCA Pattern Recognition

BO NORDÉN, ULF EDLUND and SVANTE WOLD

Department of Organic Chemistry, Institute of Chemistry, University of Umeå, S-90187 Umeå, Sweden

Thirty-two polycyclic hydrocarbons were classified as carcinogenic or not on the basis of a multivariate statistical analysis of twenty-three variables characterizing each hydrocarbon. A prediction of the level of carcinogenicity for the active compounds was also obtained on the same basis.

The interest in the carcinogenicity of polycyclic aromatic hydrocarbons (PAH) goes back to the study of Sir Percival Pott [1], who showed in 1775 that chimney sweepers had an increased tendency to have scrotal cancer. When benz-[*a*,*h*]-anthracene was synthesized [2] and benz-[*a*]-pyrene was found to be one of the cancer-inducing compounds in coal-tar,[3] around 1930, investigations concerning structure-activity relationships (SAR) of PAH's were initiated. Since then there has been a considerable interest in trying to predict the biological activity of these compounds from their chemical structure. One hypothesis about the relation between carcinogenicity and structure of PAH's has been the K and L region theory of the Pullmans.[4] They introduced as set of indices, based on electron localization theory, and related these to carcinogenicity. The Pullmans and several others considered the hydrocarbon itself as the ultimate carcinogen while others, for example Dipple *et al.*,[5] have proposed that a cationic sigma complex is the active metabolite. Recently a theory relating PAH carcinogenicity to arene epoxides, diolepoxides and "bay region" carbonium ions has appeared.[6]

A common feature in PAH SAR has been, and still is,[7] to correlate single theoretical or measured variables with the biological activity, that is carcinogenicity. In view of the complex nature of biological systems including carcinogenic processes, it is reasonable that the activity of a compound is not determined by one single variable, but rather by complex interrelations among a number of variables.[8] It is therefore interesting to base SAR's on methods of statistical data analysis which can handle several variables simultaneously, for instance methods of pattern recognition (PaRC).

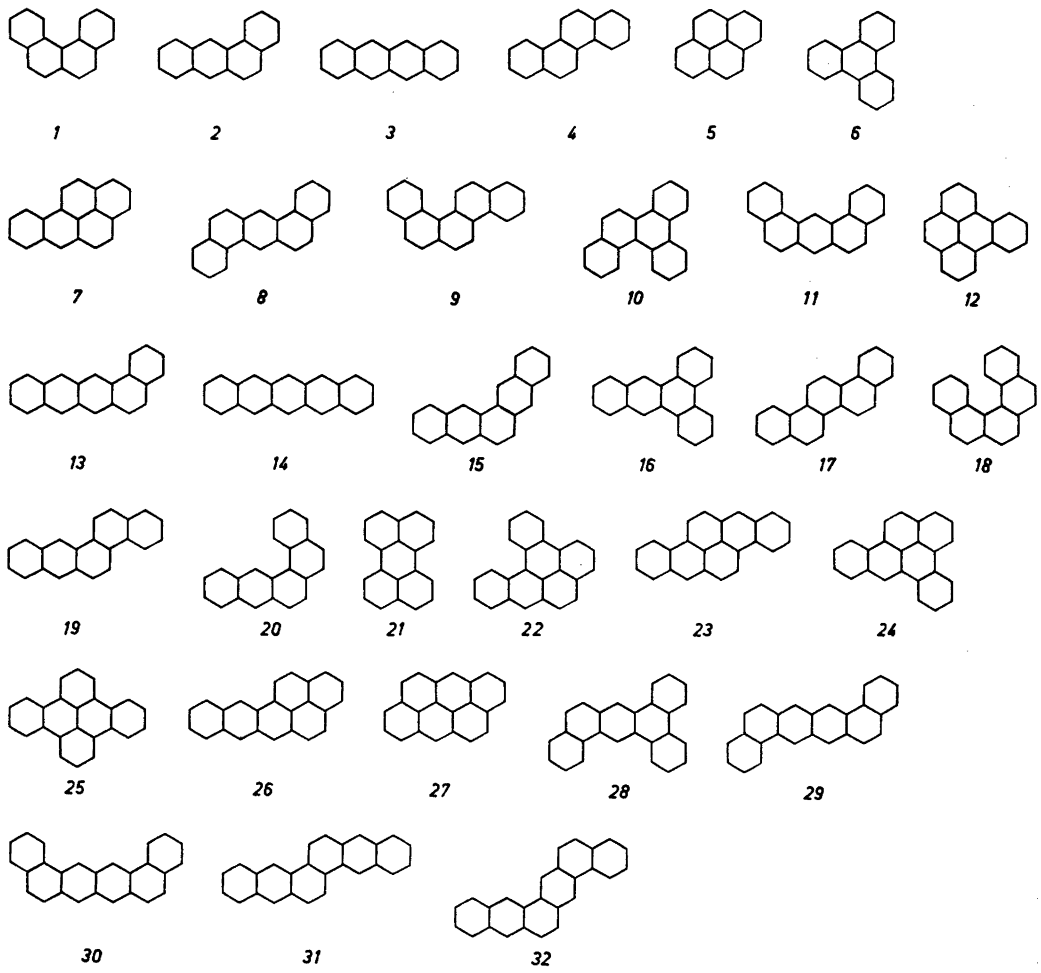
We here report the use of the SIMCA method [9] of pattern recognition in the study of the possible relation between the PAH carcinogenicity and all available variables. We have studied the same compounds as the Pullmans but excluded the four inactive compounds with three rings or less in order to avoid a trivial size factor. We also excluded anthra[3,4-*a*]anthracene since data are incomplete for that compound. This leaves us with thirty-two compounds shown in Table 1.

Variables included in the analysis are as follows:

Theoretical. The Pullmans stated that the presence of a reactive K region, whose complex index $BLE + CLE_{min}$ (bond localization energy + carbon localization energy) does not exceed a certain threshold, is favourable for the appearence of carcinogenic activity. If the molecule also contains an L region, the complex index of this region $PLE + CLE_{min}$ (para localization energy + carbon localization energy) should exceed a certain value, *e.g.* the L region should be as inactive as possible. Four of the Pullman indices, BLE_K, $(BLE + CLE_{min})_K$, PLE_L, and $(PLE + CLE_{min})_L$,[4] are included.

For compounds not having true K or L regions, values have been calculated.[10]

Table 1. Compound structures, numbers, activities and class division. (1) Model I: all compounds together in the first analysis (corresponds to step 1 in Fig. 4). (2) Second analysis (step 3 in Fig. 4). Class Ia: active compounds with moderate and high activity. Class Ib: inactive compounds predicted to have moderate or high activity by model I. The rest of the compounds (slightly active or inactive and with large θ-parameter values) are put in a test set.



Compound	Activity [a]	Class	Compound	Activity [a]	Class	Compound	Activity [a]	Class	Compound	Activity [a]	Class
1	1+		*9*	2+	Ia	*17*	0		*25*	0	
2	1+	Ia	*10*	1−2+		*18*	0		*26*	0−1+	Ib
3	0	Ib	*11*	1−2+	Ia	*19*	0	Ib	*27*	0	Ib
4	0−1+		*12*	0−1+		*20*	0	Ib	*28*	0	Ib
5	0		*13*	0	Ib	*21*	0		*29*	0	Ib
6	0		*14*	0	Ib	*22*	3−4+	Ia	*30*	0	Ib
7	4+	Ia	*15*	0	Ib	*23*	4+	Ia	*31*	0	Ib
8	2+	Ia	*16*	0−1+		*24*	3−4+	Ia	*32*	0	Ib

[a] See Refs. 7, 12, 24 and 25. The activities are weighted values from these articles.

Resonance energies calculated by Hückel MO-method, E_r.[11]

NMBO values for the sigma complexes believed to arise by attack (O_2) at the K region, $1\text{-}a_{or}$.[11]

The energy of the highest filled (or lowest empty) molecular orbital as a measure of the electron-donor (or acceptor) properties, $|k|$.[12]

The energy difference between the highest filled and succesive unfilled levels, $\Delta E_{1,2}$.[13,14]

The superdelocalizability index of the K region, I_K.[15]

The hydrophobicity estimated from hydrophobic fragment constant, *log P*.[16]

The number of symmetry-axes, A_s.[17]

The calculated charge at the most positive centre of PAH radical cations, C_{rc}.[18]

The calculated charge difference between the most and second most positive centre of radical cations, ΔC_{rc}.[18]

The calculated charge difference between the most and second most positive centre of radical cations not located at the K region, $\Delta C_{rc,non-K}$.[18]

Measured. From the absorption spectra of Clar,[19] totally six variables can be extracted, λ_{max}, *log* ε and Δ (bandwidth), for the β- and p-bands respectively.

The overlapping integral determined from the quantum intensity of the fluorescence emission of tryptophan and the molar extinction coefficient of the hydrocarbon, J_1.[12,20]

The ionization potential determined spectroscopically, I_p.[18]

For some compounds a few variables were not available. These were estimated by variable means within each class.

In summary we have included fifteen theoretical, non-measured, and eight measured variables. These include all variables reported for the 32 PAH's in the literature.

Initially the data analysis was made with all variables included while in the second run only theoretical variables were included, the latter approach being more interesting for future prospective studies.

METHODS

With SIMCA the data of each class are approximated by a separate principal components (PC) model.

$$y_{ik}^{(q)} = \alpha_i^{(q)} + \sum_{a=1}^{A} \beta_{ia}^{(q)} \theta_{ak}^{(q)} + \varepsilon_{ik}^{(q)}$$

The index q indicates that the data belong to class q. The data y_{ik}, the value of variable i measured on the object k, is described by the PC model. The parameters α_i and β_{ia}, which are estimated from the class reference set, define the position and direction of the "class". The coefficient θ_{ak}, specific for the kth object in the class, describe where in the "class" the object is situated. The residuals ε_{ik} give measures of how far from the "class" the object lies.

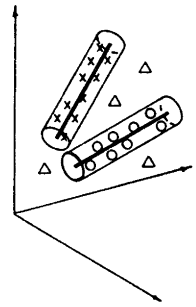


Fig. 1a. A three-dimensional space where each object is represented by a data point. The PC models describe the data with the dimensionality $A = 1$ (equivalent to a line, $A = 2$ is equivalent to a plane in space). The lines are surrounded by confidence regions that consequently enclose each class. Symbol △ represents so called outliers, not belonging to any of the two classes.

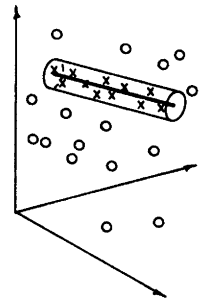


Fig. 1b. A three-dimensional space which shows two classes of objects representing an asymmetric data structure. Only one class can be described by a PC model ($A = 1$, equivalent to a line). The other class is lacking systematic structure. The line is surrounded by a confidence region and the classification of new objects is made by observing if they fall inside or outside this region. The asymmetric case of pattern recognition seems to be a rather common situation in multivariate data analysis and causes failures because it is not recognized by most of these statistical methods and certainly not by analyzing one variable at a time.

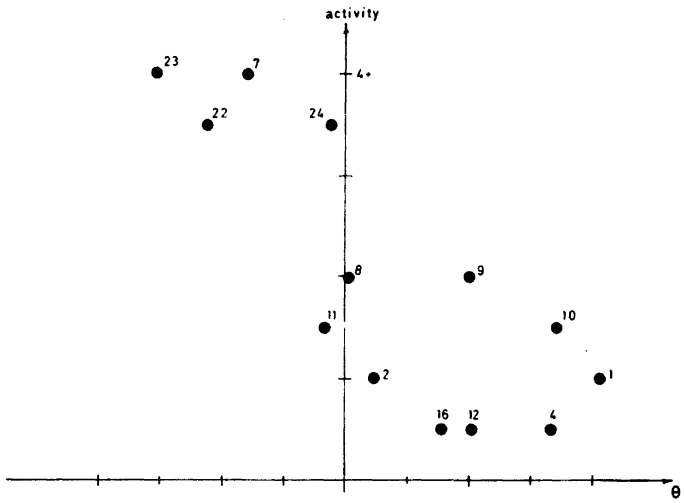


Fig. 2. Plot of the activity *vs.* θ showing the position of each object in the active group of compounds of model I. The lower the θ-parameter value the more active is the compound. Correlation coefficients: $r = 0.800$ ($p < 0.001$) including both theoretical and measured variables and $r = 0.722$ ($p < 0.01$) including only theoretical variables.

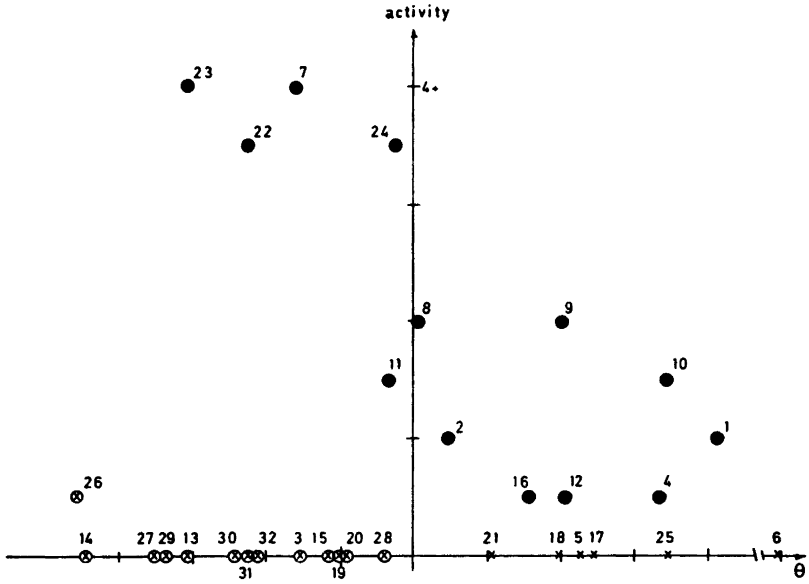


Fig. 3. A plot of the activity *vs.* θ (from the model I analysis) showing the position of all compounds. Inactive compounds with low θ-value (< 0) are predicted to have medium or high activity. The second analysis is based on this first situation and hence class Ia (active compounds) will contain compounds *2, 7–9, 11, 22–24.* Class Ib (inactive compounds) will contain compounds denoted by ⊗ (*3, 13–15, 18–20, 26, 27, 29–32*).

Thus, each "class" is described as an A-dimensional hyper plane in the measurement space (n-dimensional space, n is the number of variables).

Fig. 1a shows the situation for $A = 1$ in three dimensions.

A special case of classification is where one of the two "classes" shows no systematic struc-

ture, while the other class is well described by the PC model. This situation indicates an asymmetric data structure and is illustrated in Fig. 1b. A new object is classified according to its position inside or outside the well-described class. Both Figs. 1a and 1b show confidence regions around the lines ($A = 1$) that fit the objects best.

The principles of the method can shortly be written:

Phase 1. Use of data obtained on compounds with known classification (training or learning or reference set) for the calculation of the *parameter values* for a separate PC model for each class. We can also obtain *the distances* between the classes, and *the modelling power*, that is how much a variable participates in the mathematical description of the classes. *The discrimination power* expresses how much a variable participates in the discrimination between the classes.

Phase 2. Compounds with unknown class assignment (test set) are classified according to the earlier adapted PC model.

Phase 3. Relations between the position of an object in a class (θ-parameter value) and one "external" effect variable. In this case this corresponds to prediction of the level of the biological activity.

Thus, the analysis gives the following information:[22,23]

a. The parameter values of the class PC models.

b. A classification of compounds as active or nonactive based on their fit to the class models.

c. The relevance of each variable, *i.e.* to what extent a variable participates in the description of the classes.

d. A prediction of the level of carcinogenicity of compounds in the active class.

RESULTS

The following analysis were made: *1.* To start with, all compounds were treated together in one single model (I). The determination of A, *i.e.* the number of product terms in the equation given, is estimated using a cross-validation technique.[21] This yields $A = 1$. The parameters α, β and θ are not tabulated.[26]

The positions of the *active* compounds inside this group were related to their carcinogenicity level (Fig. 2), *i.e. activity* against θ. The same result is achieved using only theoretical variables, but with lower correlation coefficient (legend of Fig. 2).

The variables not participating in the class model I were deleted. They were: *log* $\varepsilon_{\beta,p}$, A_s, *1-*a_{or}, $\Delta_{\beta,p}$, ΔC_{rc} and $\Delta C_{rc,non-K}$. Variables that best describe model I are (in falling order): $(PLE + CLE_{min})_L$, $\lambda_{max,\beta}$, ΔE_1 and I_p.

2. Level prediction of compounds. Some of the inactive compounds were predicted to have high activity (because of their low θ-value, *i.e.* ⊗ in Fig. 3). The reason might be that inactive compounds, which are predicted to have high activity, undergo some inactivation process or have some other critical factor inhibiting their activity.

3. To investigate this we analyzed the compounds with predicted high activity as two classes. Class Ia contains compounds that in fact are active (*2*, *7*, *8*, *9*, *11*, *22*, *23* and *24*) and class Ib contains compounds predicted to have medium or high activity, but actually are inactive (⊗ in Fig. 3). The rest of the compounds (slightly active or inactive) are put in a test set. The data analysis shows that these two classes are statistically well-defined, *i.e.* the situation shown in Fig. 1a with the reservation that since $A = 2$ the reader has to imagine a plane instead of each drawn line.

The classes also give a good classification. Of the eight active and thirteen inactive compounds, one active and two of the inactive are classified as outliers (*11*, *15* and *31*), one active belongs to both classes but is closer to the inactive class (*8*). The rest (17 of 21) are correctly classified. One compound of the test set (*17*) is classified on the border of the active class, all others do not belong to any of the two classes. Compound *17*, picene, has in fact been reported as slightly active in a recent review.[24] Variables that best describe the classes Ia and Ib are: ΔE_1, $\lambda_{max,p}$, $(BLE + CLE_{min})_K$, I_K, I_p and $(PLE + CLE_{min})_L$.

Variables which do not contribute in defining the class structures are: *log* $\varepsilon_{\beta,p}$, J_1, A_s, $\Delta_{\beta,p}$ and C_{rc}.

If we only look at the active class (Ia), the following variables do best describe this class: I_p, $(BLE + CLE_{min})_K$, $\lambda_{max,p}$ and I_K.

4. Validation of the latter classification was done using a method of leaving out one compound of each class, putting them into the test set, then fitting the mathematical models to the remaining compounds of class Ia and Ib, and in the end classifying the compounds left out. This was done repeatedly so that each compound was left out once. A high prediction rate of the compounds left out indicates stable class structures.

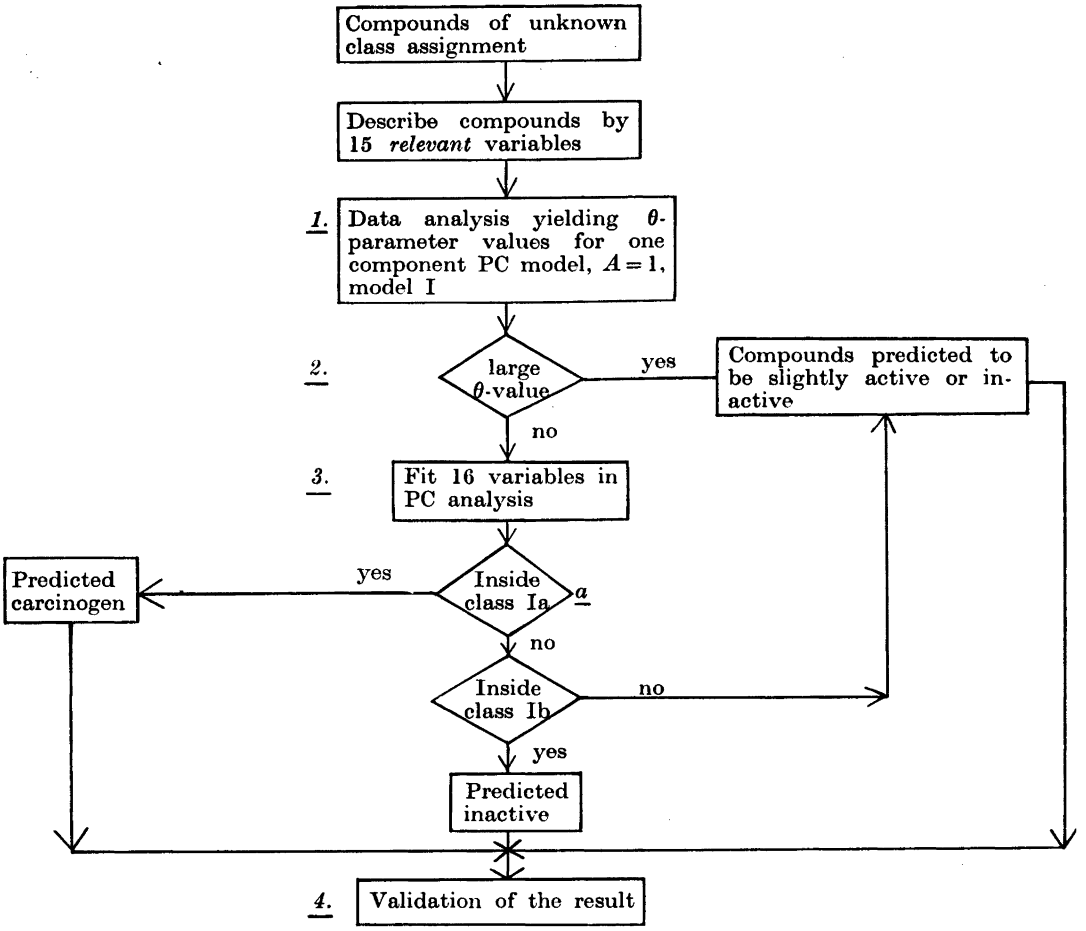


Fig. 4. Our procedure for the classification of PAH's. The numbers refer to the numbering of results in the article. [a] Classification rate (class Ia or Ib) in this work 18/21 ($\chi^2 = 6.75$, $p < 0.05$).

The validation shows that three compounds of class Ia are wrongly classified (*8*, *11* and *23*), *e.g.* three out of eight compounds are incorrectly classified in class Ia and none of thirteen in class Ib. This classification rate (18/21) is significantly better than chance ($\chi^2 = 6.75$, $p < 0.05$). The same validation procedure was used for the prediction of the level of the active compounds. The correlation coefficient between predicted and observed activity of "left out" compounds is 0.793 ($p < 0.001$).

DISCUSSION

In conclusion we have shown that it is possible to use theoretical and measured variables to

1. classify PAH's as carcinogenic or non-active and

2. make predictions about the level of the activity of a compound.

Our procedure for the classification of PAH's is shown in Fig. 4. The results indicate that two "factors" influence the carcinogenicity. The first factor influences the level of activity, the second factor in some way inhibits the carcinogenicity even where the first factor corresponds to high activity (Ib). The interpretation of these two factors is currently beyond the scope of our study.

REFERENCES

1. Pott, P. *Chirurgical Observations* (1775). Reprinted in *Natl. Cancer Inst. Monograph 10* (1963) 7.
2. Clar, E. *Ber. Dtsch. Chem. Ges. 62* (1929) 350.
3. Cook, J. W., Hewitt, C. L. and Hieger, I. *J. Chem. Soc.* (1933) 395.
4. Pullman, A. and Pullman, B. *Cancer Res. 3* (1955) 117.
5. Dipple, A., Lawley, P. D. and Brooks, P. *Eur. J. Cancer 4* (1968) 493.
6. Jerina, D. M., Lehr, R. E., Yagi, H., Hernandez, O., Dansette, P. M., Wislocki, P. G., Wood, A. W., Chang, R. L., Levin, W. and Conney, A. H. In de Serres, F. J., Fonts, J. R., Bend, J. R. and Philpot, R. M., Eds., *In Vitro Metabolic Activation in Mutagenesis Testing*, Elsevier/North Holland Biomedical Press, Amsterdam 1976, p. 159.
7. Berger, G. D., Smith, I. A., Seybold, P. G. and Serve, M. P. *Tetrahedron Lett. 3* (1978) 231.
8. Berenblum, I. In Bergmann, E. D. and Pullman, B., Eds., *Physico-Chemical Mechanisms of Carcinogenesis*, Israel Academy of Sciences and Humanities, Jerusalem 1969, p. 321.
9. Wold, S. and Sjöström, M. In Kowalski, B., Ed., *Chemometrics: Theory and Practice, ACS Symp. Ser. No. 52* (1977) 243.
10. Sung, S. S. *C. R. Acad. Sci. 274* (1972) 1597.
11. Scribner, J. D. *Cancer Res. 29* (1969) 2120.
12. Pullman, A. *Biopolymers Symposia No. 1* (1964) 47.
13. Mason, R. *Nature 181* (1958) 820.
14. Sung, S. S. *C. R. Acad. Sci. 264* (1967) 189.
15. Mainster, M. A. and Memory, J. D. *Biochim. Biophys. Acta 148* (1967) 605.
16. Nys, G. G. and Rekker, R. F. *Eur. J. Med. Chem. 9* (1974) 361.
17. Arcos, J. C. and Argus, M. F. *Chemical Induction of Cancer*, Academic, New York and London 1974, Vol. IIA.
18. Pullman, B., Pullman, A., Umans, R. and Maigret, B. In Bergmann, E. D. and Pullman, B., Eds., *Physico-Chemical Mechanisms of Carcinogenesis*, Israel Academy of Sciences and Humanities, Jerusalem 1969, p. 325.
19. Clar, E. *Polycyclic Hydrocarbons*, Academic, London and New York 1964, Vols. 1 and 2.
20. Birks, J. B. *Nature 190* (1961) 232.
21. Wold, S. *Pattern Recognition 8* (1976) 127.
22. Dunn, W. J., III, Wold, S. and Martin, Y. C. *J. Med. Chem. 21* (1978) 922.
23. Albano, C., Dunn, W., Edlund, U., Johansson, E., Nordén, B., Sjöström, M. and Wold, S. *Anal. Chim. Acta Comp. Optim.* 1978. *In press.*
24. Jones, D. W. and Matthews, R. S. *Prog. Med. Chem. 10* (1974) 159.
25. Dipple, A. In Searle, C. E., Ed., *Chemical Carcinogens, ACS Monograph 173* (1976) 245.
26. Listings of the parameter values (α, β and θ) for the two data analyses are available on request.

Received May 9, 1978.