Bull. Environ. Contam. Toxicol. (1985) 34:1–9

Bulletin of Environmental Contamination and Toxicology

Structure-toxicity Relationships of Selected Naphthalene Derivatives II. Principal Components Analysis

T. Wayne Schultz[1] and Michael P. Moulton[2]

[1]College of Veterinary Medicine, University of Tennessee, Knoxville, TN 37901-1071, and [2]Department of Zoology, University of Tennessee, Knoxville, TN 37916

Currently the best method of quantitatively predicting biological activity is by regression analysis. Regression analysis assumes that biological activity is a function of physiochemical properties and chemical structure. Therefore, biological activities including toxicity are typically modeled by equations of the general form:

$$\log C^{-1} (BR) = f (SC_1, SC_2,...,SC_n)$$

where $\log C^{-1}$ (BR), the log of the inverse of the concentrations (C) in moles/liter to cause a given biological response (BR), is the dependent variable and is predicted by means of multiple linear regression analysis with a variety of substituent constants (SC_i) used as independent variables. This relationship (i.e. equation 1) is typically tested using the following statistical model:

$$\log C^{-1}(BR) = B_0 + B_1\ SC_1 + B_2\ SC_2 +...+ B_n\ SC_n + E$$

where the null hypothesis states that

$$B_1 = B_2 = ... = B_n = 0.$$

This methodology enables one to predict the BR of a new derivative using its SC_i in conjunction with the estimates of the B_i. Also, it can aid in the identification of SC_i that do not make a significant contribution to the overall model sums of squares. Recently we have used this method to quantitate the biological activity of a series of one-position naphthalene derivatives (Schultz et al. 1983).

There is, however, a problem associated with this methodology. The problem arises from the fact that SC_i are often intercorrelated (i.e. nonorthogonal). Since the SC_i are not orthogonal, the order in which they enter the regression model is critically important. If the variance accounted for by variable 1 nearly encompasses that of variable 2, the model would suggest that variable 2 was not of any great importance. However if variable 2 was considered first, the model would provide the same conclusion as above, for variable 1. This problem becomes increasingly complex with the addition of independent variables.

An alternative to this statistical quagmire is Principal Components Analysis (PCA). In PCA, the original variables are redefined by new variables (i.e. the principal components) which are linear combinations of the original variables. The advantages of PCA are two-fold. Firstly, the new variables defined by PCA are orthogonal (i.e. not interrelated). Thus they can be entered into a multiple regression model in any order without altering their sequential sums of squares contributions. This makes interpretation much simpler.

Secondly, PCA has the potential to reduce the number of variables. If the original variables are at all correlated, PCA will identify groups of these variables that are especially correlated with one another. For example, the five original variables may in fact represent two uncorrelated groups or PC. These two PC may account for a large proportion (e.g. > 80%) of the total variance of the original variables.

Therefore, statistical inference such as a quantitative structure-activity relationship (QSAR) based on PCA not only provides a model with orthogonal independent variables, but also one that may have fewer independent variables while still identifying which independent variables account for variation in some dependent variable. It is the purpose of this investigation to examine the interrelationship between seven substituent constants using PCA, and to attempt to predict toxicity of a series of napthalene derivatives with the resulting PC.

MATERIALS AND METHODS

The biological activity data used in these investigations were obtained from Schultz et al. (1983). The biological descriptor, logarithm of the biological response (log BR), is defined as the reciprocal of the IGC_{50} value where the IGC_{50} is the concentration (mmol/L) required to inhibit the growth of Tetrahymena pyriformis strain GL-C in axenic culture by 50%. The IGC_{50} values were determined from a concentration/probit percent control absorbance regression analysis as described by Schultz (1983). The 18 napthhalene derivatives assayed were selected based on: (1) commercial availability; (2) cluster group distribution (Hansch and Leo 1979); and (3) availability of SC.

Seven SC were selected as initial variables. These include: (1) the hydrophobic parameter, PI (Fujita et al. 1964); (2) the corrected molar volume term of molar refractivity, MR (Hansch et al. 1973); (3) the polar electronic parameter, F (Swain and Lupton 1968); (4) the resonance electronic parameter, R (Swain and Lupton 1968); the discrete variables (5) the ability to accept hydrogen, Ha and (6) to donate hydrogen, Hd; and (7) the single bond fragment molecular connectivity index, $^{1}X^{V}_{sub}$ (Kier 1980). These values were obtained from Schultz et al. (1983).

Principal component analysis was used to define new orthogonal variables. These principal components were subsequently used to generate QSAR using the general linear models procedure. Each PC is an eigenvector of the correlation matrix of the original variables based on standardized measurements. Because the correlation matrix is symmetric, eigenvectors corresponding to unique eigenvalues are orthogonal (Harmon 1976). The eigenvalues of this matrix are the variances accounted for by each PC.

The original measurements are standardized so the total variance is equal to the rank of the correlation matrix. Principal components with eigenvalues less than 1 are typically ignored since these do not account for even an average amount of the total variance (e.g. Wiens and Rotenberry 1980). If the original correlation matrix is of rank n, there will generally be m eigenvalues > 1 (where n > m).

Correlation coefficients between the original variables and the PC are calculated by the equation:

$$a_{ij} = \sqrt{\lambda_i} \cdot v_{ij}$$

where a_{ij} is the correlation between the jth variable and the ith PC,

$$\lambda_i = \text{the eigenvalue of the ith PC,}$$

and

$$v_{ij} = \text{the jth element of the ith eigenvector.}$$

The matrix of a_{ij} terms is used for identifying clusters of variables. A cluster occurs when two or more of original variables have their highest correlation with a particular component.

Often interpretation of clusters of variables is enhanced by a rigid rotation of the a_{ij} matrix. One such rotation is the varimax rotation. This involves rotating the axes so that the variances of the squared a_{ij} terms with respect to each PC are maximized. Details of the procedure may be found in Harmon (1976). All analyses were computer assisted using Statistical Analysis Systems software (SAS Institute Inc, 1982).

RESULTS AND DISCUSSION

The eigenvalues, variances, and cumulative variances for the factors attained by PCA of the seven substituent constants are presented in Table 1.

Table 1 Eigenvalues and Variances of Principal Components Based on Seven Parameters

Component	1	2	3	4	5	6	7
Eigenvalue	2.427	1.892	1.237	0.797	0.382	0.190	0.074
Variance	0.347	0.270	0.177	0.114	0.055	0.027	0.011
Cumulative Variance	0.347	0.617	0.794	0.908	0.962	0.989	1.000

Notice that components 5 through 7 contribute little to the cumulative variance.

The component pattern (i.e. correlation coefficient of each parameter with each component) for the first four components is presented in Table 2.

This analysis reveals that component 1 is a clustering of PI, MR, Hd and $1_{X^v sub}$. Component 2 is dominated by Ha while components 3 and 4 are dominated by F and R, respectively. In an effort to enhance the interpretation of variable clusters a varimax rigid rotation procedure was employed.

The seven parameter rotated component pattern is presented in Table 3.

Table 2 Seven Parameter Component Pattern

	Component			
Parameter	1	2	3	4
PI	0.698*	-0.537	-0.121	-0.258
MR	0.771*	0.490	-0.302	-0.002
F	0.261	0.389	0.745*	-0.423
R	0.405	-0.012	0.578	0.697*
Ha	-0.140	0.951*	0.071	-0.101
Hd	-0.733*	0.373	-0.271	0.172
$1_{X^v sub}$	0.739*	0.412	-0.404	0.163

*parameter's major component

Table 3 Seven Parameter Rotated Component Pattern

	Component			
Parameter	1	2	3	4
PI	0.200	0.919*	0.012	0.012
MR	0.891*	0.195	0.093	0.061
F	0.020	-0.003	0.975*	0.123
R	0.081	0.018	0.115	0.982*
Ha	0.272	0.451	0.318	-0.083
Hd	-0.159	-0.285	-0.138	-0.148
$1_{X^v sub}$	0.974*	0.037	-0.024	0.064

*parameter's major component

These results show component 1 is dominated almost equally by MR and $1_X v_{sub}$. Component 2 is dominated by PI. Component 3 is dominated by F while component 4 is dominated by R. The variables Ha and Hd had their highest correlations with components 6 and 7 respectively. Recall for Table 1 that components 6 and 7 account for < 4% of the total variance in this system. This suggests (but does not prove) that Ha and Hd might have little influence in predicting log BR. Moreover, PCA generally involves only continuous variables. For these two reasons, the PCA was repeated after deleting Ha and Hd. Analysis following the removal of the discrete variables Ha and Hd yielded the same general results.

Table 4 lists the eigenvalues, variances, and cumulative variances for the factors attained by PCA of the five substituent constants.

Table 4 Eigenvalues and Variance of Principal Components Based on Five Parameters

Component	1	2	3	4	5
Eigenvalue	2.086	1.176	0.828	0.760	0.150
Variance	0.417	0.235	0.166	0.152	0.030
Cumulative Variance	0.417	0.652	0.818	0.970	1.000

In this analysis component 5 contributes relatively little to the cumulative variance. The five parameter component pattern is listed in Table 5.

Table 5 Five Parameter Component Pattern

	Component			
Parameter	1	2	3	4
PI	0.531	-0.178	0.813*	0.155
MR	0.924*	-0.132	-0.213	0.084
F	0.235	0.781*	-0.012	0.571
R	0.344	0.681*	0.150	-0.628
$1_X v_{sub}$	0.882*	-0.229	-0.306	-0.088

*parameter's major component

Analysis based on four components shows component 1 to be a clustering of MR and $1_X v_{sub}$. Component 2 is a clustering of F and R while the third principal component is dominated by PI. The five parameter rotated component pattern is given in Table 6.

Table 6 Five Parameter Rotated Component Pattern

	Component			
Parameter	1	2	3	4
PI	0.167	0.003	0.985*	0.037
MR	0.917*	0.115	0.187	0.055
F	0.040	0.992*	0.036	0.120
R	0.080	0.121	0.036	0.989*
$^1\chi^v_{sub}$	0.966*	-0.032	0.096	0.077

* parameter's major component

This rotational refinement separates R from F in component 2 by placing R in component 4.

Regression analysis was conducted on factors derived from the varimax rotation. These results are summarized in Table 7. The overall regression model is highly significant ($p < .0003$). The relative contributions of the factors decreases monotonically. Factors 4 and 5 do not contribute significantly to the model (see Table 7). The best regression model was thus:

$$\text{Log BR} = 0.6183 + 0.1993\,(\text{Factor 1}) + 0.1867\,(\text{Factor 2}) + 0.1556\,(\text{Factor 3}) \qquad r^2 = 0.826$$

This model suggests that R has little influence in predicting activity.

Table 7 Summary of Regression Model and Relative Contributions of Factors.

Source	DF	SS	PR > F
Model	5	1.7783	0.0003
Error	12	0.3754	
Factor 1	1	0.6756	0.0006
Factor 2	1	0.5927	0.0009
Factor 3	1	0.4121	0.0035
Factor 4	1	0.0874	0.1205
Factor 5	1	0.0104	0.5740

In the classical regression model (eq. 2), the B_i are estimated, based on a set of observations. With such a model, log C^{-1}(BR) for some new derivative can be estimated by computing the sum of the products of its SC_i values and the corresponding B_i values.

The PCA model is also predictive. The only difference in PCA regression modeling is that instead of using raw and correlated SC_i values, the PCA model uses orthogonal linear combinations of the SC_i values.

The SC_i values for a new derivative are easily converted to PC scores. First each SC_i value for the new derivative is standardized. To do this subtract the mean SC_i value (for the SC in question) and divide this difference by the corresponding standard deviation (means and standard deviations are listed in Table 8). This provides scaled SC_i values.

Principal component scores can then be calculated by multiplying the scaled SC_i values by the scoring coefficients (Table 9).

For example the Factor 1 score for observation number 1 is obtained by first scaling the original SC_i values. The scaled SC_i values (SC_i*) are:

$$\frac{PI_i - \overline{PI}}{PI_{SD}} \quad , \quad \frac{MR_i - \overline{MR}}{MR_{SD}} \text{ , etc.}$$

These scaled values are then multiplied by their respective scoring coefficient (K_i). The sum of these products is equal to the Factor 1 score for observation number.

$$PI_i^* \; K_{PI} + MR_i^* \; K_{MR} + \ldots + {}^1X^v i^* \; K^1X^v{}_{sub} =$$

FACTOR 1 for observation 1.

$$\frac{(1.15) - (-0.103)}{0.680} (-0.145) + \frac{(17.24) - (7.641)}{4.105} (0.449) +$$

$$\frac{(.51) - (.255)}{0.215} (-0.023) + \frac{(-0.09) - (-0.148)}{0.270} (-0.066) +$$

$$\frac{(0.733) - (0.624)}{0.327} (0.641) = 0.955$$

Table 8 Mean and Standard Deviations of SC_i

	Parameter				
	PI	MR	F	R	${}^1X^v$
MEAN	-0.103	7.641	0.255	-0.148	0.624
STD	0.680	4.105	0.215	0.270	0.327

Table 9 Scoring Coefficients

	Factor				
Parameter	F1	F2	F3	F4	F5
PI	-0.145	0.008	1.043	-0.027	-0.187
MR	0.449	-0.041	-0.100	-0.031	1.866
F	-0.023	1.028	0.009	-0.124	-0.279
R	-0.066	-0.124	-0.028	1.033	0.073
$^1\chi^v_{sub}$	0.641	0.005	-0.084	-0.046	-1.734

There are limitations to this technique. The scoring coefficients obtained for napthalene derivatives for example, can not be used to predict log C^{-1} (BR) for non-napthalene derivatives. Rather, a separate PCA should be performed, for each "family" of compounds. The PCA we performed used a highly heterogeneous set of napthalene derivatives. Thus with a new napthalene derivative log C^{-1} (BR) could be predicted by our scoring coefficients. This technique might become invalid, however, if the new napthalene derivative had SC_i values that were outside the range of values in our analysis.

The results of the PCA and subsequent multiple regression analysis yielded similar results to a regression analysis of log BR vs. the raw SC_i (Schultz et al. 1983). The first PC was defined by the variables MR and $^1\chi^v_{sub}$. The second PC was defined by the variable F, and the third PC by the variable PI. In the subsequent multiple regression model of log BR against the PC, only these three factors made a significant contribution to the overall model sum of squares (see Table 7).

A comparison of squared correlation coefficients for traditional regression models of log BR versus each of the SC_i has shown MR to be the best single descriptor for this data set followed in turn by $^1\chi^v_{sub}$, F and PI (Schultz et al. 1983). Molar refractivity accounts for 44.5% of the variation in biological activity. Multiple regression analyses (Schultz et al. 1983), on-the-other-hand, have demonstrated that Ha and PI formed the best two descriptor model followed by MR and F. The former explains 71% of the variability in log BR, whereas the latter explains 62%. The best three parameter models were Ha, PI and F and PI, MR, and F with r^2 values of 0.758 and 0.728, respectively. The independent variables PI, MR, F and R formed the best four parameter model with an r^2 value of 0.812.

The results of the PCA are in complete accord with those of the traditional regressions of log BR versus the raw SC_i. This is especially true if the discrete variable Ha is not used. For example, in the present analysis only three components are significant (see Table 7). These components are dominated by MR, F and PI, respectively, (see Table 6). These are the same parameters which form the best three variable traditonal model if Ha is not included.

The molecular descriptors MR and $^1\chi^v_{sub}$ consistently load on the same component i.e. these two parameters are highly correlated: The 18 derivatives tested in this study, have an r^2 value of 0.830.

The question arises as to why one should employ PCA if the resulting multiple regression model yields results that coincide with those of standard regression analyses? The answer is that in standard regression analyses, interpretation is confused by correlations among the independent variables. And this problem becomes more complex as the number of independent variables (SC_i) increases. PC analysis defines orthogonal variables, and these new variables can be used in multiple regression models. Interpretation of multiple regression models based on orthogonal variables is much simpler. Moreover, one could incorporate variables other than SC_i in a PCA analysis without loss of clarity of interpretation.

Acknowledgements

Supported in part by Biomedical Research Support Grant NIH-PR05845-02

REFERENCES

Fujita T, Iwasa J, Hansch C (1964) A new substituent constant, π, derived from partition coefficients. J Amer Chem Soc 86:5175-5180.

Hansch C, Leo A, Unger SH, Kim KH, Nikaitoni D, Lien EJ (1973) "Aromatic" substituent constants for structure-activity correlations. J Med Chem 16:1207-1216.

Harmon HH (1976) Modern factor analysis. 3rd ed. The Univ of Chicago Press, Chicago, 487 pp.

Kier LB (1980) Molecular connectivity as a description of structure for SAR analysis. In: Yalkowsky SH (ed) Physical Chemical Properties of Drugs, Mercel Dekker, New York, pp. 277-319.

SAS Institute Inc (1982) SAS User's Guide: Statistics, 1982 edition. SAS Institute Inc, Cary, NC, 584 pp.

Schultz TW, Dumont JN, Sankey FD, Schmoyer RL, Jr (1983) Structure activity relationships of selected naphthalene derivatives. Ecotox Environ Safety 7:191-203.

Swain CG, Lupton EC, Jr (1968) Field and response components of substituent effects. J Amer Chem Soc 90:4328-4337.

Wiens JA, Rotenberry JT (1980) Patterns of morphology and ecology in grassland and shrubsteppe bird populations. Ecol Monogr 50:287-308.

Received February 20, 1984; accepted March 19, 1984.