Correlation of Graph-Theoretical Parameters with Biological Activity

Gordon G. Cash* and Joseph J. Breen

United States Environmental Protection Agency, Office of Pollution Prevention and Toxics, Washington, D.C. 20460

Received September 29, 1992

Ośmiałowski and Kaliszan calculated graph-theoretical indices for substituted isonicotinic hydrazides and used simple and multiple regression to search (unsuccessfully) for correlations with biological activity. The present paper describes successful searches for correlation in the same data set using principal component analysis (PCA) with multivariate outlier testing and also using stepwise multiple regression. Following PCA, correlation with biological activity always appeared in the second principal component, not the first, that is, after projection of the data points into the $(n-1)$-space orthogonal to the first principal component axis. In that space, the principal component score was a more accurate predictor of biological activity than were equations provided by multiple regression or stepwise multiple regression using the underlying variables. A multivariate outlier test identified one observation as discordant, and removing that observation improved prediction further.

INTRODUCTION

Ośmiałowski and Kaliszan[1] calculated for a series of 2-substituted isonicotinic acid hydrazides (Table I) a set of 14 graph-theoretical parameters, namely, Kier and Hall generalized molecular connectivity indices ($^1\chi^v$, $^2\chi^v$);[2] Kier indices of molecular shape ($^1\kappa$, $^2\kappa$, $^3\kappa$);[3] information indices of neighborhood symmetry (IC_0, IC_1);[4] structural information content indices ($SIC_{0,1}$, $CIC_{0,1}$, $TIC_{0,1}$);[5] and the Wiener index (W).[6] Using these parameters and molecular weight (MW) as independent variables (15 in all), they applied simple and multiple regression in an unsuccessful attempt to establish a linear–logarithmic correlation with biological activity, as measured by minimum inhibitory concentration (MIC) against *Mycobacterium tuberculosis*.[7]

We have had some success in the past in finding order in seemingly random data through principal component analysis (PCA).[8] Therefore, we reanalyzed the data of Ośmiałowski and Kaliszan using PCA and found a good correlation ($r = -0.798$) with a principal component derived from five of their parameters, specifically, $^3\kappa$, IC_0, IC_1, SIC_0, and MW. Using the square root transformation of the principal component scores improved the correlation to $r = -0.891$. (The square root transformation was selected on the basis of an improved figure for the Kolmogorov–Smirnov test for normality.) The resulting equation is not necessarily the best possible relationship between the five independent variables and log (1/MIC); it merely serves as a concrete illustration of the ability of PCA to discover correlations that other statistical techniques miss. *Stepwise* multiple regression also produced good correlation with a subset of four variables.

METHODS

All stepwise multiple regressions were run with F-to-add and F-to-remove both equal to 3. Because of differences in scale, particularly between molecular weight and the graph-theoretical parameters, all principal component analyses reported in this paper were performed on correlation matrices. PCA, multivariate outlier detection, and graphics were done with SCOUT, a public-domain software package developed under contract to the U.S. Environmental Protection Agency. SCOUT may be obtained free of charge by sending a formatted, high-density diskette (either size) to Mr. Jack Teuschler, U.S. EPA, Center for Environmental Research Information, 26 W Martin Luther King Dr., Cincinnati, OH 45268.

Table I. Results of Attempts by Various Regression Techniques To Find "Best" Relationship between Various Sets of Independent Variables and log (1/MIC)

independent variable(s)	exclude obs. 10	regression technique[a]	adjusted R^2	univariate r	standard error of estimate
15	no	M	0.6642		0.4926
15 (4)[b]	no	SM	0.8793		0.2953
5[c]	no	M	0.6187		0.5249
5 (2)[b]	no	SM	0.6886		0.4743
5	yes	M	0.8507		0.3365
5 (3)[b]	yes	SM	0.8632		0.3222
PC2a[d]	no	S	0.6124	–0.7979	0.5292
PC2a	yes	S	0.7738	–0.8882	0.4141
√PC2a	no	S	0.7765	–0.8891	0.4018
√PC2a	yes	S	0.8586	–0.9317	0.3275
PC2b[e]	yes	S	0.7468	–0.8739	0.4383
√PC2b	yes	S	0.7738	–0.8882	0.4141

[a] M = multiple; SM = stepwise multiple; S = simple. [b] Number in parentheses is the number of variables included in the final multiple regression equation. [c] $^3\kappa$, IC_0, SIC_0, IC_1, and MW. [d] Second principal component scores derived from 17 observations. [e] Second principal component scores derived from 16 observations.

Many of the statistical runs reported in this paper utilized a subset of Ośmiałowski and Kaliszan's 15 variables, namely, $^3\kappa$, IC_0, SIC_0, IC_1, and MW. These were selected by inspecting Ośmiałowski and Kaliszan's correlation table for a small set that was generally free of large intercorrelations. For calculation of a second principal component that correlated well with log (1/MIC), these five variables comprised a "best" subset in the sense that either adding any other variable or taking away any one of the five made the correlation worse.

For some of the studies described below, it was necessary to project the data points into the $(n-1)$-space orthogonal to the first principal component axis and then describe the points thus projected in terms of the original parameters. That operation was accomplished as follows: If **A** is the original data matrix (Table II) and **P** is the matrix of prinipal component scores (Table III) from SCOUT, then there exists a transformation matrix $\mathbf{T} = \mathbf{P}^{-1}\mathbf{A}$, where $\mathbf{P}^{-1}$ is the generalized inverse of **P**. Data points in principal component space are

Table II. Loadings of Principal Components and Percent Variation Explained

	from 17 observations						from 16 observations					
	% var	$^3\kappa$	IC_0	SIC_0	IC_1	MW	% var	$^3\kappa$	IC_0	SIC_0	IC_1	MW
PC1	71.88	−0.3751	0.4799	0.5139	0.4487	−0.4044	75.97	−0.4027	0.4702	0.4981	0.4573	−0.3993
PC2	21.04	0.6230	0.3707	0.1329	0.4369	0.5157	16.81	0.5715	0.4084	0.1826	0.3833	0.5713
PC3	5.48	−0.5031	0.2758	0.2479	−0.3902	0.6761	5.47	−0.5925	0.2083	0.1987	−0.3912	0.6427
PC4	1.55	−0.4641	−0.2673	−0.4286	0.6617	0.3027	1.70	−0.3972	−0.2858	−0.4540	0.6877	0.2854
PC5	0.05	−0.0520	0.6962	−0.6878	−0.1333	−0.1475	0.05	−0.0489	−0.6979	−0.6877	−0.1340	−0.1403

Table III. Principal Component Scores

	from 17 observations					from 16 observations				
obs	PC1	PC2	PC3	PC4	PC5	PC1	PC2	PC3	PC4	PC5
1	1.646	−0.504	−0.288	−0.280	−0.079	1.659	−1.491	−0.166	−0.358	−0.095
2	0.702	−1.245	−0.321	0.105	0.006	0.723	−1.263	−0.189	0.064	−0.005
3	−0.226	−0.928	−0.391	0.391	−0.001	−0.204	−0.976	−0.266	0.319	−0.008
4	−1.608	−0.689	−0.289	−0.294	0.058	−1.606	−0.735	−0.217	−0.251	0.059
5	−2.563	0.090	−0.586	−0.435	−0.027	−2.580	−0.004	−0.593	−0.319	−0.020
6	1.032	−0.054	−0.074	0.231	0.075	1.039	−0.051	−0.064	0.204	0.071
7	−0.066	0.501	−0.328	0.209	0.041	−0.071	0.456	−0.363	0.249	0.043
8	1.310	−0.849	−0.191	0.077	0.000	1.324	−0.844	−0.110	0.024	−0.010
9	−0.019	1.473	−0.397	−0.022	0.018	−0.047	1.421	−0.550	0.055	0.025
10	−1.038	2.255	−0.610	−0.061	−0.018					
11	−2.262	0.118	0.149	0.275	0.026	−2.253	0.089	0.180	0.332	0.035
12	3.041	0.236	0.368	−0.410	0.076	3.026	0.334	0.278	−0.512	0.068
13	2.646	0.790	0.634	−0.374	−0.039	2.625	0.906	0.491	−0.464	−0.041
14	2.398	1.167	0.432	0.482	−0.052	2.394	1.233	0.314	0.423	−0.052
15	−2.175	−0.679	1.230	0.001	0.053	−2.153	−0.579	1.315	−0.043	0.061
16	−3.055	0.249	0.811	−0.044	−0.078	−3.053	0.286	0.810	−0.003	−0.063
17	0.235	−0.931	−0.148	0.222	−0.058	0.255	−0.943	−0.036	0.193	−0.065

transformed back into the original parameter space according to **A** = **PT** − **Y**. **Y** is a difference matrix which compensates for the fact that, by convention, PCA shifts the origin along the first principal component axis until the mean of the first principal component scores is zero. **Y** is computed as **PT** − **A**. That **Y** shifts each observation by the same amount is apparent from the fact that the rows of **Y** are all identical. Let **Q** be **P** with the first column set to all zeroes, that is, with each observation shifted in principal component space parallel to the first principal component axis until PC1 = 0 and with its position in the other four dimensions left unchanged. The data points thus projected are transformed back into the original parameter space according to **B** = **QT** − **Y**, where **B** is the new matrix of projected data. Recall that, by convention, the mean of PC1 in **P** is zero. The mean of PC1 in **Q** is also zero because all the individual PC1s are zero. Since the PC1s are the same linear combination of the underlying variables for each observation, we see the expected result that the mean for *each variable* in **B** is the same as it was in **A**.

RESULTS

When restricted to the five-parameter subset, SCOUT identified observation 10 (R = $CH_3CONHCH_2$) as a multivariate outlier at the $\alpha = 0.10$ level. Subsequent procedures were tested both with and without this observation.

Table I presents results of various attempts to discover correlation between biological activity and the graph-theoretical parameters. The adjusted R^2 is related to multivariate R^2 (and to univariate r^2) by eq 1[9]

$$\text{adjusted } R^2 = 1 - \left(\frac{n-1}{n-(p+1)}\right)(1-R^2) \qquad (1)$$

where n is the number of observations and p is the number of independent variables (not including the constant) in the regression equation. Adjusted R^2 is useful for comparing multivariate regressions with different numbers of independent variables. Adding another independent variable always increases R^2 but may decrease adjusted R^2; Table I shows that was the case for several regression pairs in the present study. Ordinarily, one would not report adjusted R^2 for a univariate regression ($p = 1$), but we included those values in Table I for purposes of comparison. Variables labeled $\sqrt{PC}$ are actually $\sqrt{PC - PC_{min}}$, since some PC scores were negative.

Table IV. Effects on Stepwise Regression Results of Projecting Data Points into 4-Space Orthogonal to First Principal Component Axis

obs used	before/after projection	variables selected by stepwise regression	adjusted R^2
17	before	IC_0, SIC_0	0.6886
17	after	IC_0, SIC_0	0.6793
16	before	$^3\kappa$, IC_0, SIC_0	0.8632
16	after	$^3\kappa$, SIC_0, MW	0.7925

Table II lists the loadings of the principal components, derived both with and without the outlier. In both cases, the first two principal components account for >90% of the variability. Table III lists the principal component scores for the 17 (or 16) observations. These scores form the **P** matrices referred to in the Methods section. Table IV describes the effects on the stepwise multiple regression runs of removing the outlier and of projecting the data points into the 4-space orthogonal to the first principal component. Table V shows the effects on the data points in the original parameter space ($^3\kappa$, IC_0, SIC_0, IC_1, MW) of the projections into 4-space. Table VI contains the measured values for the biological activity parameter from ref 7, along with the values predicted by the most successful attempts to model this parameter from the graph-theoretical predictors.

DISCUSSION

Projection. A second principal component axis is a statement about variability in the data set after projection into the (n − 1)-space orthogonal to the first principal component axis. Since a good correlation exists between the second principal component score (PC2) and log (1/MIC),

Table V. Parameters from Ref 1 and after Projection[a]

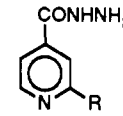
$^3\kappa$	IC_0	SIC_0	IC_1	MW
		Original Data from Ref 1		
1.89	1.74	0.426	3.22	137.14
2.10	1.68	0.388	3.21	152.17
2.39	1.62	0.358	3.19	166.20
2.93	1.57	0.334	3.01	180.23
3.57	1.53	0.315	3.00	194.25
2.36	1.78	0.404	3.33	169.18
2.90	1.72	0.375	3.30	183.21
2.08	1.75	0.411	3.28	153.16
3.34	1.78	0.389	3.36	192.50
3.94	1.74	0.365	3.35	209.23
3.09	1.56	0.315	3.06	209.27
2.06	1.98	0.485	3.41	155.13
2.27	1.98	0.485	3.41	171.58
2.29	1.94	0.458	3.54	182.14
2.52	1.58	0.332	2.89	214.24
3.21	1.54	0.314	2.92	228.27
2.24	1.66	0.379	3.20	165.08
		From PCA Including the Outlier		
2.2617	1.6243	0.3789	3.0829	153.90
2.2586	1.6306	0.3679	3.1515	159.32
2.3390	1.6359	0.3645	3.2088	163.90
2.5669	1.6831	0.3800	3.1439	163.86
2.9913	1.7102	0.3883	3.2134	168.16
2.5931	1.7074	0.3745	3.2440	179.69
2.8850	1.7247	0.3769	3.3055	182.53
2.3757	1.6579	0.3735	3.1709	166.49
3.3357	1.7813	0.3895	3.3616	192.31
3.7055	1.8130	0.3947	3.4365	198.66
2.5793	1.7190	0.3797	3.2484	186.25
2.7467	1.7662	0.3980	3.1567	186.09
2.8675	1.7939	0.4093	3.1896	198.52
2.8315	1.7714	0.3894	3.3403	206.55
2.0290	1.7329	0.3942	3.0711	192.10
2.5202	1.7548	0.4014	3.1744	197.17
2.2930	1.6435	0.3723	3.1804	167.47
		From PCA Excluding the Outlier		
2.1084	1.5071	0.3534	2.8566	141.35
2.1952	1.5785	0.3564	3.0516	154.01
2.3632	1.6486	0.3669	3.2346	165.68
2.7186	1.7955	0.4043	3.3618	176.15
3.2305	1.8921	0.4279	3.5650	187.70
2.4968	1.6341	0.3585	3.1024	171.82
2.8907	1.7299	0.3781	3.3155	183.03
2.2542	1.5642	0.3531	2.9901	156.52
3.3338	1.7866	0.3910	3.3702	192.38
2.7935	1.8762	0.4136	3.5534	203.55
2.4583	1.5552	0.3526	2.7472	162.82
2.6155	1.6115	0.3702	2.8351	178.25
2.6051	1.6039	0.3532	3.0156	188.22
2.2366	1.8822	0.4262	3.3615	208.77
2.8082	1.9685	0.4476	3.5885	220.52
2.2736	1.6242	0.3678	3.1441	165.73

[a] Projections are in the 4-space orthogonal to the first principal component axis as derived both including and excluding the multivariate outlier.

and since PC2 is nothing but a linear combination of the original parameters, one might anticipate also finding a good correlation between those parameters and log (1/MIC) after projection into the same $(n-1)$-space. To pursue this exercise is to investigate whether variability within the full parameter space does not predict biological activity, but variability within the defined subspace does. The correlation between PC2 and log (1/MIC) seems to indicate such a situation exists here.

Values for the original parameters after projection were obtained as described in the Methods section. Note the relationship among Tables II, III, and V. For a given observation, a parameter value is smaller after projection if the product of the first principal component loading for that parameter and the first principal component score for that observation is positive. If the product is negative, projection makes the parameter value larger. To our considerable surprise, regression results for the original parameters in the $(n-1)$-space were about the same as in the full n-space, or even worse.

Table VI. Measured and Predicted Biological Activities



obs	R	log (1/MIC) from ref 7	eq 2	eq 3	eq 4	eq 6
1	H	−0.041	−0.256	−0.329	−0.001	−0.288
2	CH_3	−0.716	−1.071	−0.923	−0.911	−0.972
3	C_2H_5	−1.324	−1.141	−1.279	−1.360	−1.307
4	n-C_3H_7	−1.742	−1.601	−1.772	−1.618	−1.889
5	i-C_4H_9	−2.653	−2.436	−2.322	−2.265	−2.347
6	CH_3O	−2.185	−2.296	−2.189	−2.160	−2.290
7	C_2H_5O	−2.655	−2.762	−2.704	−2.541	−2.772
8	NH_2	−1.161	−1.242	−1.208	−1.451	−1.237
9	CH_3CONH	−3.332	−2.828	−3.521	−3.097	−3.565
10	$CH_3CONHCH_2$	−2.386	−2.909			
11	$(C_2H_5)_2N$	−2.856	−3.069	−2.444	−2.285	−2.440
12	F	−2.415	−2.197	−2.239	−2.367	−2.36
13	Cl	−2.593	−2.568	−2.404	−2.718	−2.317
14	NO_2	−2.569	−2.595	−2.768	−2.934	−2.636
15	C_6H_5	−1.699	−1.809	−1.691	−1.629	−1.648
16	$C_6H_5CH_2$	−1.585	−1.421	−2.243	−2.376	−2.053
17	CH_2=CH	−1.544	−1.255	−1.038	−1.356	−0.976

Figures 1 and 2 illustrate our attempts to portray graphically the results of projecting the data points into the 4-space projection. Figure 1 shows one effect of the projection. After projection, the points appear roughly coplanar in the space defined by $^3\kappa$, IC_0, and MW. One might interpret this figure to mean that the projection into the 4-space approximates a projection into a 2-space and displaying the graph of those three underlying variable axes captures that fact. Figure 2 reinforces that interpretation, showing much less apparent change in the space defined by IC_0, SIC_0, and IC_1. Inspection of Table II, however, reveals that the axis variables in Figure 1 are the least important contributors to the first principal component, while those in Figure 2 are the most important, exactly the opposite of what one would expect if the initial interpretation of the figures were correct. In fact, Table II is correct, and the interpretation of the figures is not. Thus, in this exercise, as in other operations in higher-dimensional spaces, one must be extremely cautious in relying on graphical visualization aids of this type, which may aid in visualizing an erroneous interpretation. In this instance, it was much easier to see near-coplanarity of points than to interpret whether the plane in which the points nearly lie was relevant to the problem under study.

Regression. The highest adjusted R^2 and lowest standard error of the estimate (see Table I) were obtained from stepwise multiple regression on the full set of 15 independent variables and 17 observations examined by Ośmiałowski and Kaliszan. Using all 15 variables, multivariate outlier tests did not identify observation 10 as anomalous, so there was no objective justification for excluding it. Stepwise multiple regression produced a "best" equation with four variables:

$$\log (1/\mathrm{MIC}) = -1.029(0.378)^1\kappa - 9.757(2.701)\mathrm{IC}_0 - 3.144(1.361)\mathrm{CIC}_0 + 0.0148(0.004)W + 29.372$$

$$n = 17,\ s = 0.2953,\ R^2 = 0.9095,\ F = 30 \qquad (2)$$

The facts that this subset has only one variable in common

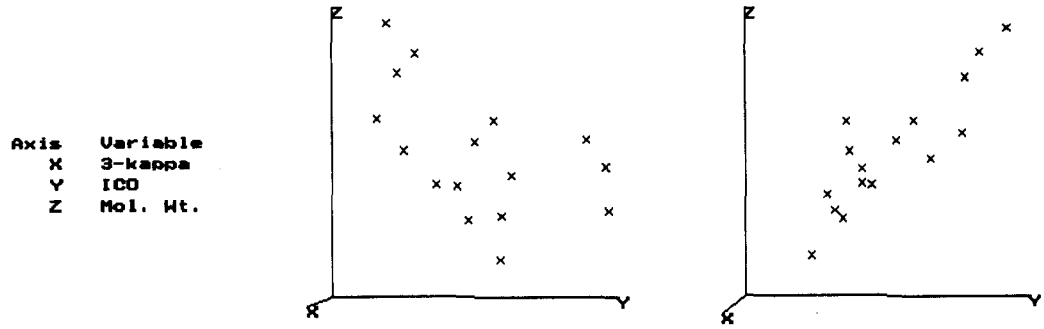


Figure 1. View of data points in the space defined by $^3\kappa$, IC_0, and MW before (left) and after (right) projection into the 4-space orthogonal to the first principal component axis.

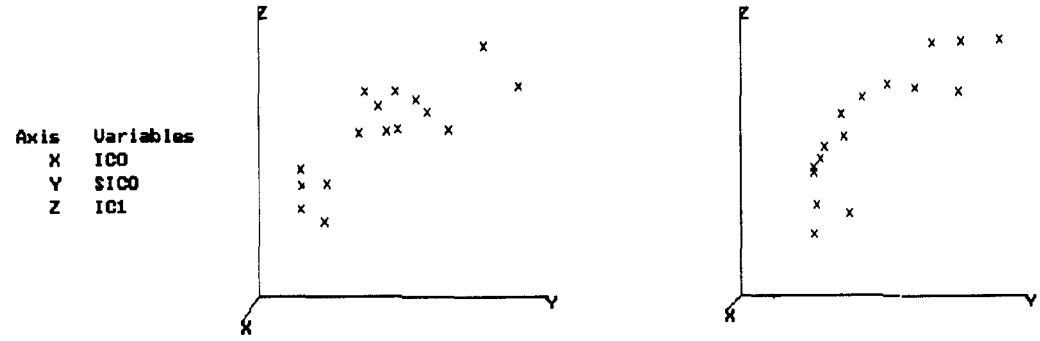


Figure 2. View of data points in the space defined by IC_0, SIC_0, and IC_1, before (left) and after (right) projection into the 4-space orthogonal to the first principal component axis.

with the subset we chose by inspecting the correlation table ($^3\kappa$, IC_0, SIC_0, IC_1, MW) and that both sets gave predictions of roughly equal quality indicate that there must be a great deal of duplication of information within the full set of 15 variables. Indeed, two-thirds of the numbers in Ośmiałowski and Kaliszan's correlation table exceed 0.8.

In studies such as the present one, in which the ratio of observations to independent variables is uncomfortably close to one, it is highly desirable to find some objective means to eliminate variables. This desire, however must be balanced against the need not to dispose of valuable information. Stepwise multiple regression is supposed to weigh these needs and select a "best" compromise. Another approach is to compute regressions on all possible subsets of variables. With 15 variables, however, the number of possible subsets exceeds 30 000.

All the other techniques that were nearly as good at predicting log (1/MIC) (adjusted $R^2 > 0.8$ and standard error of the estimate < 0.4) utilized the five-parameter subset, for which there was justification for excluding observation 10. Therefore, all of the following results are based on 16 observations. In comparing results of the different statistical techniques, particularly the different standard errors of the estimate, it would be well to remember that the claimed accuracy of the *measured* concentration (MIC) values is ±25%.[7] It is, therefore, questionable whether standard errors of the estimate of log (1/MIC) that differ from each other by less than 0.1 are significantly different at all. The next best predictions resulted from simply running the same stepwise multiple regression technique on the subset. That technique selected three variables and gave eq 3 as the best fit.

$$\log (1/\mathrm{MIC}) = -0.785(0.339)^3\kappa - 16.627(3.680)\mathrm{IC}_0 + 37.530(11.248)\mathrm{SIC}_0 + 14.097$$

$$n = 16,\ s = 0.3222,\ R^2 = 0.8905,\ F = 33 \qquad (3)$$

Equation 3 at least has one fewer parameter than eq 2, and, having started from only five parameters, it took much less time to compute.

We performed PCA on the data in the 5-space defined by the five-parameter subset. The correlation between log (1/MIC) and the first principal component was poor, but was much better for the second principal component. As mentioned above in the Methods section, the five-parameter subset was "best" for looking at the principal component correlation. In looking at all possible one-parameter additions and deletions from the subset, we noted that for all such operations the best correlation by far was between log (1/MIC) and the second principal component. While admitting that the precise physical meaning of this result eludes us, we speculate that the second principal component captures some aspect of molecular shape and size important in biological activity, perhaps in binding to an enzyme. It would be very significant indeed if the same sort of relationship were seen for other series of compounds. Failure to find the relationship, on the other hand, might simply reflect differences in enzyme systems.

Although the set of second principal component scores passed the Kolmogorov–Smirnov test for normality, an improved value for that test was obtained for the square root transform of those scores. The transform consists of finding the minimum value in the range, subtracting that value from each score to render them all ≥ 0, and then replacing each one by its square root. The best result in terms of correlation with log (1/MIC) was obtained by computing the principal component scores from all 17 observations and then dropping the outlier. The result is eq 4. The equation relating PC2a

$$\log (1/\mathrm{MIC}) = -1.798(0.187)\sqrt{\mathrm{PC2a}} + 0.0044$$

$$n = 16,\ s = 0.3275,\ R^2 = 0.8680,\ F = 92 \qquad (4)$$

to the original parameters is eq 5. The coefficients of the parameters are the second column of the inverse of the T

$$PC2a = 1.0349^3\kappa + 2.5304IC_0 + 2.3879SIC_0, + 2.3534IC_1 + 0.0205MW - 19.2702 \quad (5)$$

matrix ($\mathbf{T}^{-1}$), as $\mathbf{T}$ is defined above in the Methods section. If $\mathbf{y}$ is any row of the $\mathbf{Y}$ matrix (the rows are all the same), the constant is the second element of the vector defined by $\mathbf{y}\mathbf{T}^{-1}$.

Finally, reasonably good correlation with log (1/MIC) was obtained by running multiple regression (not stepwise) on the five parameters. The result was eq 6.

$$\log (1/MIC) = -0.768(0.355)^3\kappa - 22.795(7.941)IC_0 + 53.845(20.732)SIC_0 + 0.939(1.685)IC_1 + 0.011(0.011)MW + 13.355$$

$$n = 16, s = 0.3365, R^2 = 0.8507, F = 18 \quad (6)$$

The best relationship between log (1/MIC) and PC2a (not the square root transformation) is

$$\log (1/MIC) = 0.886(0.123)PC2a - 2.067$$

$$n = 16, s = 0.4141, R^2 = 0.7890, F = 52 \quad (7)$$

Substituting eq 5 into eq 7 gives

$$\log (1/MIC) = -0.917^3\kappa - 2.242IC_0 - 2.116SIC_0 - 2.085IC_1 - 0.018MW + 15.009 \quad (8)$$

We find it curious that eq 8 bears no apparent relationship at all to eq 6, even though their standard errors of the estimate are not wildly different.

The above represents many attempts by various techniques to demonstrate correlation between a suite of graph-theoretical parameters and a measure of biological activity.

The SCOUT multivariate statistics package was useful in exploring a variety of principal component analyses and manipulations of the principal component scores. Indeed, PCA produced a correlation from a subset of five out of 15 original parameters that was nearly indistinguishable in quality from the best obtained from the full parameter set by more conventional statistical techniques. One of the salient features of SCOUT is that its ease of use allows this sort of exploration without major investment of resources. We hope the examples presented in this paper will encourage more widespread use of multivariate techniques to seek out relationships among seemingly disparate types of data.

REFERENCES AND NOTES

(1) Ośmiałowski, K.; Kaliszan, R. Studies of Performance of Graph Theoretical Indices in QSAR Analysis. *Quant. Struct.-Act. Relat.* **1991**, *10*, 125–134.

(2) (a) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure-Activity Analysis*; Research Studies Press: Letchworth, 1986. (b) Kier, L. B.; Hall, L. H. Molecular Connectivity. 7. Specific Treatment of Heteroatoms. *J. Pharm. Sci.* **1976**, *65*, 1806–1809.

(3) Kier, L. B. Shape Indexes of Orders One and Three from Molecular Graphs. *Quant. Struct.-Act. Relat.* **1986**, *5*, 1–7.

(4) Sarkar, R.; Roy, A. B.; Sarkar, P. K. Topological Information Content of Genetic Molecules. 1. *Math. Biosci.* **1978**, *39*, 299–312.

(5) Basak, S. C.; Monsrud, L. J.; Rosen, M. E.; Frane, C. M.; Magnuson, V. R. A comparative study of lipophilicity and topological indices in biological correlation. *Acta Pharm. Jugosl.* **1986**, *36*, 81–95.

(6) Wiener, H. Relation of the Physical Properties of the Isomeric Alkanes to Molecular Structure: Surface Tension, Specific Dispersion, and Critical Solution Temperature in Aniline. *J. Phys. Colloid Chem.* **1958**, *52*, 1082–1089.

(7) Seydel, J. K.; Schaper, K.-J.; Wempe, E.; Cordes, H. P. Mode of Action and Quantitative Structure–Activity Correlations of Tuberculostatic Drugs of the Isonicotinic Acid Hydrazide Type. *J. Med. Chem.* **1976**, *19*, 483–492.

(8) Cash, G. G.; Breen, J. J. Principal Component Analysis and Spatial Correlation: Environmental Analytical Software Tools. *Chemosphere* **1992**, *24*, 1607–1623.

(9) Neter, J.; Wasserman, W. *Applied Linear Statistical Models*; Richard C. Irwin, Inc.: Homewood, IL, 1974; p 229.