Optimization of a Regression Model for a Quantitative Structure. Mutagenicity Relationship of Some Natural Amino Acids

Minati Kuanar and Bijay K. Mishra*

Center of Studies in Surface Science and Technology, Department of Chemistry, Sambalpur University, Jyoti Vihar-768019, India

(Received June 2, 1997)

Multivariate statistical tools, such as principal-component analysis and multiple-regression analysis, were used for a large number of structural and empirical parameters of some eleven natural amino acids for predicting mutagenicity. An optimized regression model was derived from some selective principal components and solubility data. Plots related to the principal components resulted in the ordination of amino acids as well as the structural and empirical parameters.

In an epidemiological study of the incidence of cancer among migrants in Hawaii from Japan, Matsushima and Sugimura[1] have shown that environmental factors, especially factors present in food, play an important role in causing cancer. They have detected a high mutagenic potential in smoke from broiled fish.[2] Naggo et al.[3] have also found high mutagenic activity due to the charred surface of sardine and beefsteak. From the mutagenicity of smoke condensates obtained by a pyrolysis study of various biological materials, lysozymes and histones have been found to lead the list with maximum mutagenicity. Compared to DNA and RNA, the two proteins have 20—100 times more mutagen activity.

A quantitative description of the variation of protein and biological activities with the structure has been extensively studied.[4] Mc Cann et al.[5] and Sugimura et al.[6] have found a close relationship between mutagenicity and carcinogenicity. To screen possible carcinogenicity in our daily food, a mutagenicity test using *Salmonella typhimurium* is well established.[7]

The mutagenicity of tar obtained from the pyrolysis of various amino acids on *Salmonella typhimurium* TA 98 has been reported by Naggo et al.[8] Possible condensates for the mutagenic activity have also been proposed by isolating various nitrogen-containing heterocycles from the pyrolusates.[9—15] Considering amino acids as being progeneric materials for cancers, we have made an attempt to correlate the structure of amino acids with the mutagenicity of tars obtained from the pyrolysis of amino acids.

Graph theoretical molecular descriptors have already been used in a quantitative structure-activity relationship to explain various biological and physicochemical activities.[16—23] Recently, Pogliani[24] used some of these descriptors to explain the physicochemical properties of natural amino acids. In the present study we determined some topological indices (TIs) of amino acids, and used those to optimize a regression model for predicting the mutagenicity of eleven natural amino acids. The solubility of amino acids in water has also been included as an independent variable in the regression model.

Data Bases

The mutagenic activities of the tars produced from the pyrolysis of amino acids in *S. typhimurium* TA 98 with S9 were taken from the literature.[1] Mutagenic activities are expressed as the mutation rate ($\ln(R)$) in log (revertants/mg). The solubilities of natural amino acids in water were collected from the literature.[25] The mutagenic activities and solubility data are given in Table 1.

Graph Theoretical Parameters. Three different sets of graph theoretical parameters were used for the study.

Molecular Connectivity Indices (D, D^v, ${}^0\chi$, ${}^0\chi^v$, ${}^1\chi$, and ${}^1\chi^v$): These indices can be computed as follows. The parameter (D) is obtained by using Eq. 1,[26]

$$D = \sum \delta_i, \tag{1}$$

where δ_i represents the count of non-hydrogen σ-bond electrons contributed by atom i,[27]

$$\delta_i = \sigma_i - H_i, \tag{2}$$

H_i being the number of hydrogen atoms attached to i. D^v is defined as[26]

$$D^v = \sum \delta_i^v, \tag{3}$$

where δ_i^v represents the count of all non-hydrogen valence electrons contributed by atom i,[27,28]

$$\delta_i^v = \delta_i + p_i + n_i. \tag{4}$$

Here, σ, p, and n are the sigma, p_i, and lone-pair electrons, respectively.

The zero-order connectivity index (${}^0\chi$,[26,27]) is obtained by using

$${}^0\chi = \sum (\delta_i)^{-1/2} \tag{5}$$

Table 1. Mutagenecity (*S. typhimurium* TA 98 pyrolysate products), Solubility (S) in Water ($g\,kg^{-1}$) of 11 Natural Amino Acids (AA), and Their Molecular Connectivity Indices

No.	AA	$\log R_{ev}$/nmol	S	D	D^v	$^0\chi$	$^0\chi^v$	$^1\chi$	$^1\chi^v$
1.	Ser	422	18400	12	28.0	5.862	3.664	3.181	1.774
2.	Arg	181	4950	22	42.0	9.560	6.709	5.537	3.600
3.	Thr	97	3100	14	30.0	6.732	4.535	3.553	2.219
4.	Ala	167	2980	10	22.0	5.155	3.510	2.643	1.627
5.	Gln	42	1600	18	38.0	8.146	5.410	4.537	2.804
6.	Met	56	890	16	26.7	7.276	6.146	4.181	4.044
7.	Tyr	0.5	199	26	48.0	9.845	6.974	6.092	3.857
8.	Phe	29	148	24	42.0	8.975	6.604	5.698	3.722
9.	His	43	104	22	42.0	8.268	5.819	5.198	3.155
10.	Asn	25	98	16	36.0	7.439	4.703	4.037	2.304
11.	Val	58	0.9	14	26.0	6.732	5.088	3.553	2.538

Similarly, $^0\chi^v$ is the zero-order valence connectivity index,[26,28]

$$^o\chi^v = \sum(\delta_i^v)^{-1/2}. \quad (6)$$

The first-order connectivity index[26] ($^1\chi$) and the first-order valence connectivity index[28] ($^1\chi^v$) are obtained by using

$$^1\chi = \sum(\delta_i\delta_j)^{-1/2}, \quad (7)$$

$$^1\chi^v = \sum(\delta_i^v\delta_j^v)^{-1/2}. \quad (8)$$

Information Theoretical Topological Indices (IC, SIC, CIC): To obtain these parameters, a total molecular graph (where hydrogen is not suppressed) is initially constructed for the molecule. For each vertex (atom), partition coordinates are assigned according to the bonding-connected atoms with their immediate neighborhood. The coordinates bear information concerning the types of bonds between the concerned atom and adjacent atoms. The coordinates are then classified according to their partition coordinates. From the probability of the class and the number of atoms in the molecular graph, IC is calculated using Shannon's formula,[29,30]

$$IC = -\sum_{i=1}^{k} P_i \log_2 P_i, \quad (9)$$

where $P_i=n_i/n$ is the probability that a randomly selected element will lie in the i^{th} partitioning class, n is the number of atoms in the molecular graph and k is the number of partitioning class. The structural information content (SIC) and complementary information content (CIC) are obtained by using[30]

$$SIC = IC/\log_2 n, \quad (10)$$

$$CIC = 1/n\sum^{k} n_i \log_2 n_i. \quad (11)$$

Wiener Indices: (W and $\overline{W}$) and Information Indices (I_D^W and $I_D^{\overline{W}}$): The Wiener number (W),[31] the first topological index reported in the chemical literature, is obtained by adding the entries in the upper triangular distance-submatrix of the hydrogen-suppressed chemical graph. The entries of the matrix, being $d_{ij}=d_{ji}$, represent the number of bonds between vertices i and j by the shortest path. From the distance matrix, W is computed as[31]

$$W = \sum d_{ij}/2, \quad (12)$$

The mean Wiener number ($\overline{W}$) is determined as

$$\overline{W} = 2W/n(n-1). \quad (13)$$

By a statistical treatment of the topological distances of the chemical graph, the information index (I_D^W) is obtained through the formalism of information theory,[32]

$$I_D^W = W\log_2 W - (k_d)(d\log_2 d), \quad (14)$$

where the distance (d) appears k_d times in the partition. The mean information index ($\bar{I}_D^W$) can be calculated as

$$\bar{I}_D^W = I_D^W/W. \quad (15)$$

KOKOS Parameters: ($^1\Gamma$, $^2\Gamma$, $^3\Gamma$, $^4\Gamma$, $^5\Gamma$, and $^6\Gamma$): The KOKOS parameters were obtained from the literature.[24]

Principal Component Analysis. The Principal component analysis (PCA) uses a linear combination of a set of n-variables to derive a new set of p-principal components, where $p \leq n$ and the principal components are orthogonal to each other. Initially, a matrix is constructed consisting of the correlations (covariances) among the variables of interest. The eigenvalues and eigenvectors of the matrix are then determined. The thus-obtained eigenvectors are orthogonal, and the sum of their eigenvalues equals the original number of variables. Each eigenvector is a linear combination of the original variables, and represents a principal component (PC). The process can be viewed as one in which the first principal component (PC_1) axis is constructed to account for a maximum amount of variance in the data; the second component (PC_2) axis accounts for the maximum amount of the remaining variance under the constraint that it be orthogonal to the first principal component, and so forth, until all component axes are constructed. Thus, the total number of principal components is the same as the number of concerned variables. The extraction of significant principal components to be used in a regression analysis from the total number can be made by various techniques. In the social sciences, there is a method for retaining PCs having eigenvalues of one percent, five percent or ten percent, or having magnitude 'one' or more than one.[33] Cattell[34] has advocated the use of a scree-

test in which the eigenvalues are plotted against the principal component number. The principal components, constituting the horizontal straight line at the lower part of the plot, are excluded. During the extraction technique the consideration of eigenvalues greater than unity has been considered to be too stringent for a large number of physical data.[23] In the present study, a scree plot and eigenvalues with more than one percent were considered for the extraction of significant PCs.

During the optimization process, insignificant parameters were removed from the model by considering the resulting F-test, t-test, and correlation coefficient (r) values. All of the calculations for PCA and a multiple-regression analysis were made by using SAS software package.

Results and Discussion

In the present study, thirteen basic topological parameters, six empirical parameters derived from various physical and chemical properties and solubility as a parameter (Tables 1, 2, and 3), were used to predict mutagenicity of eleven natural amino acids. In correlation studies of the biological activity, the solubility in water or the partition coefficient between a lipid and water of the candidate molecule plays an important role.[35] In a homologous series of compounds (where there is no change in the functional groups and there is a change only in the methylene units) only the topological parameters

Table 2. Information Content Parameters (IC, SIC, CIC), Wiener and Information Indices (W, $\overline{W}$, I_D^W, $\bar{I}_D^W$) of 11 Natural Amino Acids (AA)

No.	AA	IC	SIC	CIC	W	$\overline{W}$	I_D^W	$\bar{I}_D^W$
1.	Ser	3.039	0.798	0.768	46	2.190	195.6	4.251
2.	Arg	3.088	0.657	1.612	244	3.697	1505	6.149
3.	Thr	3.052	0.747	1.036	65	2.321	303.4	4.668
4.	Ala	2.931	0.792	0.769	29	1.933	109.9	3.788
5.	Gln	3.041	0.704	1.280	136	3.022	719.4	5.290
6.	Met	3.022	0.699	1.300	102	2.833	507.1	4.971
7.	Tyr	3.099	0.676	1.485	264	3.385	1688	6.395
8.	Phe	2.888	0.639	1.635	212	3.213	1238	5.838
9.	His	3.504	0.811	0.818	166	3.018	927	5.585
10.	Asn	3.101	0.759	0.986	92	2.556	455.7	4.953
11.	Val	2.774	0.653	1.474	65	2.321	303.4	4.668

Table 3. KOKOS ${}^{i}\Gamma$ Indices of 11 Natural Amino Acids (AA)

AA	${}^{1}\Gamma$	${}^{2}\Gamma$	${}^{3}\Gamma$	${}^{4}\Gamma$	${}^{5}\Gamma$	${}^{6}\Gamma$
Ser	−1.21	−1.19	−0.33	−0.46	−0.54	0.22
Arg	1.16	−0.57	−1.52	−1.07	−0.28	−0.13
Thr	−0.67	−0.97	0.10	−0.36	0.57	0.86
Ala	−1.44	−0.47	0.11	0.32	−0.51	−0.86
Gln	0.22	−1.24	−0.46	−1.05	0.19	−0.42
Met	0.44	0.20	0.72	1.00	0.45	0.24
Tyr	1.34	1.16	1.04	−0.07	1.02	1.21
Phe	1.09	1.6	1.24	1.16	0.88	0.48
His	0.52	−0.46	−0.18	−0.13	−0.56	−0.10
Asn	−0.34	−1.25	−0.06	−0.96	−1.00	−1.19
Val	−0.34	0.42	0.77	1.38	1.84	1.66

can be used to predict the biological activities or physicochemical properties of the molecules.[36—39] Recently, Basak and Grunwald[40] have reported on the use of topological parameters along with some physicochemical parameters for predicting the mutagenicity of some aromatic and heteroaromatic amines.

Topological and KOKOS Parameters. The topological parameters calculated for three distinct sets are tabulated in Tables 1 and 2. The connectivity parameters (D, D^{v}, ${}^{0}\chi$, ${}^{0}\chi^{v}$, ${}^{1}\chi$, and ${}^{1}\chi^{v}$) are derived by considering the connectivity of atoms with a maximum of one bond connection. The parameters (IC, SIC, and CIC) are derived by considering the probability of the partition coordinates assigned to each vertex of a nonhydrogen depleted molecular graph, while the distance parameters (W, $\overline{W}$, I_D^W and $\bar{I}_D^W$) are obtained from the distance matrix of the molecule. Cross-correlation matrices of the topological parameters of each set are given in Table 4. Regression models (Eqs. 16, 17, and 18) were derived by using all of the parameters of each set for predicting the mutagenicity.

$$\begin{aligned}\log(R_{ev}/\mathrm{mg}) = {} & 21805 \pm 23592 - (4884.3 \pm 11563)D \\ & - (2318.3 \pm 4374.2)D^{v} + (6280.0 \pm 36743)^{0}\chi \\ & - (26264 \pm 48509)^{0}\chi^{v} + (52320 \pm 47097)^{1}\chi \\ & + (4005.7 \pm 25508)^{1}\chi^{v} \\ r = 0.8100 \quad & F = 1.272 \end{aligned} \tag{16}$$

($t = 0.92, -0.42, -0.53, 0.17, -0.54, 1.11, 0.16$ respectively)

$$\begin{aligned}\log(R_{ev}/\mathrm{mg}) = {} & -(5557530 \pm 351880) - (52061 \pm 28184)\mathrm{IC} \\ & + (785800 \pm 465420)\mathrm{SIC} + (127330 \pm 80361)\mathrm{CIC} \\ r = 0.6638 \quad & F = 1.838 \end{aligned} \tag{17}$$

($t = -1.58, -1.85, 1.69, 1.58$ respectively)

$$\begin{aligned}\log(R_{ev}/\mathrm{mg}) = {} & (17766 \pm 52021) - (420.10 \pm 568.83)W \\ & + (12924 \pm 15867)\overline{W} + (58.214 \pm 67.732)I_D^W \\ & - (7476.4 \pm 13542)\bar{I}_D^W \\ r = 0.5433 \quad & F = 0.6345 \end{aligned} \tag{18}$$

($t = 0.34, -0.74, 0.81, 0.86, -0.55$ respevtively)

The connectivity parameters show a relatively good correlation (r=0.8100, F=1.272) with mutagenicity. While optimizing the regression model for predicting the solubility of nineteen natural amino acids, Pogliani[24] showed an improvement in the regression model by considering the reciprocal of the topological parameters. The reciprocal of the first set, in the present study, improved the correlation from 0.8100 to 0.9098. Similarly, the reciprocal of the parameters derived from the distance matrices improved from 0.5433 to 0.7125. However, there was no improvement in the case of the reciprocal of TIs of information theoretic indices.

The KOKOS parameters reported by Kidera et al.[41] are the result of several multivariate statistical analyses to 188 physical properties of amino acids. These parameters, in the present study, have an intercorrelationship of 0.03 to

Table 4. Cross Correlation Matrix for Each Set of Parameter

Set I. Connectivity Indices (D, D^v, ${}^0\chi$, ${}^0\chi^v$, ${}^1\chi$, and ${}^1\chi^v$)

	D	D^v	${}^0\chi$	${}^0\chi^v$	${}^1\chi$	${}^1\chi^v$
D	1.00					
D^v	0.95	1.00				
${}^0\chi$	0.97	0.93	1.00			
${}^0\chi^v$	0.91	0.77	0.93	1.00		
${}^1\chi$	0.99	0.94	0.98	0.93	1.00	
${}^1\chi^v$	0.82	0.63	0.82	0.96	0.84	1.00

Set II. Information Theoretic Parameters (IC, SIC, and CIC)

	IC	SIC	CIC
IC	1.00		
SIC	0.54	1.00	
CIC	−0.41	−0.99	1.00

Set III. Wiener and Information Indices (W, $\overline{W}$, I_D^W, $\bar{I}_D^W$)

	W	$\overline{W}$	I_D^W	$\bar{I}_D^W$
W	1.00			
$\overline{W}$	0.96	1.00		
I_D^W	0.99	0.94	1.00	
$\bar{I}_D^W$	0.98	0.96	0.97	1.00

Set IV. KOKOS Parameters (${}^1\Gamma$, ${}^2\Gamma$, ${}^3\Gamma$, ${}^4\Gamma$, ${}^5\Gamma$, ${}^6\Gamma$)

	${}^1\Gamma$	${}^2\Gamma$	${}^3\Gamma$	${}^4\Gamma$	${}^5\Gamma$	${}^6\Gamma$
${}^1\Gamma$	1.00					
${}^2\Gamma$	0.59	1.00				
${}^3\Gamma$	0.13	0.76	1.00			
${}^4\Gamma$	0.03	0.74	0.79	1.00		
${}^5\Gamma$	0.31	0.68	0.64	0.66	1.00	
${}^6\Gamma$	0.25	0.58	0.51	0.54	0.88	1.00

0.88 (Table 4). For all of these six KOKOS parameters (Table 3), when subjected to a multiple-regression analysis for predicting mutagenicity, the correlation coefficient was found to be 0.7908 with an F value of 1.112 (Eq. 19). When the reciprocal of these parameters was used as independent variables, 'r' was found to be 0.7957 with an F value of 1.151.

$$\log(R_{ev}/\mathrm{mg}) = (45061 \pm 2445.0) - (4098.0 \pm 3440.0)^1\Gamma + (3462.1 \pm 5249.6)^2\Gamma - (3193.9 \pm 4126.3)^3\Gamma - (1600.1 \pm 4223.5)^4\Gamma - (6169.8 \pm 4945.5)^5\Gamma + (6365.9 \pm 4220.8)^6\Gamma$$
$$r = 0.7908 \quad F = 1.112 \qquad (19)$$

(t = 1.84, −1.19, 0.66, −0.77, −0.38, −1.25, 1.51 respectively)

Principal-Component Analysis of the TIs and KOKOS Parameters. To reduce the dimentionality of the parameters and to generate orthogonal parameters in each set, PCA was undertaken. From the scree-plot criterion and eigenvalues with more than one percent, the first six PCs were found to be significant. All six principal components were evaluated in each set. Almost all TIs show a good correlation with either the first or second PC of each set. We have made an attempt to ordinate the TIs on the basis of their contribution to the PCs. All nineteen parameters were subjected to PCA; it was found that the first nine PCs can explain all of the variances (cumulative variance being 100% Table 5). However, the first six PCs (which have a cumulative variance of 99.3%) were correlated with all nineteen parameters; the correlation coefficients are given in Table 6. When the correlation coefficients of these parameters with PC_1 and PC_2 are plotted, the points are distributed in various quadrants of the plot (Fig. 1). Some of the important features of this plot are:

(i) The parameters D, D^v, ${}^0\chi$, and ${}^1\Gamma$ are found to be in one quadrant, and can be accommodated in one group.

Table 5. Eigenvalues, Percent of Variance, and Cumulative Percent of Variance Derived from PCA

PCs	Eigenvalues	(%) of variance	Cumulative variance
1.	11.820	62.2	62.20
2.	4.4700	23.5	85.70
3.	1.2450	6.60	92.30
4.	0.6825	3.60	95.90
5.	0.4528	2.40	98.30
6.	0.2044	1.10	99.30
7.	0.0789	0.40	99.70
8.	0.0334	0.20	99.90
9.	0.0147	0.10	100.0
10.	0.0008	0.00	100.0
11.	0.0000	0.00	100.0
12.	0.0000	0.00	100.0
13.	0.0000	0.00	100.0
15.	−0.0000	0.00	100.0
16.	0.0000	0.00	100.0
17.	−0.0000	0.00	100.0
18.	−0.0000	0.00	100.0
19.	−0.0000	0.00	100.0

(ii) The parameters ${}^0\chi^v$ and ${}^1\chi^v$ are found to be in the same quadrant at close proximity.

(iii) The IC, SIC, and CIC parameters have a wide distribution on the plot, indicating the least correlation with each other.

(iv) The KOKOS parameters, except for ${}^1\Gamma$, are found in a group in the same quadrant.

Thus, parameters D, D^v, ${}^0\chi$, ${}^0\chi^v$, ${}^1\chi^v$, W, $\overline{W}$, I_D^W, $\bar{I}_D^W$, and ${}^1\Gamma$ form one group of parameters; information theoretic parameters form a different group. However, the rest of the KOKOS parameters form another group.

When the first principal component (PC_1) and the second principal component (PC_2) of the eleven natural amino acids are plotted, all of the amino acids are found to be dispersed in the four different quadrants (Fig. 2 and Table 7).

Table 6. Correlation Coefficients of TIs and KOKOS Parameters with First Six PCs

Parameters	Principal components					
	PC_1	PC_2	PC_3	PC_4	PC_5	PC_6
D	−0.9607	0.1727	0.1740	−0.0277	0.1076	−0.0031
D^v	−0.855	0.4029	0.1093	−0.1253	0.2613	0.0605
$^0\chi$	−0.9696	0.2152	−0.0565	−0.0432	0.0495	0.0655
$^0\chi^v$	−0.9781	−0.0311	0.0158	0.0723	0.1819	0.0472
$^1\chi$	−0.9723	0.1910	0.1064	−0.0027	0.0568	0.0034
$^1\chi^v$	−0.9005	−0.0958	0.0970	0.1902	−0.3316	0.0908
IC	−0.1671	0.7725	0.5176	−0.1926	−0.2500	0.0259
SIC	0.7019	0.5159	0.4738	−0.0842	−0.0026	−0.0534
CIC	−0.7999	−0.4085	−0.4293	0.0706	−0.0211	0.0221
W	−0.9630	0.2101	0.0003	0.0026	0.0993	−0.1223
$\overline{W}$	−0.9406	0.2817	−0.1304	0.0577	−0.1063	−0.0251
I_D^W	−0.9525	0.1998	−0.0039	−0.0164	0.1274	−0.1593
$\bar{I}_D^W$	−0.9734	0.2076	−0.0048	−0.0768	0.0256	0.0250
$^1\Gamma$	−0.9876	0.0643	0.0179	0.0701	−0.1121	0.0484
$^2\Gamma$	−0.6420	−0.6395	0.3013	0.1758	0.1361	−0.1930
$^3\Gamma$	−0.1733	−0.7585	0.5126	0.1145	0.2270	0.2494
$^4\Gamma$	−0.0548	−0.8724	0.3460	0.2129	−0.1708	−0.1483
$^5\Gamma$	−0.3842	−0.8446	−0.0859	−0.3167	−0.0344	0.0944
$^6\Gamma$	−0.3380	−0.6947	0.0658	−0.6079	−0.999	−0.0874

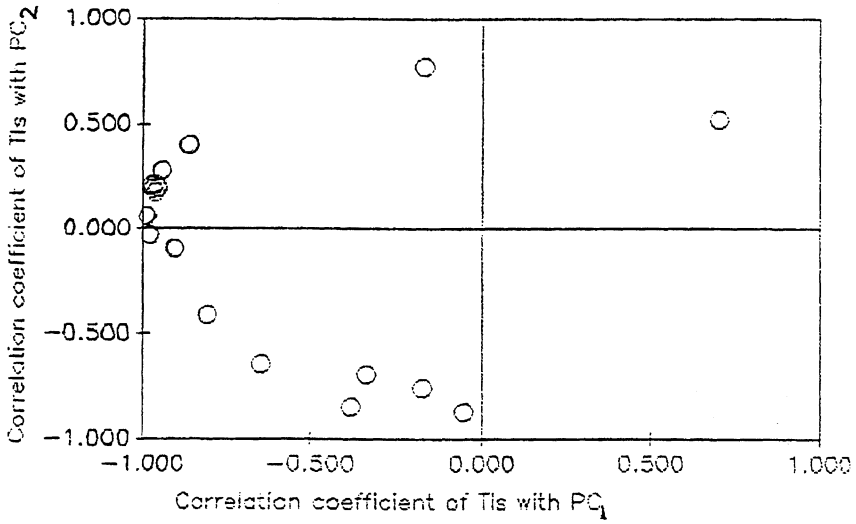

Fig. 1. Distributions of TIs in various quadrants of the plot of r values of TIs with PC_1 and PC_2.

Interestingly, the structurally related amino acids considered in this work are found to be present in the same quadrant. The two aromatic amino acids (phenylalanine and tyrosine) are found in the left-bottom quadrant, the two aliphatic amino acids (alanine and valine) are found in the right-bottom quadrant and the two basic amino acids (arginine and histidine) are found in the left-top quadrant. Thus, the structural characteristics seem to assume a quantitative coding through the principal components.

Optimization of Regression Model. The solubility (S; g kg^{-1} water), when considered as a parameter, is found to correlate well with the mutagenicity value with a correlation coefficient of 0.97 (F=133.6). Further, it is found that each set of parameters collectively have some contribution towards mutagenicity. To determine the loading of each set of parameters towards mutagenicity, the solubility along with the PC_1 and/or PC_2 of each set have been correlated with the mutagenicity. However, by adding the PCs of each set to the solubility parameter, no significant improvement in the regression model is observed. Hence, the PCs derived by considering all of the nineteen parameters were used as independent variables in the regression model. The number of variables was reduced by discarding any PCs having 't' values less than one in the regression model. Finally a regression model (20) was obtained with a correlation coefficient of 0.9924 and an F value of 151.3.

$$\log(R_{ev}/\text{mg}) = (-2532.8 \pm 397.83) + (53.842 \pm 3.1235)\text{S} - (295.28 \pm 92.382)\text{PC}_1 + (3146.5 \pm 713.10)\text{PC}_6$$
$$r = 0.9924 \quad F = 151.3 \qquad (20)$$
(t = −6.37, 17.24, −3.20, 4.41 respectively)

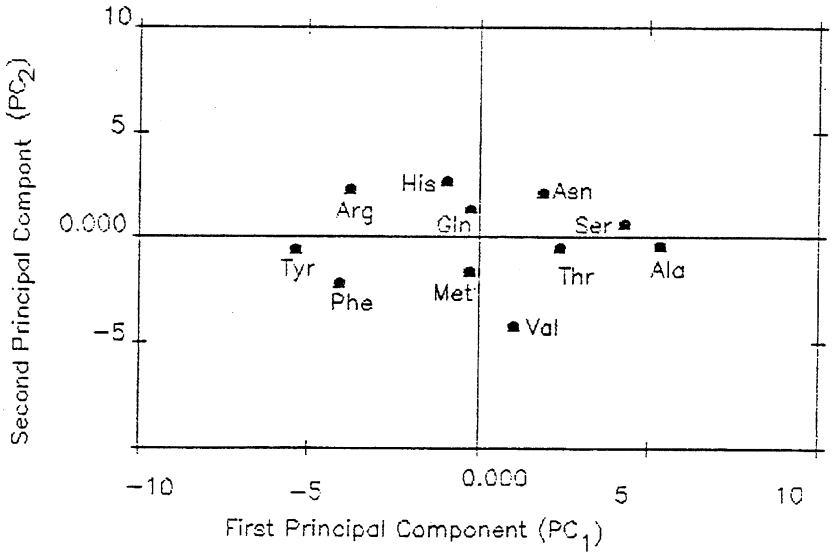

Fig. 2. Distribution of amino acids in a two dimensional space of PC_1 and PC_2.

Table 7. First and Second Principal Components of the 11 Natural Amino Acids

Amino acids	Principal components	
(AA)	PC_1	PC_2
Ser	4.280	0.694
Arg	−3.812	2.322
Thr	2.371	−0.435
Ala	5.326	−0.340
Gln	−0.282	1.371
Met	−0.301	−1.610
Tyr	−5.367	−0.533
Phe	−4.110	−2.145
His	−0.986	2.727
Asn	1.855	2.107
Val	1.027	−4.159

The mutagenicities of the eleven natural amino acids were calculated using Eq. 20 and plotted against the observed values (Fig. 3).

Graph theoretical indices were generated from the molecular graph, and, hence, lack a three-dimensional structural contribution to the molecular descriptor. Further, the PCs are constructed mathematically, and, hence, necessarily, no physical significance can be attributed to them.[42] However, each PC has some loading from each topological parameter, and, thus, in optimizing the regression model the PCs are found to have an important role to play.

One of the authors (MK) thanks CSIR, New Delhi, for awarding a Senior Research Fellowship.

Appendix

Illustration of Information Content Parameters (IC, SIC and CIC): The calculation of Information content parameters (IC, SIC, and CIC) of the amino acid, serine (Chart 1) is illustrated below. The hydrogen depleted molecular graph of serine (Chart 2) and the total molecular graph of serine and the partition co-ordinate values assigned to the bonding connected atoms with its immediate neighbourhood are given in Chart 3 and Table A1.

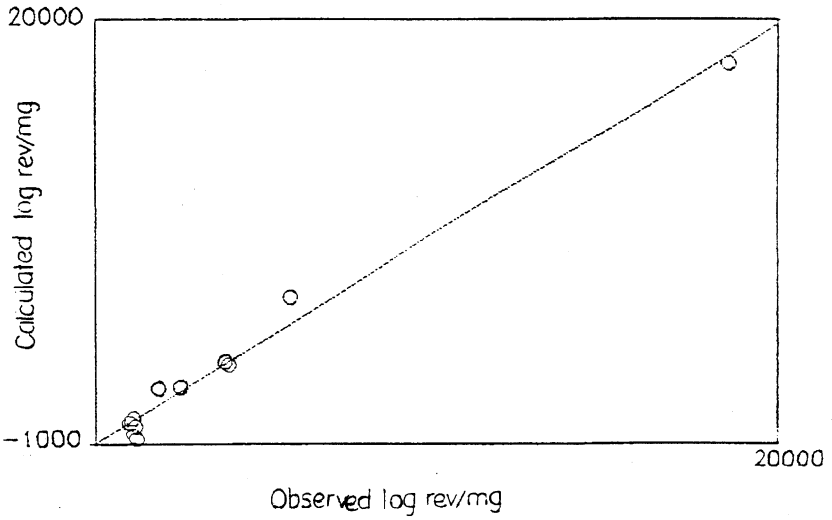

Fig. 3. Plot of observed and calculated mutagenicity of eleven natural amino acids using Eq. 20.

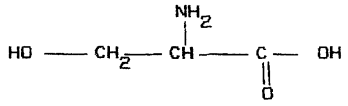

Chart 1.

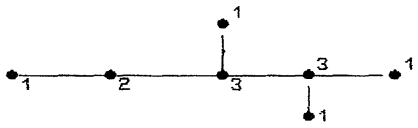

Chart 2.

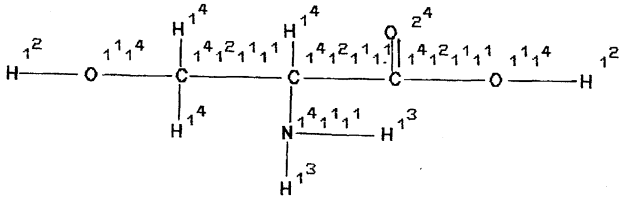

Chart 3. Coordinate attached structure of serine.

Table A1. Partitioning of Serine

Partition class and coordinate		Number of atoms [a] in partitioned class	Probablity ($P_i=n_i/n$)
I	1^2	2	2/14
II	1^3	2	2/14
III	1^4	3	3/14
IV	2^4	1	1/14
V	$1^1 1^4$	2	2/14
VI	$1^4 2^2 1^2$	1	1/14
VII	$1^4 1^1 1^1$	1	1/14
VIII	$1^4 1^4 1^3 1^1$	1	1/14
IX	$1^4 1^2 1^1 1^1$	1	1/14

a) Total number of atoms in the molecule is 14. The IC, SIC and CIC parameters are obtained as $IC = -\sum^k P_i \log_2 P_i = 3\times 2/14 \log_2 (14/2) + 3/14 \log_2 (14/3) + 5/14 \log_2 (14) = 3.039$, $SIC = IC/\log_2 n = IC/\log_2 14 = 0.798$, $CIC = 1/n \sum^k n_i \log_2 n_i = 3\times 2/14 \log_2 (2) + 3/14 \log_2 (3) + 5\times 1/14 \log_2 (1) = 0.768$.

References

1) T. Matsushima and T. Sugimura, "Progress in Mutation Research," ed by A. Kappas, Elsevier Biomedical Press, North-Holland (1981), Vol. 2, p. 49.

2) T. Sugimura, M. Naggo, T. Kawachi, M. Honda, T. Yahagi, Y. S. S. Sato, N. Matsukura, T. Matsushima, A. Shirai, M. Sawamura, and H. Matsumoto, "Origins of Human Cancer," ed by H. H. Hiatt, Cold Spring, Harbor (1977), p. 1561.

3) M. Naggo, M. Honda, Y. Seino, T. Yahagi, T. Kawachi, and T. Sugimura, *Cancer Lett.*, **2**, 335 (1977).

4) M. Charton, *J. Chim. Phys. Phys.-Chim. Biol.*, **89**, 1689 (1992).

5) J. Mc Cann, N. E. Spingarn, J. Kobori, and B. N. Ames, *Proc. Natl. Acad. Sci. U.S.A.*, **72**, 979 (1975).

6) T. Sugimura, S. Sato, M. Naggo, T. Yahagi, T. Matsushima, Y. Seino, M. Takeuchi, and T. Kawachi, "Fundamentals in Cancer Prevention," ed by P. N. Magee, Jpn. Sci. Soc. Press, Baltimore (1976), p. 191.

7) T. Matsushima, T. Sugimura, M. Naggo, T. Yahagi, A. Shiraj, and M. Sawamura, "Factor Modulating Mutagenicity in Microbia Tests," in "Short Term Test System for Detecting Carcinogens," ed by K. H. Horpoth, Springer, Berlin (1980), p. 273.

8) M. Naggo, T. Yahagi, T. Kawachi, Y. Seino, M. Honda, N. Matsukura, T. Sugimura, K. Wakabayashi, K. Tsuji, and T. Kosuge, "Mutagens in Foods, and Aspecially Pyrolysis Products of Protein," in "Progress in Genetic Toxicology," ed by D. Scott, Elsevier, Amsterdam (1977), p. 259.

9) T. Sugimura, T. Kawachi, M. Naggo, T. Yahagi, Y. Seino, T. Okamoto, K. Shudo, T. Kosuge, K. Tsuji, K. Wakabayashi, Y. Iitaka, and A. Itai, *Proc. Jpn. Acad.*, **53**, 58 (1977).

10) T. Yamanoto, K. Tsuji, T. Kosuge, T. Okamoto, K. Shudo, K. Takeda, Y. Iitake, K. Yamaguchi, Y. Seino, T. Yahagi, M. Naggo, and T. Sugimura, *Proc. Jpn. Acad., Ser. B*, **54B**, 248 (1978).

11) K. Wakabayashi, K. Tsuji, T. Kosuge, K. Takeda, K. Yamaguchi, T. Iitaka, T. Okamoto, T. Yahagi, M. Naggo, and T. Sugimura, *Proc. Jpn. Acad., Ser. B*, **54B**, 569 (1978).

12) D. Yoshida, T. Matsumoto, R. Yoshimura, and T. Matsuzaki, *Biochem. Biophys. Res. Commun.*, **83**, 915 (1978).

13) H. Kasai, S. Nishimura, K. Wakabayashi, M. Naggo, and T. Sugimura, *Proc. Jpn. Acad., Ser. B*, **56B**, 382 (1980).

14) H. Kasai, Z. Yamaizumi, K. Wakabayashi, M. Naggo, T. Sugimura, S. Yokoyama, T. Miyazawa, N. E. Spingarn, J. H. Weisburger, and S. Nishimura, *Proc. Jpn. Acad.*, **56**, 278 (1980).

15) H. Kasai, S. Nishimura, M. Naggo, Y. Takahashi, and T. Sugimura, *Cancer Lett.*, **7**, 343 (1979).

16) M. Kuanar and B. K. Mishra, *J. Serb. Chem. Soc.*, **62**, 289 (1997).

17) L. Pogliani, *J. Phys. Chem.*, **100**, 18065 (1996).

18) L. Pogliani, *J. Chem. Inf. Comput. Sci.*, **36**, 1082 (1996).

19) B. Lucic, S. Nikolic, and N. A. Trinajstic, *Croat. Chem. Acta*, **68**, 435 (1995).

20) S. C. Basak and G. D. Grunwald, *SAR QSAR Env. Res.*, **3**, 289 (1995).

21) S. C. Basak, S. Bertelsen, and G. D. Grunwald, *Toxicol. Lett.*, **79**, 239 (1995).

22) P. Khadikar, S. Karmarkar, S. Joshi, and I. Gutman, *J. Serb. Chem. Soc.*, **61**, 89 (1996).

23) D. E. Needham, I. C. Wei, and P. G. Seybold, *J. Am. Chem. Soc.*, **110**, 4186 (1988).

24) L. Pogliani, *Croat. Chem. Acta*, **69**, 95 (1995).

25) "Hand Book of Chemistry and Physics," ed by D. R. Lide, CRC Press, Boca Raton, Florida (1992), p. 7.1.

26) L. Pogliani, *J. Phys. Chem.*, **97**, 6731 (1993).

27) L. B. Kier and L. H. Hall, *J. Pharm. Sci.*, **70**, 583 (1981).

28) M. Randic, A. Sabaljic, S. Nikolic, and N. Trinajstic, *Int. J. Quantum Chem: Quantum Biology Symp.*, **15**, 267 (1988).

29) C. E. Shannon, *Bell. Syst. Tech. J.*, **27**, 379 (1948).

30) S. C. Basak, D. K. Harris, and V. R. Magnuson, *J. Pharm. Sci.*, **73**, 429 (1984).

31) H. Wiener, *J. Am. Chem. Soc.*, **69**, 17 (1947).

32) D. Bonchev and N. Trinajstic, *J. Chem. Phys.*, **67**, 4517 (1977).

33) J. Kim and C. W. Mueller, in "Factor Analysis: Statistical Methods and Practical Issues," ed by E. M. Uslaner, Sage University Press (1978).

34) R. B. Cattell, *Biometrics*, **21**, 190 and 405 (1965).

35) C. Hansch, D. Hoekman, and H. Gao, *Chem. Rev.*, **96**, 1045 (1996).

36) M. Kuanar, R. K. Mishra, and B. K. Mishra, *Indian J. Chem., Sect. A*, **35A**, 1026 (1996).

37) L. H. Hall and L. B. Kier, *J. Chem. Inf. Comput. Sci.*, **35**,

1039 (1995).
38) S. C. Basak and V. R. Magnuson, *Arzneim–Forsch./Drug Res.*, **33**, 501 (1983).
39) J. Galvez, R. Garcia-Domenech, J. V. de Julian-Ortiz, and R. Soler, *J. Chem. Inf. Comput. Sci.*, **35**, 272 (1995).
40) S. C. Basak and G. D. Grunwald, *Chemosphere*, **31**, 2529 (1995).
41) A. Kidera, Y. Konishi, M. Oka, T. Ooi, and H. A. Scherga, *Protein Chem.*, **4**, 23 (1985).
42) M. Chastrette, M. Rajzmann, M. Channon, and F. K. Purcell, *J. Am. Chem. Soc.*, **107**, 1 (1985).