Indian Journal of Chemistry
Vol. 48A, July 2009, pp. 921-929

A molecular structure descriptor derived from bond-disconnection: Application to quantitative structure property relationships

Sagarika Sahoo[a], Sabita Patel[b,]*, Sukalyan Dash[c] & B K Mishra[a,] *

[a]School of Chemistry, Centre of Studies in Surface Science and Technology, Sambalpur University, Jyoti Vihar, Burla 768 019, India
[b]Department of Chemistry, National Institute of Technology, Rourkela 769 008, India
[c]Department of Chemistry, University College of Engineering, Burla 768 018, India
Email: bijaym@hotmail.com

Received 30 April 2009; revised and accepted 27 May 2009

A set of novel molecular structure descriptors, viz., vertex weighted walk parameter, has been derived by using a bond-disconnection protocol in hydrogen depleted vertex weighted molecular graph. The parameters are correlated with other topological indices by constructing the correlation matrix and by correlating with the principal components of these indices. The applicability of these parameters in quantitative structure property relationships has been investigated for boiling points, molar volume at 20 ºC, molar refractions at 20 ºC, heats of vaporization at 25 ºC, critical temperature and surface tension at 20 ºC. Both the single parametric model and multiparametric model have been used for the QSPR studies. Successive exclusion of variables technique has been used to optimize the multiparametric models. The high collinearity of the observed and predicted properties reveals the applicability of the proposed parameters.

Keywords: Theoretical chemistry, Graph theory, Molecular descriptors, Topological indices, Connectivity parameters, Vertex weighted walk parameter, Structure-activity relationships

With the advent of information technology, chemical sciences is now replete with data pools for which, codification, classification and proper utilization of the data necessitate an upward growth. With the proliferated data set, selection of data for optimization has become a potential problem and suffers an enhanced risk of chance correlation. Rucker *et al.*[1] have used a randomization technique for validation of quantitative structure activity/property relationship (QSA/PR) models. Recently, JAVA language has been used to construct a program (OSIRIS) for a complete processing of drug discovery[2] which covers all information from compound synthesis via biological testing, secondary screening evaluation, chemistry-aware data visualization, physicochemical property prediction, three-dimensional pharmacophore comparisons, interactive modeling and computing grid based ligand-protein docking. Tripos Topomer Search Technology (TTST)[3], a novel technique used as a measure of compound similarity, is found to be more promising than the classical QSAR[4] for pharmacophore modeling. Recently, Bender *et al.*[5] employed a diverse subset of the MDDR (Molecular Database Library Drug Data Report) database to present a large scale investigation for virtual screening studies (where orthogonal descriptors are usually chosen for diverse virtual screening hit lists) and to understand descriptor behaviour.

A new Monte Carlo variable selection (MCVS) method was proposed by Konovalov *et al.*[6] and was applied to the blood-brain barrier and human intestinal absorption problems using more than 1600 electronic remote versions of DRAGON molecular descriptors. Only a single descriptor TPSA (NO) (Topological Polar Surface Area using N and O polar Contribution) and ALOGP (atom-type summation logarithm of partition coefficient) could be interpreted as a casual biochemical QSAR relationship for the BBB (Blood Brain Barrier) and HIA (Human Intestinal Absorption) problems respectively yielding low P-values. The MCVS method is equally applicable to the multiple-linear-regression (MLR)-based or non-MLR-based QSAR models.

Wester *et al.*[7] have enumerated graph representations of scaffold topologies for up to eight ring molecules and four-valence atoms, thus providing coverage of the lower portion of the chemical space of small molecules. They have examined scaffold topology distributions for several databases, viz.,

ChemNavigator[8] and PubChem[9], for commercially available chemicals, the Dictionary of Natural Products, a set of 2742 launched drugs, WOMBAT (World of Molecular Bioactivity)[10], a database of medicinal chemistry compounds, and two subsets of PubChem, "actives" and DSSTox (Distributed Structure-searchable Toxicity)[11] comprising toxic substances.

The molecular descriptors generated by Shape Signatures method[12] was utilized by Chekmarev *et al.*[13] with support vector machines (SVM) and Kohonen self-organizing maps (Kohonen SOM)[14] techniques, which perform better in classification problems related to the analysis of highly clustered and heterogeneous property spaces. Such models are utilized to predict the potential for cardiotoxicity in drug discovery and elucidating the QSPR model for properties of pharmaceutical interest like aqueous solubility (Log *S*), melting point (T_m), and octanol-water partition coefficient (Log *P*)[15].

With the development of novel statistical tools and proliferated molecular descriptors, the QSAR or QSPR studies have become more complicated. However, the simple graph theoretical parameters have made significant contributions to solve the problems through simple statistical tools like multiple regression analysis. Based on the above, a novel topological parameter has been deduced from the structural graph of some hydrocarbons and has been used for QSPR of alkane isomers.

Theoretical

Topological descriptors

The topological descriptors have been derived for C-3 to C-9 hydrocarbons by using the methods reported below.

The molecular connectivity index (χ) as proposed by Randic[16], Kier and Hall[17] was calculated from the hydrogen suppressed molecular graph. A valence 'δ' value was assigned to the constituting atom by considering the number of non-hydrogen atoms bonded to it.

The 'n' order connectivity index, ${}^n\chi$, is the sum of all connecting bonds,

$$ {}^n\chi = \Sigma\, C_k \quad \ldots (1)$$

where the connectivity value for first order (connecting atom i and j) is computed as $C_k = 1/\sqrt{\delta_i\delta_j}$, second order (connecting bonds i-j-k) as $C_k = 1/\sqrt{\delta_i\delta_j\delta_k}$ and third order (connecting bonds i-j-k-l) as $C_k = 1/\sqrt{\delta_i\delta_j\delta_k\delta_l}$.

In order to evaluate the molecular topological indices the following algebraic operations on adjacency [A], distance [D], reciprocal of distance [H], walk [W] and valence [V] matrices were carried out[18-20].

$$ A_2 = \Sigma\,\Sigma\,[A^2] \quad \ldots (2)$$

$$ S_D = \Sigma\,\Sigma\,[AD]_i \quad \ldots (3)$$

$$ S_H = \Sigma\,\Sigma\,[AH]_{ij} \quad \ldots (4)$$

$$ S_W = \Sigma\,\Sigma\,[AW]_{ij} \quad \ldots (5)$$

$$ WW^{(1)} = [W]/2 \quad \ldots (6)$$

$$ DW^{(1)} = [D]/2 \quad \ldots (7)$$

$$ HW^{(1)} = [H]/2 \quad \ldots (8)$$

The above defined equations were used to obtain the following molecular topological indices:

$$ {}^1W = \tfrac{1}{2}\,\Sigma\,\Sigma\, d_{ij} \quad \ldots (9)$$

where d_{ij} are the elements of [D].

$$ {}^1H = \tfrac{1}{2}\,\Sigma\,\Sigma\, h_{ij} \quad \ldots (10)$$

where h_{ij} are the elements of the matrix.

$$ R = \tfrac{1}{2}\,\Sigma\,\Sigma\, w_{ij} \quad \ldots (11)$$

where w_{ij} are the elements of the walk matrix

$$ DI = \Sigma[DW^{(1)}\,(A+D)] \quad \ldots (12)$$

$$ {}^2W = \{DI - (S_D/2)\}/2 \quad \ldots (13)$$

$$ HI = \Sigma[HW^{(1)}\,(A+H)] \quad \ldots (14)$$

$$ WI = \Sigma[WW^{(1)}\,(A+W)] \quad \ldots (15)$$

$$ MTI = \Sigma\,\Sigma[A^2+AD]_{ij} \quad \ldots (16)$$

$$ VD = \Sigma\,\Sigma(V \times D)_{ij} \quad \ldots (17)$$

$$ VW = \Sigma\,\Sigma(V \times W)_{ij} \quad \ldots (18)$$

$$ VWI = \Sigma\,\Sigma\,[W \times (V \times W)]_{ij} \quad \ldots (19)$$

$$ VDI = \Sigma\,\Sigma\,[DW^{(1)}\,(V \times D)]_{ij} \quad \ldots (20)$$

$$ VMTI = \Sigma\,\Sigma[V^2+VD]_{ij} \quad \ldots (21)$$

A new set of vertex weighted walk (VWW) descriptors considering the disconnection of bonds in

the hydrogen depleted vertex weighted molecular graph has been derived.

In a vertex weighted graph, a valence δ^v value for each atom can be assigned as:

$$\delta_i^v = \sigma_i + p_i + n_i \quad \ldots (22)$$

where δ_i^v is the number of non-hydrogen valence electrons contributed by atom i, and σ_i and π_i are the number of sigma and pi bonds, and n_i is the number of non-bonding electrons[21,22].

Various order of VWW* can be calculated by considering generation of fragments (ζ) after disconnection of a single bond (i,j), two consecutive single bonds (i,k), three consecutive single bonds (i,l), four consecutive single bonds (i,m), etc.,

$$\zeta_{i,j/k/l/m} = \left(\prod_{i=1}^{n} \delta_i^v\right)^{1/n} \quad \ldots (23)$$

'n' being the number of atoms in that fragment.

$$C_p = \prod_{i=1}^{n} \zeta_{i,j/k/l/m} \quad \ldots (24)$$

$$\mathrm{VWW}^* = \sum_{p} \quad \ldots (25)$$

p = number possible bond disconnection and accordingly * = 1, 2, 3 or 4 for one, two, three and four consecutive bond disconnections respectively or order of VWW. An example of calculation of VWW^1 has been presented in Table 1.

The possible disconnections on the valence-weighted molecular graph are presented in Fig. 1.

From Table 1, $VWW^1 = \Sigma C_p = 12.272$.

The possible disconnections for second, third and fourth order VWW are represented in Figs 2 – 4.

Table 1—Possible disconnection for VWW^1 calculation

Possible disconnections	[a] $\zeta_{i,j} = \left(\prod_{i=1}^{n} \delta_i^v\right)^{1/n}$	$C_p = \prod_{i=1}^{n} \zeta_{i,j}$
a	$\zeta_{i,j}^{1} = (1\times2\times3\times4\times1\times1\times1)^{1/7} = 1.575$ $\zeta_{i,j}^{2} = (1)^{1/1} = 1$	$1.575\times1 = 1.575$
b	$\zeta_{i,j}^{1} = (1\times2\times3\times4\times1\times1\times1)^{1/7} = 1.575$ $\zeta_{i,j}^{2} = (1)^{1/1} = 1$	$1.575\times1 = 1.575$
c	$\zeta_{i,j}^{1} = (1\times3\times1)^{1/3} = 1.442$ $\zeta_{i,j}^{2} = (2\times4\times1\times1\times1)^{1/5} = 1.516$	$1.442\times1.516 = 2.186$
d	$\zeta_{i,j}^{1} = (1\times1\times3\times2)^{1/4} = 1.565$ $\zeta_{i,j}^{2} = (4\times1\times1\times1)^{1/4} = 1.414$	$1.565\times1.414 = 2.213$
e	$\zeta_{i,j}^{1} = (1\times1\times3\times2\times4\times1\times1)^{1/7} = 1.575$ $\zeta_{i,j}^{2} = (1)^{1/1} = 1$	$1\times1.575 = 1.575$
f	$\zeta_{i,j}^{1} = (1\times1\times3\times2\times4\times1\times1)^{1/7} = 1.575$ $\zeta_{i,j}^{2} = (1)^{1/1} = 1$	$1\times1.575 = 1.575$
g	$\zeta_{i,j}^{1} = (1\times1\times3\times2\times4\times1\times1)^{1/7} = 1.575$ $\zeta_{i,j}^{2} = (1)^{1/1} = 1$	$1\times1.575 = 1.575$

[a] n is the number of atoms present in the fragment

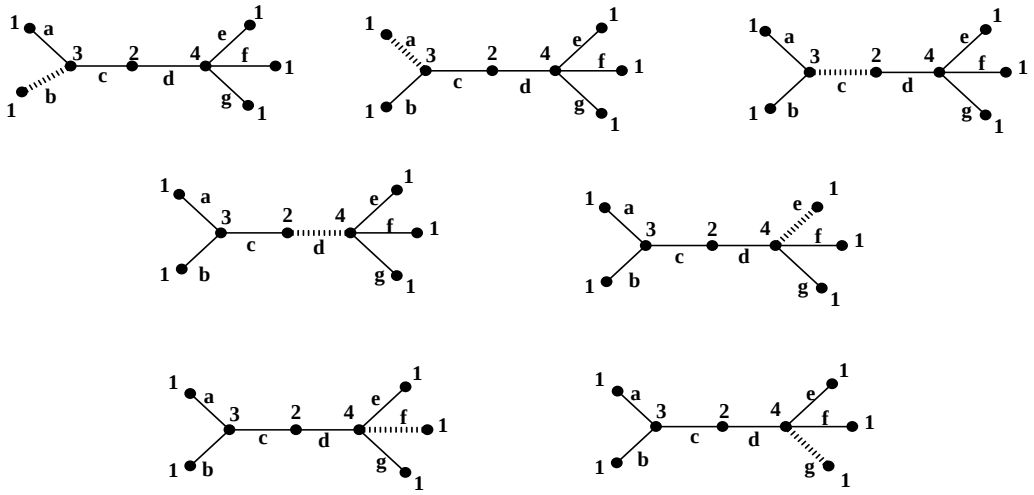

Fig. 1—First order VWW graph.

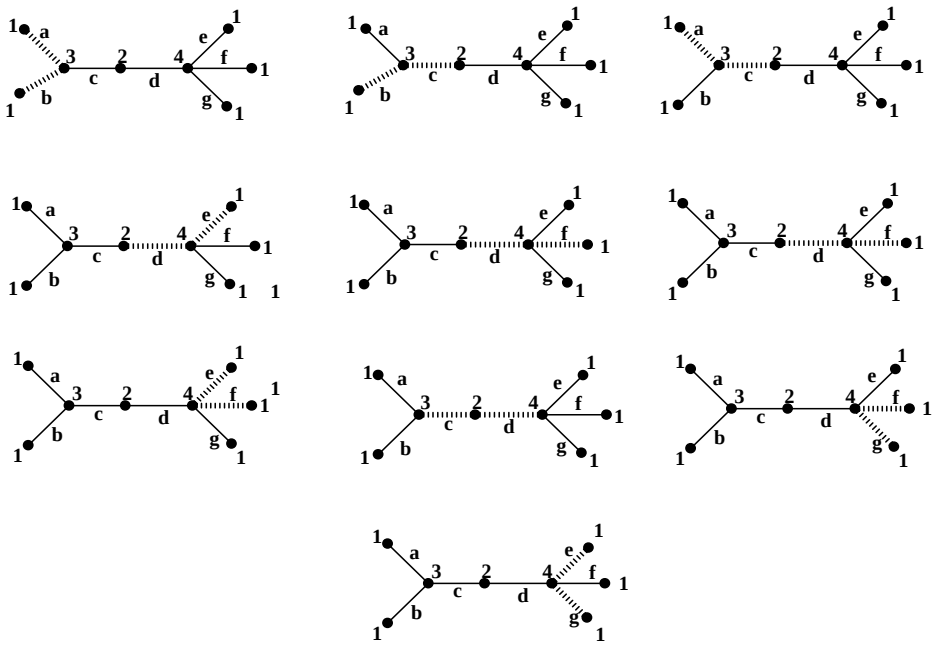

Fig. 2—Second order VWW graph.

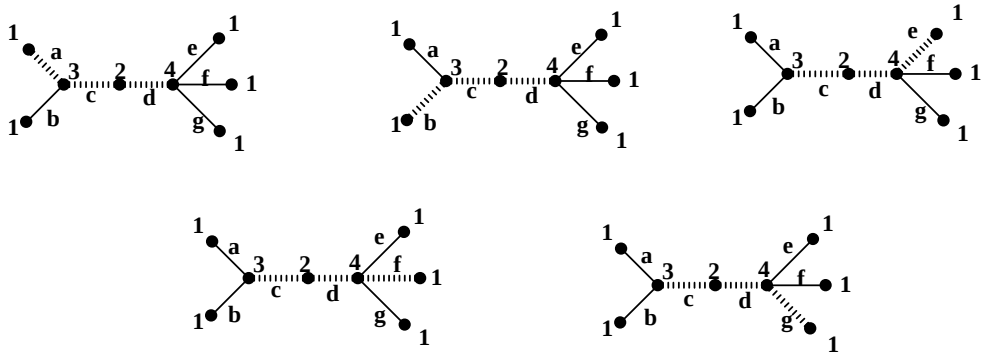

Fig. 3—Third order VWW graph.

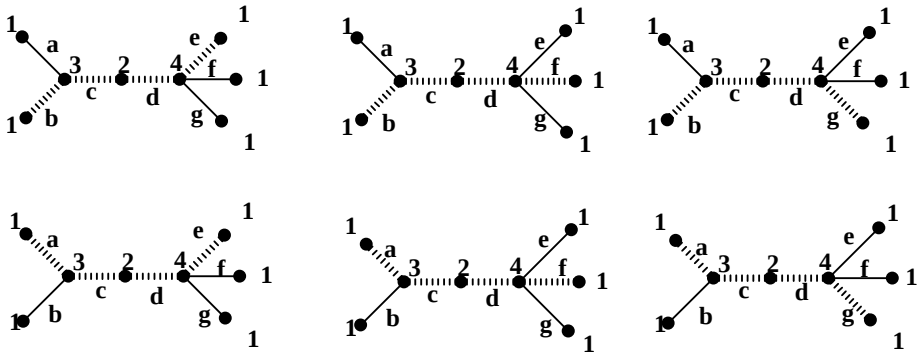

Fig. 4—Fourth order VWW graph.

The VWW^1, VWW^2, VWW^3 and VWW^4 have been calculated for C-3 to C-9 hydrocarbons and are listed in Table 2.

Physico-chemical properties of alkanes

The physico-chemical properties selected for the present investigation are boiling points (BP), molar volume (MV) at 20 ºC, molar refractions at 20 ºC (MR), heats of vaporization at 25 ºC (HV), critical temperature (T_c) and surface tension at 20 ºC (ST)[23].

Methodology

During the quantitative structure property/activity relationship (QSP/AR) studies, various indices are usually incorporated successively into the regression model followed by analysis for optimization. In the present work, a successive exclusion of variable (SEV) technique has been suggested where, from a basic regression model, successive regression models were derived by exclusion of variable having the minimum Student-'t' value. The resultant statistical

Table 2—VWW^1, VWW^2, VWW^3, VWW^4 of C-3 to C-9 alkanes

No.	Alkane	VWW^1	VWW^2	VWW^3	VWW^4	No.	Alkane	VWW^1	VWW^2	VWW^3	VWW^4
1	3	2.828	1.000	0.000	0.000	38	2233MMMM4	10.916	22.992	22.679	0
2	4	5.175	2.828	1.000	0.000	39	9	20.305	17.019	13.827	10.759
3	2M3	4.327	5.196	0.000	0.000	40	2M8	18.521	19.985	14.964	11.972
4	5	7.853	5.175	2.828	1.000	41	3M8	18.41	21.245	18.605	13.543
5	2M4	6.735	8.158	3.464	0.000	42	4M8	18.381	21.641	19.83	14.965
6	22M3	5.657	9.524	0.000	0.000	43	3E7	18.144	22.647	24.074	16.569
7	6	10.759	7.853	5.832	2.828	44	4E7	18.076	22.955	24.909	19.932
8	2M5	9.434	10.965	6.341	3.464	45	22MM7	16.579	22.396	13.985	11.075
9	3M5	9.358	11.734	6.716	1.732	46	23MM7	16.816	23.984	27.16	17.058
10	22MM4	8.063	13.194	4.762	0.000	47	24MM7	16.778	24.100	20.317	23.998
11	23MM4	8.287	13.456	12.00	0.000	48	25MM7	16.781	23.797	19.293	14.286
12	7	13.827	10.605	7.853	5.175	49	26MM7	16.871	22.642	15.903	12.923
13	2M6	12.33	13.861	9.104	6.341	50	33MM7	16.437	24.216	18.867	11.967
14	3M6	12.229	14.906	12.21	6.716	51	34MM7	16.689	25.255	32.31	22.93
15	3E5	12.069	15.834	15.419	3.634	52	35MM7	16.691	24.953	22.684	23.626
16	22MM5	10.738	16.3	8.149	4.762	53	44MM7	16.407	24.638	20.305	15.179
17	23MM5	10.927	17.359	19.01	6.000	54	23ME6	16.565	25.197	33.333	22.869
18	24MM5	10.990	16.488	9.992	12.00	55	24ME6	16.568	25.042	23.837	24.828
19	33MM5	11.251	17.461	11.217	1.587	56	33ME6	16.208	26.004	24.604	12.996
20	223MMM4	9.605	16.698	16.497	0.000	57	34EM6	16.478	26.198	37.323	26.19
21	8	17.019	13.828	10.758	7.854	58	223MMM6	15.065	25.888	29.259	15.758
22	2M7	15.37	16.87	11.97	9.103	59	224MMM6	15.036	25.7	17.811	22.768
23	3M7	15.261	18.051	15.43	10.348	60	225MMM6	15.098	24.68	14.743	11.767
24	4M7	15.238	18.318	16.328	12.584	61	233MMM6	15.030	26.565	30.192	13.354
25	3E6	15.03	19.299	19.944	12.139	62	234MMM6	15.260	27.239	38.346	32.926
26	22MM6	13.592	19.386	11.076	8.149	63	235MMM6	15.331	26.181	27.042	23.307
27	23MM6	13.798	20.723	23.506	8.473	64	244MMM6	15.000	26.318	19.221	21.718
28	24MM6	13.784	20.586	16.07	17.278	65	334MMM6	15.113	27.338	34.134	16.487
29	25MM6	13.865	19.536	12.924	9.992	66	33EE5	16.008	27.456	29.547	10.447
30	33MM6	13.459	20.928	15.449	5.609	67	223MME5	14.890	24.493	30.158	16.635
31	34MM6	13.702	21.636	27.192	14.78	68	233MME5	14.855	27.849	35.407	13.295
32	3E2M5	13.624	21.625	27.944	14.32	69	234MEM5	15.166	27.198	39.414	33.805
33	3E3M5	13.297	22.238	19.491	5.045	70	2233(M)5	13.498	27.688	32.129	7.559
34	223MMM5	12.226	22.418	24.409	8.248	71	2234(M)5	13.767	27.659	34.324	28.573
35	224MMM5	12.272	21.535	11.766	16.497	72	2244(M)5	13.540	26.280	13.469	22.679
36	233MMM5	12.199	22.899	25.005	5.499	73	2334(M)5	13.755	28.273	38.638	19.049
37	234MMM5	12.483	22.815	29.308	20.785						

parameters like R^2, F and RMS (residual mean square) were analyzed. With decrease in independent variable, R^2 value decreases and F and RMS value increases if the exclusion leads to optimization. The regression model with maximum F and minimum RMS values were considered for the optimized regression model.

Results and Discussion

The VWW parameters have been derived with a view to quantify the interaction of one atom with other atom within the molecule present at varying distances which may contribute or be able to explain the experimental factors like inductive effect, mesomeric effect, NOE (Nuclear Overhauser Effect), etc. We are considering up to fourth order of the VWW since the magnitude of the interaction of the atoms becomes negligible with increase in the number of bonds[24]. These interactions of various bond distances contribute to the physical and chemical properties of the compound.

When the TI values are the same for the two different molecular graphs, they are said to be degenerate. Balaban[25] observed a high degeneracy for TI values like $^1\chi$ and ^{1}W. In many cases the degeneracy can be explained by the similar pattern of molecular graphs. Furthermore, in alkane isomers the TI values, $^1\chi$, ^{1}W, ^{2}W, R, DI, WI, MTI, VD, VW, VDI and VWI decrease with increasing branching and the values of TI such as $^2\chi$, ^{1}H, ^{2}H, HI, VMTI increase with increase in branching. The analysis of VWW data reveals that VWW^1 decreases with increasing branching while VWW^2 and VWW^3 increase with increase in branching in the alkane chain. To check the interrelationships of the variables, the cross-correlation matrix of the parameters of alkane isomers is represented in Table 3.

Analysis of the cross-correlation matrix reveals the following:

1 A good inter-relationship is observed between almost all the molecular descriptors.

2 The new parameters, VWW^1 and VWW^2, show high correlation coefficient with ^{1}W, ^{2}W, DI, R, MTI, VD, VW, VDI, VWI and $^1\chi$, while its third and fourth order indices show comparatively lesser correlation coefficient.

To have better insight towards the interrelationships, the TIs have been subjected to principal component analysis (PCA). PCA reduces the size of the descriptors and at the same time proposes some new orthogonal descriptors for the molecules. Each descriptor has more or less contribution from all the test descriptors with variable loading. The results of PCA are given in the Table 4.

Table 3—Simple correlation matrix table of TIs of alkane isomers

	^{1}W	^{2}W	DI	H^1	H^2	HI	R	WI	MTI	VD	VW	VDI	VWI	VMTI	$^1\chi$	$^2\chi$	$^3\chi$	VWW^1	VWW^2	VWW^3	VWW^4
^{1}W	1																				
^{2}W	0.98	1																			
DI	0.98	1	1																		
H^1	0.92	0.83	0.84	1																	
H^2	0.89	0.80	0.80	0.99	1																
HI	0.89	0.80	0.80	0.99	1	1															
R	0.98	0.99	0.99	0.82	0.77	0.77	1														
WI	0.91	0.97	0.97	0.69	0.65	0.65	0.97	1													
MTI	1	0.98	098	0.92	0.88	0.89	0.98	0.91	1												
VD	0.99	0.99	0.99	0.90	0.87	0.87	0.98	0.93	0.99	1											
VW	0.99	0.99	0.99	0.88	0.86	0.86	0.98	0.94	0.99	1	1										
VDI	0.99	1	1	0.87	0.84	0.84	0.99	0.95	0.99	1	1	1									
VWI	0.94	0.99	0.99	0.76	0.72	0.72	0.99	0.99	0.95	0.96	0.97	0.97	1								
VMTI	0.96	0.93	0.93	0.96	0.96	0.96	0.9	0.82	0.96	0.97	0.96	0.95	0.87	1							
$^1\chi$	0.97	0.91	0.91	0.94	0.9	0.9	0.91	0.81	0.96	0.94	0.91	0..93	0.85	0.92	1						
$^2\chi$	0.63	0.53	0.53	0.81	0.82	0.83	0.51	0.41	0.63	0.61	0.62	0.57	0.48	0.75	0.59	1					
$^3\chi$	0.65	0.54	0.55	0.81	0.83	0.83	0.52	0.38	0.64	0.62	0.59	0.59	0.45	0.72	0.76	0.52	1				
VWW^1	0.96	0.94	0.94	0.86	0.81	0.81	0.95	0.88	0.96	0.94	0.92	0.94	0.9	0.88	0.98	0.49	0.64	1			
VWW^2	0.82	0.72	0.72	0.97	0.98	0.98	0.68	0.55	0.82	0.8	0.78	0.76	0.62	0.90	0.85	0.83	0.84	0.74	1		
VWW^3	0.65	0.55	0.55	0.81	0.83	0.83	0.5	0.36	0.64	0.64	0.60	0.59	0.44	0.74	0.72	0.55	0.90	0.59	0.86	1	
VWW^4	0.73	0.69	0.69	0.75	0.74	0.75	0.65	0.56	0.73	0.74	0.72	0.72	0.61	0.76	0.72	0.53	0.52	0.66	0.75	0.73	1

Analysis of the data of Table 4 reveals the following:

1 The PC1 can explain 79.1% of variance of the total descriptors. The cumulative percent of variance with second PC is found to be 89.1%. Subsequent addition of each principal component increases the cumulative percent of variance.

2 When the PC1 and PC2 of 73 alkane isomers are plotted against each other it is found that the alkanes are clubbed in specific domain according to their molecular size. All nonane isomers are clubbed in one region; similarly the octanes, heptanes, hexanes, pentanes, butanes and propane are also arranged (Fig. 5).

3 The contribution of each descriptor towards the PCs is also determined from the simple correlation coefficient with all the TIs and it is found that all the TIs contribute their maximum towards the first principal component, while $^3\chi$ and VWW^3 contribute moderately to the second PC.

4 The correlation coefficients of TIs with respect to the PC1 are plotted against the correlation coefficient of PC2 (Fig. 6). The plot further classifies the TIs into different classes.

Table 4—Eigen values, percent of variance and cumulative percent of variance used in principal component analysis

Principal component (PC)	Eigen values	Variance (%)	Cumulative variance (%)
PC1	16.60	79.1	79.1
PC2	2.090	10.0	89.1
PC3	1.024	4.90	93.9
PC4	0.621	3.00	96.8
PC5	0.442	2.10	98.9
PC6	0.116	0.60	99.5
PC7	0.049	0.20	99.7
PC8	0.028	0.10	99.9

Monovariable regression model

All the TIs are subjected to multiple regression analysis for a single parametric equation,

$$\text{Physicochemical property} = a\,TI + b \quad \ldots (26)$$

where 'a' is the sensitivity of the TI towards the physicochemical property of the testing molecules and 'b' is a constant. In monovariable regression model, the alkane isomers are found to have good regression coefficient values (Table 5).

The boiling point values correlate well with the $^1\chi$, having F value of 2379.4 with R = 0.98 indicating the level of confidence of the regression model to be 99.9%. High correlation with properties like T_c (R = 0.96, F = 948.04), HV (R >0.9951, F = 8984.33), ST (R = 0.91, F = 303.62) have also been observed. Molar volume values correlate well with ^{1}W values (R = 0.97, F = 1142.5), while molar refraction correlate well with ^{1}H values (R = 0.98, F = 1751.95).

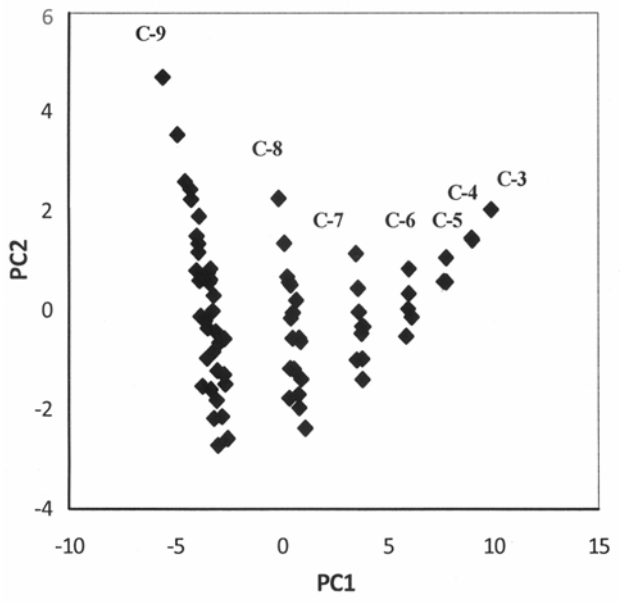

Fig. 5—Plot of first principal component versus second principal component.

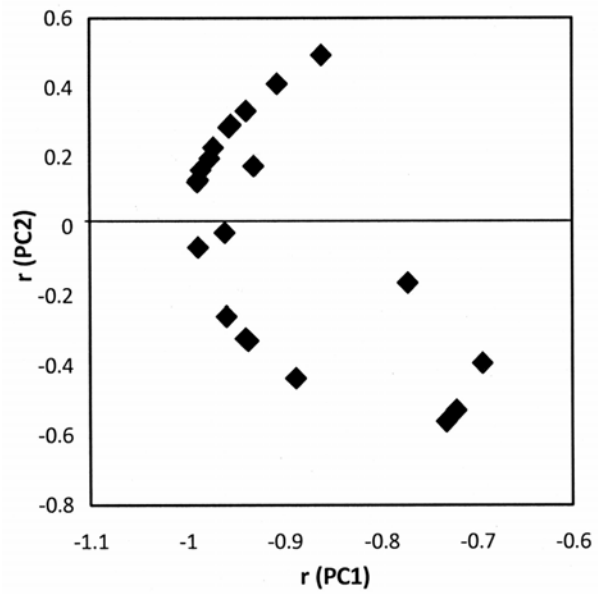

Fig. 6—Plot of the correlation coefficient of PC1 versus PC2 of TIs.

Optimization of regression model with multivariable equations

Though the monovariable equations are found to have significant predictability, attempts have been taken to obtain equations with higher confidence level by using a multiparametric regression model.

The BP and TC of 73 alkanes, MV, MR, HV of 69 alkanes and ST of 68 alkanes have been correlated with the 15 molecular descriptors (^{1}W, ^{2}W, DI, ^{1}H, ^{2}H, HI, R, WI, MTI, VD, VW, VDI, VWI, VMTI and VWW). The t-values were considered for optimization, and the variables having least t-value were excluded and then further subjected to regression analysis. Improved regression models were indicated by increasing F and R^2 values and with minimum residual mean square (RMS) value. A representative example of successive exclusion of variables for optimizing the regression model has been reported for BP in Table 6. For the BP, TC, MV, MR, HV and ST, respective optimized regression models represented by Eqs 27-32 have been proposed.

Table 5—Regression coefficient (R) of monovariable regression model

TIs	R					
	BP	MV	MR	HV	ST	Tc
^{1}W	0.93	0.97	0.96	0.96	0.81	0.90
^{2}W	0.85	0.92	0.90	0.92	0.74	0.80
DI	0.86	0.92	0.90	0.92	0.74	0.81
^{1}H	0.96	0.96	0.98	0.91	0.87	0.96
^{2}H	0.92	0.92	0.95	0.86	0.85	0.93
HI	0.93	0.93	0.96	0.87	0.85	0.94
R	0.86	0.91	0.88	0.91	0.71	0.81
WI	0.74	0.82	0.78	0.82	0.60	0.67
MTI	0.93	0.97	0.96	0.96	0.80	0.89
VD	0.90	0.96	0.95	0.95	0.79	0.86
VW	0.87	0.95	0.94	0.92	0.76	0.83
VDI	0.88	0.94	0.93	0.94	0.77	0.84
VWI	0.79	0.87	0.83	0.86	0.65	0.73
VMTI	0.91	0.97	0.98	0.92	0.83	0.90
$^1\chi$	0.98	0.95	0.96	0.99	0.91	0.96
$^2\chi$	0.66	0.69	0.68	0.45	0.41	0.67
$^3\chi$	0.81	0.59	0.68	0.70	0.85	0.86
VWW^1	0.94	0.90	0.89	0.97	0.82	0.90
VWW^2	0.89	0.86	0.90	0.79	0.80	0.91
VWW^3	0.76	0.63	0.70	0.69	0.82	0.80
VWW^4	0.69	0.74	0.75	0.70	0.61	0.68

Table 6—Optimization of regression model for predicting boiling point of alkanes[a]

t_{min}	Excluded var.	R^2	F	RMS
0.18	^{2}H	0.9972	1453.4	5.7734
0.26	DI	0.9971	1590.2	5.7213
-0.39	HI	0.9971	1747.3	5.6024
-1.35	R	0.9971	1880.2	5.6790
1.93	VD	0.9964	1980.2	5.9305
1.71	VMTI	0.9967	2134.3	6.1134
2.49	WI	0.9964	2220.4	6.6082
-0.43	VWI	0.9964	2569.9	6.5251
6.52	VW	0.9940	1836.1	10.629

[a]Basic regression model: Explanatory variables: 15 molecular descriptors; N = 73, R^2 = 0.9972, F = 1336.6, RMS = 5.8711.

BP (^{o}C) =(14.2216±2.0954)^{1}W+(0.3134±0.0385) ^{2}W
+(13.2056±0.8453) ^{1}H–(3.8061±0.5187) MTI
+(0.0092±0.0014) VW–(0.1045±0.0121) VDI
+ (6.5133±1.0587) VWW–(81.9501±3.5564)
... (27)

(N= 73, R^2 = 0.9964, F = 2569.9, RMS = 6.5251)

T_c (^{o}C) = (12.581±0.5113) ^{1}H - (0.4451±0.0375) MTI
+ (12.769±0.8514) VWW + (42.144±7.1847)
... (28)

(N = 73, R^2 = 0.9821, F = 1259.4, RMS = 45.772)

MV(cm^3) = -(6.5208±0.7363) ^{1}W-(0.0684±0.0073) DI
+ (18.937±1.6152) ^{1}H–(1.1551±0.0978) HI
+ (1.7360±0.1646) MTI + (0.0399±0.0048)
VDI + (0.2815±0.3727) VWW
+ (51.958±4.1902) ... (29)

(N = 69, R^2 = 0.9984, F = 4763.5, RMS = 0.6060)

MR = – (0.0011±0.0001) DI + (2.8728±0.1367)^{1}H
– (0.1214±0.0072) HI + (0.0388±0.0033) MTI
– (0.1183±0.0186) VWW + (12.146±0.3145)
... (30)

(N = 69, R = 0.9999, R^2 = 0.9997, F = 50279.2,
RMS = 0.0074)

HV (kJ mol^{-1}) = (0.0487±0.0031) HI
– (0.0225±0.0039) MTI
+ (1.7916±0.0930) VWW
+ (1.0612±0.6568) ... (31)

(N = 69, R^2 = 0.9890, F = 2038.7, RMS = 0.3168)

ST (dyne cm^{-1}) = (2.3229±0.2969) ^{1}W
+ (0.0233±0.0027) DI
+ (0.1018±0.0072) HI
– (0.6560±0.0738) MTI
+ (0.0008±0.0002) VW
– (0.0143±0.0018) VDI
+ (0.6183±0.1524) VWW
+ (9.3768±0.7224) … (32)

(N = 68, R^2 = 0.9755, F = 341.58, RMS = 0.1017).

Table 7—Predicted MV, MR, HV and ST values of the compounds 3, 4, 2M3, 22M3 and 2233M4 from optimized regression models

Properties	Eqn	3	4	2M3	22M3	2233M4
MV (cm^3)	8	84.1702	100.016	102.328	—-—	158.255
MR	9	17.6972	21.3105	21.4692	---—	38.8607
HV (kJ mol^{-1})	10	6.52270	11.0790	9.75660	---—	9.75660
ST (dyne cm^{-1})	11	11.4026	13.6411	12.4647	13.3659	20.6838

By using the proposed optimized equations, calculated physical properties have been plotted against the observed values. The linearity of the plots justifies the applicability of the regression models.

The experimental values of MV, MR, HV and ST of propane, butane, 2-methyl propane and 2,2,3,3-tetramethylbutane are not available in the literature. Using the optimized regression equations (Eqs 29-32), the values above have been successfully predicted and are listed in Table 7.

Conclusions

The graph theoretical parameters discussed above are found to be capable of distinguishing chemical structures of various isomers of an alkane from one another. The principal components of the molecular descriptors in all the compounds are due to the contributions of individual descriptors of the concerned compounds. Though PCs seem to be abstract values, the plot of *r*(PC1) versus *r*(PC2) exhibits a distinct demarcation between various alkane series. Furthermore, the novel parameters VWW are found to correlate well with different molecular properties of the alkane series.

Acknowledgement

The authors thank University Grants Commission, New Delhi, for financial assistance through Departmental Research Support (DRS) programme.

References

1 Rucker C, Rucker G & Meringer M, *J Chem Inf Model*, 47 (2007) 2345.
2 Sander T, Freyss J, von Korff M, Reich J R. & Rufener C, *J Chem Inf Model*, 49 (2009) 232.
3 Cramer R D, Jilek R J & Andrews K M, *J Mol Graph Model*, 20 (2002) 447.
4 Nisius B & Goller A H, *J Chem Inf Model*, Article ASAP (2009).
5 Bender A, Jenkins J L, Scheiber J, Chetan S, Sukuru K, Glick M & Davies J W, *J Chem Inf Model*, 49 (2009) 108.
6 Konovalov D A, Sim N, Deconinck E, Heyden Y V & Coomans D, *J Chem Inf Model*, 48 (2008) 370.
7 Wester M J, Pollock S N, Coutsias E A, Allu T K, Muresan S & Oprea T I, *J Chem Inf Model*, 48 (2008) 1311.
8 iResearch Library, ChemNavigator.com, Inc 2006, http://www.chemnavigator.com/ (accessed Dec 7, 2007).
9 PubChem, National Center for Biotechnology Information, 2006, http://pubchem.ncbi.nlm.nih.gov/(accessed Dec 7, 2007).
10 Olah M, Rad R, Ostopovici L, Bora A, Hadaruga N, Hadaruga D, Moldovan R, Fulias A, Mracec M & Oprea T, *Chemical Biology: From Small Molecules to Systems Biology and Drug Design*, edited by S L Schreiber, T M Kapoor & G Wess (Wiley-VCH: New York) 2007.
11 *Distributed Structure-Searchable Toxicity (DSSTox)*, 2007 (US Environmental Protection Agency) http://epa.gov/ncct/dsstox/ (accessed Dec 7, 2007).
12 Zauhar R J, Moyna G, Tian L L Z & Welsh W J, *J Med Chem*, 46 (2003) 5674.
13 Chekmarev D S, Kholodovych V, Balakin K V, Ivanenkov Y, Ekins S & Welsh W J, *Chem Res Toxicol*, 21 (2008) 1304.
14 Kohonen T, *Self-Organizing Maps*, 3rd Edn (Springer Verlag, New York) 2000.
15 Hughes L D, Palmer D S, Nigsch F & Mitchell J B O, *J Chem. Inf Model*, 48 (2008) 220.
16 Randic M, *J Am Chem Soc*, 97 (1975) 6609.
17 Kier L B & Hall L H, *Molecular Connectivity in Chemistry and Drug Research*, (Wiley, New York) 1986.
18 Diudea M V, Topan M & Graovac A, *J Chem Inf Comput Sci*, 34 (1994) 1072.
19 Schultz H P, *J Chem Inf Comput Sci*, 29 (1989) 227.
20 Kuanar M & Mishra B K, *J Serb Chem Soc*, 62 (1997) 289.
21 Kier L B & Hall L H, *J Pharm Sci*, 70 (1981) 583.
22 Randic M, Sabljic A, Nikolic S & Trinajstic N, *Int J Quant Chem; Quant Biol Symp*, 15 (1988) 267.
23 (a) *CRC Handbook of Chemistry and Physics*, 73rd Edn, (CRC Press, Boca Raton FL)1992-1993; (b) Needhan D E, Wei I C & Seybold P G, *J Am Chem. Soc*, 110 (1986) 4186.
24 Smith M B & March J, *March's Advanced Organic Chemistry*, 6th Edn (Wiley-Interscience, Hoboken, NJ) 2007.
25 Balaban A T, *J Chem Inf Comput Sci*, 34 (1994) 398.