Molecular Similarity and Estimation of Molecular Properties

Subhash C. Basak* and Gregory D. Grunwald

Natural Resources Research Institute, The University of Minnesota, 5013 Miller Trunk Highway, Duluth, Minnesota 55811

Received September 8, 1994[⊗]

Five molecular similarity methods have been used to select K nearest neighbors of chemicals for K = 1−10, 15, 20, 25. The properties of the selected neighbors have been used to estimate properties of two sets of chemicals: (a) normal boiling point of a group of 139 hydrocarbons and (b) mutagenicity of a set of 95 aromatic and heteroaromatic amine compounds. The similarity methods are based on calculated topological indices and atom pairs. The results show that each of these methods give reasonable estimates of molecular properties investigated in this paper.

1. INTRODUCTION

In recent years, there has been an upsurge of interest in the use of molecular similarity methods in drug design, selection of analogs for chemicals, and estimation of properties.[1−10] In drug design, similarity/dissimilarity based methods have been very useful in the rational selection of candidate chemicals from large databases.[9] In risk assessment, we often have to select analogs of chemicals of interest when adequate sets of test data are not available.[11] The properties of the selected "similar" neighbors can be used to estimate the hazard posed by the chemical of interest.

The use of molecular similarity methods is based on the structure−property similarity principle. This notion states that similar structures usually have similar properties.[1,3] The principle has been used in ordering a wide range of chemical phenomena ranging from atomic properties to macromolecular behavior.[12] Intermolecular similarity can be defined in terms of the number of structural features and their mutual arrangements which two chemical species have in common.[13] The structural feature(s) used to quantify similarity will vary with the level of organization to which the chemical species belong, viz., atomic, molecular, macromolecular, etc. It is also dependent on the mode of representation of the species, choice of the set of structural descriptors, and the selection of the particular mathematical function used to quantify similarity from the chosen set of descriptors.

Recently, a number of different methods have been developed for quantitative molecular similarity analysis (QMSA) of chemicals.[1−8,10] Any similarity method gives us an ordered set of molecules (analogs) structurally related to the chemical of interest. Properties of the K selected neighbors (analogs) can then be used to estimate properties of the candidate chemical.[14] Similarity of chemicals has been quantified using empirical and nonempirical properties or parameters.[1,4−8,10,12] In view of the fact that the majority of new and existing chemicals will have scanty experimental data, QMSA methods based on nonempirical parameters would be most useful for hazard assessment and drug design.

Graph theoretical parameters like topological indices and substructures have been used in the quantification of molecular similarity.[1−10,12] In a recent paper, Basak and Grunwald[14] used properties of the closest neighbor of chemicals to estimate properties. The similarity methods used in the study were based on topological indices and atom pairs. It was of interest to see whether a higher number of K neighbors ($K > 1$) results in better property estimation. Therefore, in this paper, we have carried out a comparative study of five QMSA techniques in predicting (a) boiling points of a set of 139 hydrocarbons and (b) mutagenic potency of 95 aromatic and heteroaromatic amines. Results of these analyses are presented along with a critical evaluation of the applicability of these QMSA techniques in estimating molecular properties.

2. METHODS

2.1. Databases. The 139 hydrocarbons used in this study were taken from literature and included 73 C_3−C_9 alkanes,[15] 29 alkylbenzenes,[16] and 37 polycyclic aromatic hydrocarbons (PAHs).[17] A listing of the hydrocarbons is presented in Table 1.

The 95 aromatic and heteroaromatic amines used to study mutagenic potency were taken from the literature.[18] The mutagenic activities of these compounds in *S. typhimurium* TA98 + S9 microsomal preparation had been collected and are expressed as the mutation rate, ln(R), in log(revertants/nmol). Table 2 lists the compounds used for this study.

2.2. Topological Indices and Atom Pairs. The calculation of the topological indices (TIs) and atom pairs[2] have previously been described in detail.[5] However, four additional indices developed by A. T. Balaban[19−21] were calculated and used in these analyses. Balaban denoted these as J indices and they are based upon the distance sums s_i of a chemical graph. J is defined as

$$J = q(\mu + 1)^{-1} \sum_{i,j \text{ edges}} (s_i s_j)^{-1/2} \qquad (1)$$

where the cyclomatic number μ (or number of rings in graph) is $\mu = q - n + 1$, with q adjacencies or edges and n vertices. In the original definition of J, the term s_i referred to either the row distance sum for vertex i in the distance matrix (**D**) or the multigraph distance matrix (**M**):

$$s_i = \sum_j d_{ij} \qquad (2)$$

[⊗] Abstract published in *Advance ACS Abstracts*, December 15, 1994.

0095-2338/95/1635-0366$09.00/0

Table 1. Normal Boiling Points (°C) for 139 Hydrocarbons

no.	compound	bp	no.	compound	bp
1	*n*-propane	−42.1	71	2,2,3,4-tetramethylpentane	133.0
2	n-butane	−0.5	72	2,2,4,4-tetramethylpentane	122.3
3	2-methylpropane	−11.7	73	2,3,3,4-tetramethylpentane	141.6
4	*n*-pentane	36.1	74	benzene	80.1
5	2-methylbutane	27.8	75	toluene	110.6
6	2,2-dimethylpropane	9.5	76	ethylbenzene	136.2
7	*n*-hexane	68.7	77	*o*-xylene	144.4
8	2-methylpentane	60.3	78	*m*-xylene	139.1
9	3-methylpentane	63.3	79	*p*-xylene	138.4
10	2,2-dimethylbutane	0.7	80	*n*-propylbenzene	159.2
11	2,3-dimethylbutane	58.0	81	1-methyl-2-ethylbenzene	165.2
12	*n*-heptane	98.4	82	1-methyl-3-ethylbenzene	161.3
13	2-methylhexane	90.0	83	1-methyl-4-ethylbenzene	162.0
14	3-methylhexane	91.8	84	1,2,3-trimethylbenzene	176.1
15	3-ethylpentane	93.5	85	1,2,4-trimethylbenzene	169.4
16	2,2-dimethylpentane	79.2	86	1,3,5-trimethylbenzene	164.7
17	2,3-dimethylpentane	89.8	87	*n*-butylbenzene	183.3
18	2,4-dimethylpentane	80.5	88	1,2-diethylbenzene	183.4
19	3,3-dimethylpentane	86.1	89	1,3-diethylbenzene	181.1
20	2,2,3-trimethylbutane	80.9	90	1,4-diethylbenzene	183.8
21	*n*-octane	125.7	91	1-methyl-2-*n*-propylbenzene	184.8
22	2-methylheptane	117.6	92	1-methyl-3-*n*-propylbenzene	181.8
23	3-methylheptane	118.9	93	1-methyl-4-*n*-propylbenzene	183.8
24	4-methylheptane	117.7	94	1,2-dimethyl-3-ethylbenzene	193.9
25	3-ethylhexane	118.5	95	1,2-dimethyl-4-ethylbenzene	189.8
26	2,2-dimethylhexane	106.8	96	1,3-dimethyl-2-ethylbenzene	190.0
27	2,3-dimethylhexane	115.6	97	1,3-dimethyl-4-ethylbenzene	188.4
28	2,4-dimethylhexane	109.4	98	1,3-dimethyl-5-ethylbenzene	183.8
29	2,5-dimethylhexane	109.1	99	1,4-dimethyl-2-ethylbenzene	186.9
30	3,3-dimethylhexane	112.0	100	1,2,3,4-tetramethylbenzene	205.0
31	3,4-dimethylhexane	117.7	101	1,2,3,5-tetramethylbenzene	198.2
32	2-methyl-3-ethylpentane	115.6	102	1,2,4,5-tetramethylbenzene	196.8
33	3-methyl-3-ethylpentane	118.3	103	naphthalene	218.0
34	2,2,3-trimethylpentane	109.8	104	acenaphthalene	270.0
35	2,2,4-trimethylpentane	99.2	105	acenaphthene	279.0
36	2,3,3-trimethylpentane	114.8	106	fluorene	294.0
37	2,3,4-trimethylpentane	113.5	107	phenanthrene	338.0
38	2,2,3,3-tetramethylbutane	106.5	108	anthracene	340.0
39	*n*-nonane	150.8	109	4*H*-cyclopenta[*def*]phenanthrene	359.0
40	2-methyloctane	143.3	110	fluoranthene	383.0
41	3-methyloctane	144.2	111	pyrene	393.0
42	4-methyloctane	142.5	112	benzo[*a*]fluorene	403.0
43	3-ethylheptane	143.0	113	benzo[*b*]fluorene	398.0
44	4-ethylheptane	141.2	114	benzo[*c*]fluorene	406.0
45	2,2-dimethylheptane	132.7	115	benzo[*ghi*]fluoranthene	422.0
46	2,3-dimethylheptane	140.5	116	cyclopenta[*cd*]pyrene	439.0
47	2,4-dimethylheptane	133.5	117	chrysene	431.0
48	2,5-dimethylheptane	136.0	118	benz[*a*]anthracene	425.0
49	2,6-dimethylheptane	135.2	119	triphenylene	429.0
50	3,3-dimethylheptane	137.3	120	napthacene	440.0
51	3,4-dimethylheptane	140.6	121	benzo[*b*]fluoranthene	481.0
52	3,5-dimethylheptane	136.0	122	benzo[*j*]fluoranthene	480.0
53	4,4-dimethylheptane	135.2	123	benzo[*k*]fluoranthene	481.0
54	2-methyl-3-ethylhexane	138.0	124	benzo[*a*]pyrene	496.0
55	2-methyl-4-ethylhexane	133.8	125	benzo[*e*]pyrene	493.0
56	3-methyl-3-ethylhexane	140.6	126	perylene	497.0
57	3-methyl-4-ethylhexane	140.4	127	anthanthrene	547.0
58	2,2,3-trimethylhexane	133.6	128	benzo[*ghi*]perylene	542.0
59	2,2,4-trimethylhexane	126.5	129	indeno[1,2,3-*cd*]fluoranthene	531.0
60	2,2,5-trimethylhexane	124.1	130	indeno[1,2,3-*cd*]pyrene	534.0
61	2,3,3-trimethylhexane	137.7	131	dibenz[*a,c*]anthracene	535.0
62	2,3,4-trimethylhexane	139.0	132	dibenz[*a,h*]anthracene	535.0
63	2,3,5-trimethylhexane	131.3	133	dibenz[*a,j*]anthracene	531.0
64	2,4,4-trimethylhexane	130.6	134	picene	519.0
65	3,3,4-trimethylhexane	140.5	135	coronene	590.0
66	3,3-diethylpentane	146.2	136	dibenzo[*a,e*]pyrene	592.0
67	2,2-dimethyl-3-ethylpentane	133.8	137	dibenzo[*a,h*]pyrene	596.0
68	2,3-dimethyl-3-ethylpentane	142.0	138	dibenzo[*a,i*]pyrene	594.0
69	2,4-dimethyl-3-ethylpentane	136.7	139	dibenzo[*a,l*]pyrene	595.0
70	2,2,3,3-tetramethylpentane	140.3			

For distance matrix D, each matrix element d_{ij} represents the distance from vertex i to vertex j. The diagonal entries are all zero, and the distance between any two adjacent vertices would be one. All other entries in this matrix would be the number of edges or bonds traversed in the shortest path from i to j. For the multigraph distance matrix **M**, for

Table 2. Mutagenicity (*S. typhimurium* TA98 with Metabolic Activation) of 95 Aromatic and Heteroaromatic Amines

no.	compound	log rev./nmol	no.	compound	log rev./nmol
1	2-bromo-7-aminofluorene	2.62	49	2,6-dichloro-1,4-phenylenediamine	−0.69
2	2-methoxy-5-methylaniline	−2.05	50	2-amino-7-acetamidofluorene	1.18
3	5-aminoquinoline	−2.00	51	2,8-diaminophenazine	1.12
4	4-ethoxyaniline	−2.30	52	6-aminoquinoline	−2.67
5	1-aminonaphthalene	−0.60	53	4-methoxy-2-methylaniline	−3.00
6	4-aminofluorene	1.13	54	3-amino-2′-nitrobiphenyl	−1.30
7	2-aminoanthracene	2.62	55	2,4′-diaminobiphenyl	−0.92
8	7-aminofluoranthene	2.88	56	1,6-diaminophenazine	0.20
9	8-aminoquinoline	-1.14	57	4-aminophenyldisulfide	−1.03
10	1,7-diaminophenazine	0.75	58	2-bromo-4,6-dinitroaniline	−0.54
11	2-aminonaphthalene	-0.67	59	2,4-diamino-*n*-butylbenzene	−2.70
12	4-aminopyrene	3.16	60	4-aminophenylether	−1.14
13	3-amino-3′-nitrobiphenyl	−0.55	61	2-aminobiphenyl	−1.49
14	2,4,5-trimethylaniline	−1.32	62	1,9-diaminophenazine	0.04
15	3-aminofluorene	0.89	63	1-aminofluorene	0.43
16	3,3′-dichlorobenzidine	0.81	64	8-aminofluoranthene	3.80
17	2,4-dimethylaniline	−2.22	65	2-chloroaniline	−3.00
18	2,7-diaminofluorene	0.48	66	3-amino-α,α,α-trifluorotoluene	−0.80
19	3-aminofluoranthene	3.31	67	2-amino-1-nitronaphthalene	−1.17
20	2-aminofluorene	1.93	68	3-amino-4′-nitrobiphenyl	0.69
21	2-amino-4′-nitrobiphenyl	−0.62	69	4-bromoaniline	−2.70
22	4-aminobiphenyl	−0.14	70	2-amino-4-chlorophenol	−3.00
23	3-methoxy-4-methylaniline	−1.96	71	3,3′-dimethoxybenzidine	0.15
24	2-aminocarbazole	0.60	72	4-cyclohexylaniline	−1.24
25	2-amino-5-nitrophenol	−2.52	73	4-phenoxyaniline	0.38
26	2,2′-diaminobiphenyl	−1.52	74	4,4′-methylenebis(*o*-ethylaniline)	−0.99
27	2-hydroxy-7-aminofluorene	0.41	75	2-amino-7-nitrofluorene	3.00
28	1-aminophenanthrene	2.38	76	benzidine	−0.39
29	2,5-dimethylaniline	−2.40	77	1-amino-4-nitronaphthalene	−1.77
30	4-amino-2′-nitrobiphenyl	−0.92	78	4-amino-3′-nitrobiphenyl	1.02
31	2-amino-4-methylphenol	−2.10	79	4-amino-4′-nitrobiphenyl	1.04
32	2-aminophenazine	0.55	80	1-aminophenazine	−0.01
33	4-aminophenylsulfide	0.31	81	4,4′-methylenebis(*o*-fluoroaniline)	0.23
34	2,4-dinitroaniline	−2.00	82	4-chloro-2-nitroaniline	−2.22
35	2,4-diaminoisopropylbenzene	−3.00	83	3-aminoquinoline	−3.14
36	2,4-difluoroaniline	−2.70	84	3-aminocarbazole	−0.48
37	4,4′-methylenedianiline	−1.60	85	4-chloro-1,2-phenylenediamine	−0.49
38	3,3′-dimethylbenzidine	0.01	86	3-aminophenanthrene	3.77
39	2-aminofluoranthene	3.23	87	3,4′-diaminobiphenyl	0.20
40	2-amino-3′-nitrobiphenyl	−0.89	88	1-aminoanthracene	1.18
41	1-aminofluoranthene	3.35	89	1-aminocarbazole	−1.04
42	4,4′-ethylenebis(aniline)	−2.15	90	9-aminoanthracene	0.87
43	4-chloroaniline	−2.52	91	4-aminocarbazole	−1.42
44	2-aminophenanthrene	2.46	92	6-aminochrysene	1.83
45	4-fluoroaniline	−3.32	93	1-aminopyrene	1.43
46	9-aminophenanthrene	2.98	94	4,4′-methylenebis(*o*-isopropylaniline)	−1.77
47	3,3′-diaminobiphenyl	−1.30	95	2,7-diaminophenazine	3.97
48	2-aminopyrene	3.50			

each entry d_{ij}, $1/b$ is used, where b is the conventional bond order (b = 1, 2, 3, and 1.5 for single, double, triple, and aromatic bonds, respectively). To account for periodicity of chemical properties for heteroatoms, Balaban proposed two J variants[21]: J^X which includes corrections for heteroatom electronegativities and J^Y which has corrections for heteroatom covalent radii.

Table 3 lists the 102 TIs used in this paper. The set of TIs were generated for all compounds in each data set. Several of these indices were completely correlated ($r = 1.0$) and were dropped. Coincidentally, this left 90 TIs for use in the analyses for both data sets. The indices were calculated by POLLY.[22] The set of atom pairs was generated by APPROBE.[23]

2.3. Data Reduction. Initially, all TIs were transformed by the natural log of the index plus one. This was done since the scale of some TIs may be several orders of magnitude greater than other TIs.

Since many of the TIs are highly intercorrelated, the original 90 dimensioned space of these indices can be represented by a subspace without significant loss of information. One such method, available under SAS as the VARCLUS[24] procedure, is to divide the indices into disjoint subsets of variables based on the correlation matrix.

The VARCLUS procedure divides the set of TIs into disjoint clusters so that each cluster is essentially unidimensional. Initially, all TIs start in one cluster, and the first and second principal components (PCs) are determined from the correlation matrix. Then the cluster is split by assigning all TIs most correlated with the first PC to one cluster and all TIs most correlated with the second PC to another cluster. The first and second PCs for each new cluster are determined after the original cluster is split. Each new cluster is split into two new clusters until each cluster has only one significant PC, i.e., the eigenvalue for the second PC of subset of clustered TIs is less than one. The output of this procedure is the first principal component (PC) from each cluster of TIs. These PCs were subsequently used in the similarity measures described below.

Table 3. Topological Index Symbols and Definitions

I^{W}_{D}	information index for the magnitudes of distances between all possible pairs of vertices of a graph
$\bar{I}^{W}_{D}$	mean information index for the magnitude of distance
W	Wiener index = half-sum of the off-diagonal elements of the distance matrix of a graph
I^{D}	degree complexity
H^{V}	graph vertex complexity
H^{D}	graph distance complexity
$\overline{IC}$	information content of the distance matrix partitioned by frequency of occurrences of distance h
O	order of neighborhood when IC_r reaches its maximum value for the hydrogen-filled graph
I_{ORB}	information content or complexity of the hydrogen-suppressed graph at its maximum neighborhood of vertices
O_{ORB}	maximum order of neighborhood of vertices for I_{ORB} within the hydrogen-suppressed graph
M_1	a Zagreb group parameter = sum of square of degree over all vertices
M_2	a Zagreb group parameter = sum of cross-product of degrees over all neighboring (connected) vertices
IC_r	mean information content or complexity of a graph based on the rth ($r = 0-6$) order neighborhood of vertices in a hydrogen-filled graph
SIC_r	structural information content for rth ($r = 0-6$) order neighborhood of vertices in a hydrogen-filled graph
CIC_r	complementary information content for rth ($r = 0-6$) order neighborhood of vertices in a hydrogen-filled graph
$h\chi$	path connectivity index of order $h = 0-6$
$h\chi_C$	cluster connectivity index of order $h = 3-6$
$h\chi_{Ch}$	chain connectivity index of order $h = 3-6$
$h\chi_{PC}$	path-cluster connectivity index of order $h = 4-6$
$h\chi^{b}$	bonding path connectivity index of order $h = 0-6$
$h\chi^{b}_{C}$	χ^{b}_{C}bonding cluster connectivity index of order $h = 3-6$
$h\chi^{b}_{Ch}$	χ^{b}_{Ch}bonding chain connectivity index of order $h = 3-6$
$h\chi^{b}_{PC}$	χ^{b}_{PC}bonding path-cluster connectivity index of order $h = 4-6$
$h\chi^{v}$	valence path connectivity index of order $h = 0-6$
$h\chi^{v}_{C}$	χ^{v}_{C}valence cluster connectivity index of order $h = 3-6$
$h\chi^{v}_{Ch}$	χ^{v}_{Ch}valence chain connectivity index of order $h = 3-6$
$h\chi^{v}_{PC}$	χ^{v}_{PC}valence path-cluster connectivity index of order $h = 4-6$
P_h	number of paths of length $h = 0-10$
J	Balaban's J index based on distance
J^{B}	Balaban's J index based on multigraph bond orders
J^{X}	Balaban's J index based on relative electronegativities
J^{Y}	Balaban's J index based on relative covalent radii

In a standard principal component analysis, each of the orthogonal axes derived from the set of variables is composed of linear combinations of all the original variables. This can make interpretation of the new axes difficult. By using disjoint clusters of variables, axis interpretation can be enhanced, although orthogonality can no longer be guaranteed.

In addition to using the cluster or PC variables derived above, we selected from each cluster the TI which was most correlated with the cluster to which it belonged. These TIs were then used in the similarity measures as well.

2.4. Similarity Measures. Five measures of intermolecular similarity were used in these studies. The first similarity measure was an associative coefficient using the atom pairs. Similarity (S) between two molecules i and j is defined as[19]

$$S_{ij} = 2C/(T_i + T_j) \tag{3}$$

where C is the number of atom pairs common to molecule i and j. T_i and T_j are the total number of atom pairs in i and j, respectively. The numerator is multiplied by two to account for the atom pairs belong to both molecules.

The remaining four similarity measures are based on Euclidean distance (ED) within an n-dimensional space. ED between molecules i and j is defined as

$$ED_{ij} = [\sum_{k=1}^{n}(D_{ik} - D_{jk})^2]^{1/2} \tag{4}$$

where n equals the number of dimensions and D_{ik} equals the datavalue of the kth dimension for compound i.

The dimensions are the clusters (PCs) or TIs as selected by the procedure outlined in section 2.3. The four ED based approaches included (a) TI_s, scaled TIs, (b) TI_u, unscaled TIs, (c) PC_s, scaled cluster variables, and (d) PC_u, unscaled cluster variables. The two methods using scaled dimensions involved rescaling the variables to have mean zero and variance one.

2.5. *K* Neighbor Selection and Property Estimation. Using the similarity methods described in section 2.4, intermolecular similarity of the chemicals within each data set was quantified. For each chemical, the K nearest neighbors were determined using each of the similarity techniques, where K was 1−10, 15, 20, and 25.

For the hydrocarbons, the mean observed boiling point of the K nearest neighbors for a compound was used as the estimated boiling point for the compound. The correlation (r) of observed boiling point with estimated boiling point and the standard error of the estimates were used to assess the relative efficacy of the five similarity methods.

Since the probe compound is essentially removed from the database when selecting analogs, i.e., the probe is not a candidate neighbor to itself, the estimation of the property is the same as if the probe were a complete unknown.

The analog selection and property prediction method used for the set of hydrocarbons was used to estimate the mutagenic activity, ln(R), of the aromatic and heteroaromatic amines.

3. RESULTS

3.1. Variable Clustering and TI Selection. There were 10 significant clusters derived from the variable clustering of 90 TIs for the 139 hydrocarbons. These 10 clusters explained a total of 90.5% of the total variation within the original TIs.

Table 4. Topological Indices Selected for Similarity Measures and the Correlation of the TI with the Cluster from which It Was Selected

	hydrocarbons		aromatic amines	
Cluster	TI	r	TI	r
1	P_0	0.991	$^4\chi$	0.985
2	IC_6	0.984	SIC_4	0.984
3	$^5\chi^v_{PC}$	0.922	$^4\chi^b_{PC}$	0.926
4	$^5\chi^v_{Ch}$	0.998	CIC_1	0.967
5	$^5\chi^v_C$	0.986	P_0	0.990
6	$^6\chi$	0.987	$^5\chi_{Ch}$	0.940
7	$^3\chi^v_C$	0.986	IC_4	0.987
8	CIC_1	0.956	J^B	0.954
9	SIC_3	0.951	SIC_2	0.896
10	SIC_1	0.916	O	0.944

The TIs selected for use in the TI_S and TI_U similarity methods are listed in Table 4. Shown in Table 4 are the TI selected and the correlation of the TI with the cluster from which it was selected. Each of these TIs were the variable most correlated with the cluster from which it was selected.

For the 95 amine substituted aromatic and heteroaromatic chemicals, there were 10 significant clusters derived from the 90 TIs. These clusters explained 90.5% of the total variation within the original TIs.

As for the hydrocarbons, Table 4 shows the TI's selected for use in the TI_S and TI_U similarity methods.

3.2. *K*-Neighbor Property Estimation. Figure 1 presents the correlation (r) and standard error of prediction (SE) for boiling points of the hydrocarbons for the K values examined (1−10, 15, 20,25). Each line in the figure represents one of the five similarity methods. Table 5 shows the best boiling point model obtained for each similarity method. Shown in the table are the K value used for the similarity method, the correlation of the mean boiling point of the K neighbors with the observed boiling points, and the residual standard deviation (SE) seen with the predictive model. As can be seen from Figure 1 and Table 5, best boiling point estimates for each method occur when K is equal to two or three. The range of K from two to approximately five appear to give the best boiling point estimates.

Figure 2 presents the correlation and SE of prediction for mutagenic potency, ln(R), where R is the number of revertants per nanomole. Table 6 shows the best potency model obtained for each similarity method. The best mutagenic rate estimates were for K in the range of two to five for the four ED based methods. For the AP method, the best estimates were from K from five to eight with the best mutagenic rate estimates being at K equal to five.

For both data sets, using ED with scaled TIs gave superior results and the AP method slightly inferior results. These differences, however, were only slight.

4. DISCUSSION

The major goals of this study were to investigate the effectiveness of similarity measures based on graph theoretic parameters in estimating properties and to explore the effect of increasing the number of analogs used for property estimation on the efficacies of the various methods. To this end, we estimated normal boiling points of a set of 139 hydrocarbons and mutagenic potencies for a group of 95 aromatic and heteroaromatic amines. Testing these methods on diverse properties gives us a better assessment of their general applicability. Furthermore, each of these data sets has published models upon which we could draw comparisons.

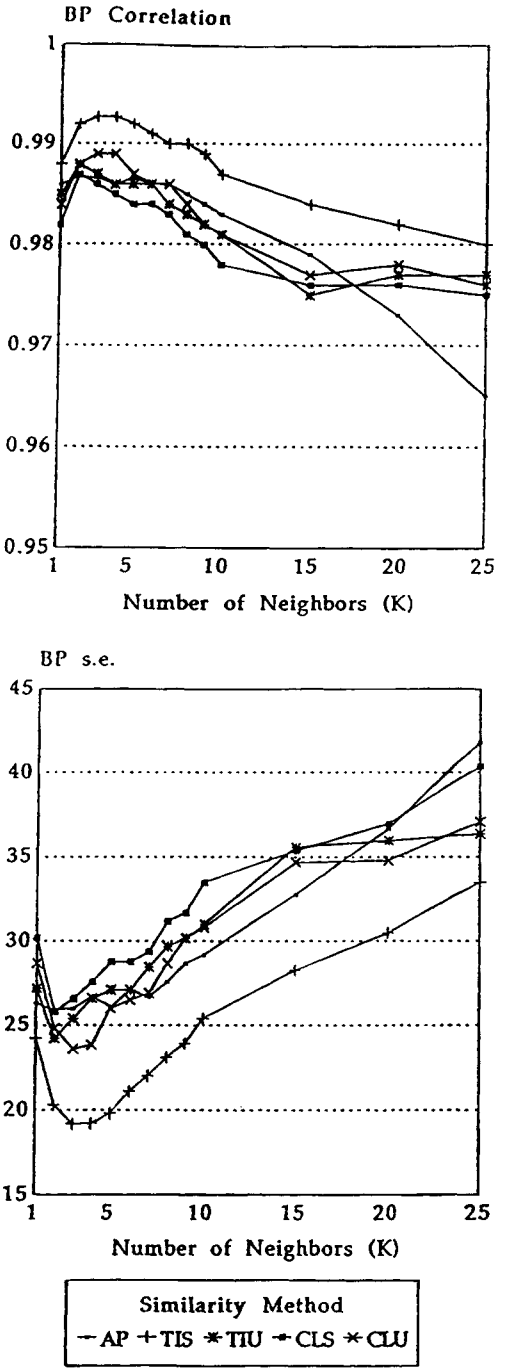

Figure 1. Pattern of correlation (r) and standard error of the estimates (SE) according to K nearest neighbor selection for 139 hydrocarbon boiling points.

Table 5. Comparison of the Five Similarity Methods for Prediction of Boiling Point for 139 Hydrocarbons

similarity method	K	r	SE
AP	2	0.987	25.9
TI_S	3	0.993	19.2
TI_U	2	0.988	24.2
PC_S	2	0.987	25.8
PC_U	3	0.989	23.6

The results in Tables 5 and 6 show that for both data sets, all five similarity methods give reasonable estimates of the properties. In Basak *et al.*,[25] a principal component regres-

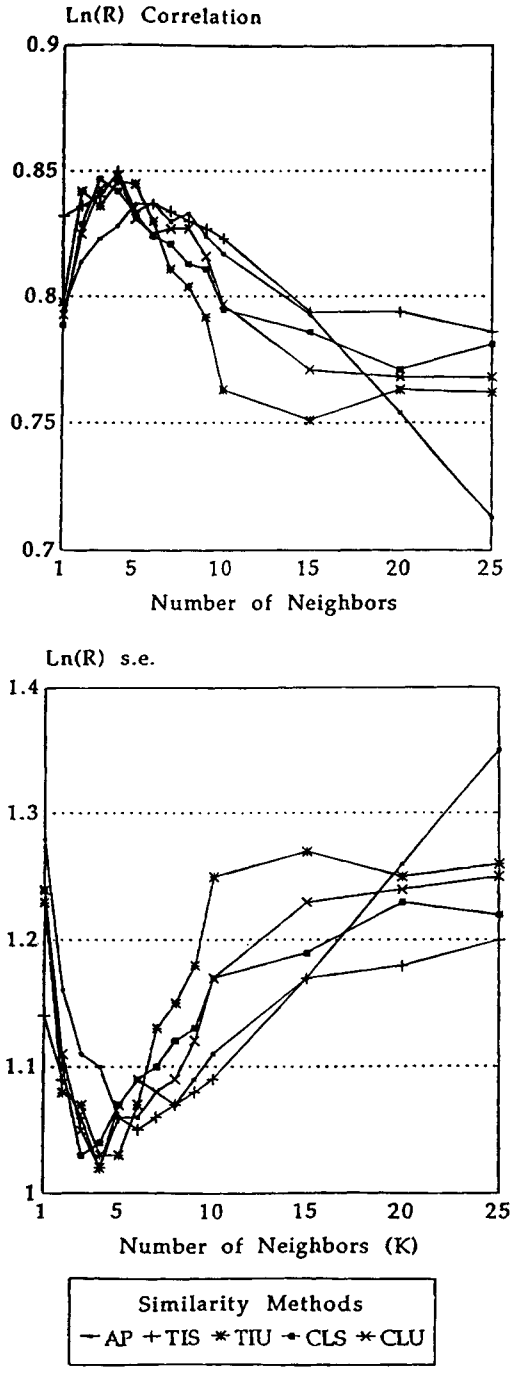

Figure 2. Pattern of correlation (r) and standard error of the estimates (SE) according to K nearest neighbor selection for 95 aromatic amine mutagenic rates [ln(R)].

Table 6. Comparison of the Five Similarity Methods for Prediction of Mutagenic Potency, ln(R), for 95 Aromatic and Heteroaromatic Amines

similarity method	K	r	SE
AP	5	0.837	1.06
TI_S	4	0.850	1.02
TI_U	4	0.846	1.03
PC_S	3	0.847	1.03
PC_U	4	0.848	1.02

sion (PCR) model derived from topological indices for this set of hydrocarbons had a correlation coefficient of 0.989 with a SE of 23.9 °C. The results seen here are comparable, with the correlation ranging from 0.987 to 0.993 and SE ranging from 19.2° to 25.9° (see Table 5). The somewhat high standard errors seen with these compounds can be attributed to the structural diversity between alkanes, alkyl benzenes, and polycyclic aromatic hydrocarbons.[25]

In the original study of the aromatic amines by Debnath *et al.*,[18] four physicochemical parameters were used to model mutagenicity. The model had a correlation coefficient of 0.898 and a standard error of 0.86. In Table 6, the correlation ranges from 0.837 for the AP method to 0.850 for the TI_S method. The standard errors for these methods were 1.06 down to 1.02, respectively.

The most interesting result was the effect of changing the number of analogs used in the estimation of properties. In all cases, a relatively small number of analogs (2−5) give best estimates of normal boiling point and mutagenicity. In a recent paper, Basak and Grunwald[14] used the nearest neighbor (K = 1) to estimate molecular properties. The results of this paper indicate higher numbers of neighbors represent the neighborhoods of property space better as compared to the nearest neighbor.

The same methods of estimation using principles of similarity were used for two diverse properties, vis., boiling points of hydrocarbons and mutagenicity of aromatic amines. We were interested in seeing how each of the methods performs as a generalized method. Further studies with other properties and larger, diverse data sets are needed to assess the general applicability of the methods and to test the effect of increasing the neighborhood space on the accuracy of estimated properties.

ACKNOWLEDGMENT

The authors are delighted to dedicate this paper to Professor Alexandru T. Balaban in appreciation of his pioneering work in chemical graph theory. In addition, the authors are grateful to Sharon Bertelsen for technical support. This is contribution number 136 from the Center for Water and the Environment of the Natural Resources Research Institute. Research reported in this paper was supported, in part, by cooperative agreement CR 819621 from the United Stated Environmental Protection Agency and by Grant F49620-94-1-0401 from the United States Air Force.

REFERENCES AND NOTES

(1) Johnson, M. A.; Maggiora, G. M. *Concepts and Applications of Molecular Similarity*; Wiley: New York, 1990.
(2) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom Pairs as Molecular Features in Structure-Activity Studies: Definition and Applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64−73.
(3) Johnson, M. A.; Basak, S. C.; Maggiora, G. A Characterization of Molecular Similarity Methods for Property Prediction. *Mathl. Comput. Modelling* **1988**, *11*, 630−634.
(4) Basak, S. C.; Magnuson, V. R.; Niemi, G. J.; Regal, R. R. Determining Structural Similarity of Chemicals Using Graph-Theoretic Indices. *Discrete Appl. Math.* **1988**, *19*, 17−44.
(5) Basak, S. C.; Bertelsen, S.; Grunwald, G. D. Application of Graph Theoretical Parameters in Quantifying Molecular Similarity and Structure−Activity Studies. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 270−276.
(6) Basak, S. C.; Grunwald, G. D. Use of Topological Space and Property Space in Selecting Structural Analogs. *Math. Modelling Sci. Comput.*, in press.
(7) Willet, P.; Winterman, V. A Comparison of Some Measures for the Determination of Intermolecular Structural Similarity: Measures of Intermolecular Structural Similarity, *Quant. Struct-Act. Relat.* **1986**, *5*, 18−25.
(8) Fisanick, W.; Cross, K. P.; Rusinko, III, A. Similarity Searching on CAS Registry Substances 1. Molecular Property and Generic Atom Triangle Geometric Searching. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 664−674.
(9) Lajiness, M. S. In *Computational Chemical Graph Theory*; Rouvray, D. H., Ed.; Nova: New York, 1990, pp 299−316.

(10) Wilkins, C. L.; Randić, M. A Graph Theoretic Approach to Structure-Property and Structure−Activity Correlations. *Theor. Chim. Acta (Berl.)* **1980**, *58*, 45−68.
(11) Auer, C. M.; Nabholz, J. V.; Baetcke, K. P. Mode of Action and the Assessment of Chemical Hazards in the Presence of Limited Data: Use of Structure-Activity Relationships (SAR) Under TSCA, Section 5. *Environ. Health Perspect.* **1990**, *87*, 183−197.
(12) Basak, S. C.; Grunwald, G. D. Estimation of Lipophilicity from Structural Similarity. *New J. Chem.*, in press.
(13) Austel, V. In *Topics in Current Chemistry*; Charton, M., Motoc, I., Eds.; Springer-Verlag: Berlin, 1983; Vol. 114, pp 7−19.
(14) Basak, S. C.; Grunwald, G. D. Molecular Similarity and Risk Assessment: Analog Selection and Property Estimation Using Graph Invariants. *SAR QSAR Environmental Res.*, in press.
(15) Needham, D. E.; Wei, I. C.; Seybold, P. G. Molecular Modelling of the Physical Properties of Alkanes. *J. Am. Chem. Soc.* **1988**, *110*, 4186−4194.
(16) Mekenyan, O.; Bonchev, D.; Trinajstic, N. Chemical Graph Theory: Modelling the Thermodynamic Properties of Molecules. *Int. J. Quantum Chem.* **1980**, *18*, 369−380.
(17) Karcher, W. *Spectral Atlas of Polycyclic Aromatic Hydrocarbons, Vol. 2*. Kluwer Academic: Dordrecht/ Boston/London, 1988; pp 16−19.
(18) Debnath, A. K.; Debnath, G.; Shusterman, A. J.; Hansch, C. A QSAR Investigation of the Role of Hydrophobicity in Regulating Mutagenicity in the Ames Test: 1. Mutagenicity of Aromatic and Heteroaromatic Amines in *Salmonella typhimurium* TA98 and TA100. *Environ. Mol. Mutagen.* **1992**, *19*, 37−52.
(19) Balaban, A. T. Highly Discriminating Distance-Based Topological Index. *Chem. Phys. Lett.* **1982**, *89*, 399−404.
(20) Balaban, A. T. Topological Indices Based on Topological Distances in Molecular Graphs. *Pure Appl. Chem.* **1983**, *55*, 199−206.
(21) Balaban, A. T. Chemical Graphs. Part 48. Topological Index J for Heteroatom-Containing Molecules Taking into Account Periodicities of Element Properties. *MATCH* **1986**, *21*, 115−122.
(22) Basak, S. C.; Harriss, D. K.; Magnuson, V. R. POLLY: Copyright of the University of Minnesota, 1988.
(23) Basak, S. C.; Grunwald, G. D. APPROBE: Copyright of the University of Minnesota, 1994.
(24) SAS Institute Inc. In *SAS/STAT User's Guide, Release 6.03 Edition*. SAS Institute Inc.: Cary, NC, 1988; Chapter 34, pp 949−965.
(25) Basak, S. C.; Niemi, G. J.; Veith, G. D. Predicting Properties of Molecules Using Graph Invariants. *J. Math. Chem.* **1991**, *7*, 243−272.

CI940103Z