Characterization of Isospectral Graphs Using Graph Invariants and Derived Orthogonal Parameters

Krishnan Balasubramanian[†] and Subhash C. Basak*,[‡]

Department of Chemistry, Arizona State University, Tempe, Arizona 85287-1604, and
Natural Resources Research Institute, University of Minnesota, Duluth, Duluth, Minnesota 55811

Received July 1, 1997

Numerical graph theoretic invariants or topological indices (TIs) and principal components (PCs) derived from TIs have been used in discriminating a set of isospectral graphs. Results show that lower order connectivity and information theoretic TIs suffer from a high degree of redundancy, whereas higher order indices can characterize the graphs reasonably well. On the other hand, PCs derived from the TIs had no redundancy for the set of isospectral graphs studied.

1. INTRODUCTION

Graph theoretical and topological techniques have been harnessed in numerous practical applications in recent years. In particular, the use of graph theoretical techniques for the characterization of structures and for the exploration of structure−property relations have received considerable attention.[1−24] The intimate relation between the structure of a molecule and its activity has been the topic of exploration for many years. Several novel techniques based primarily on graph theory and topology have been proposed for predicting activities from the structure, and such techniques have been successfully applied to molecules of pharmacological relevance.

Since graph theoretical techniques are based on the topological connectivity of a molecule rather than its three-dimensional molecular structure, there is always a question as to the suitability of a graph theoretically based technique for the characterization or prediction of properties that may depend on more complex factors than simple connectivity. For this reason techniques based on the three-dimensional molecular geometry have been proposed.[21−23]

A recognized problem with graph-theoretically based technique is in dealing with graphs called isospectral graphs.[24−27] Isospectral graphs are graphs with the same characteristic polynomial which is simply the secular determinant of the adjacency matrix of a graph. Thus isospectral graphs would have the same graph eigenvalues or spectra, which could be visualized as the Huckel energy levels associated with the molecule corresponding to the graphs under consideration. The isospectral graphs have thus received much attention due to their "pathological" nature. Prior to the discovery of isospectral graphs it was surmised that the characteristic polynomials or spectra might uniquely characterize graphs, but examples of isospectral graphs revealed that there are pairs of nonisomorphic graphs which are topologically distinct and yet they have the same characteristic polynomials and spectra. As a result of this isospectral graphs pose several problems. As discussed in the work of Liu et *al.*,[24] some of the vertex partitioning algorithms fail for isospectral graphs. Likewise, the topologically based indices such as the Wiener index[3] become identical for isospectral graphs.

Basak et al.[28] used a combination of graph invariants to characterize a large collection of complex graphs. The principal component analysis (PCA) which is performed on the basis of these indices and the Euclidian distance method have provided a promising avenue for the characterization of structures and structure−activity relationships. Thus, it is interesting to explore if these techniques are satisfactory for isospectral graphs which are considered to be pathological in a graph theoretical sense. The objective of this study is to consider a series of isospectral graphs for the purpose of computing these indices and the PCA on those indices. We show that while lower-order indices often fail to discriminate isospectral graphs, the PCs derived from indices discriminate all isospectral graphs considered here.

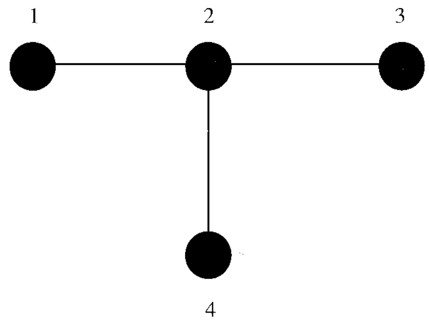
2. CALCULATION OF GRAPH THEORETICAL PARAMETERS

The calculation of the topological indices (TIs) used in this study has previously been described in detail.[1] The TIs for the isospectral pairs of graphs were calculated by POLLY.[2] The POLLY 2.3 version is capable of calculating 97 TIs from the SMILES line notation input of chemical structures. The TIs calculated by POLLY 2.3 include the Wiener index,[3] connectivity indices,[4,5] and information theoretic indices defined on distance matrices of graphs[6,7] as well as a set of parameters derived on the neighborhood complexity of vertices in hydrogen-filled molecular graphs.[8−11] We describe below the methods for the calculation of the TIs used in this paper.

The Wiener index W,[3] the first topological index reported in the chemical literature, may be calculated from the distance matrix $\mathbf{D}(\mathbf{G})$ of a hydrogen-suppressed chemical graph G as the sum of the entries in the upper triangular distance submatrix. The distance matrix $\mathbf{D}(\mathbf{G})$ of a nondirected graph $\mathbf{G}$ with n vertices is a real symmetric $n \times n$ matrix with

* Corresponding author.
[†] Arizona State University.
[‡] University of Minnesota.

S0095-2338(97)00052-8 CCC: $15.00
Published on Web 04/21/1998



G_1

Figure 1. Hydrogen suppressed graph of isobutane.

elements d_{ij} equal to the distance between vertices v_i and v_j in **G**. Each diagonal element d_{ii} of **D**(**G**) is zero. We give below the distance matrix **D**($\mathbf{G}_1$) of the unlabeled hydrogen-suppressed graph $\mathbf{G}_1$ of isobutane (Figure 1):

$$D(\mathbf{G}_1) = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} (1) & (2) & (3) & (4) \end{array} \\ \begin{bmatrix} 0 & 1 & 2 & 2 \\ 1 & 0 & 1 & 1 \\ 2 & 1 & 0 & 2 \\ 2 & 1 & 2 & 0 \end{bmatrix} \end{array}$$

W is calculated as

$$W = {}^1/_2 \sum_{ij} d_{ij} = \sum_h h \cdot g_h \qquad (1)$$

where g_h is the number of unordered pairs of vertices whose distance is h.

Randić's[4] connectivity index as well as the higher-order path, cluster, and path-cluster types of simple and valence connectivity indices developed by Kier and Hall[5] were calculated by the computer program POLLY.[2] P_h parameters, the number of paths of length h ($h = 0-10$) in the hydrogen-suppressed graph, are calculated using standard algorithms.

Information-theoretic topological indices are calculated by the application of information theory to chemical graphs. An appropriate set A of n elements is derived from a molecular graph **G** depending upon certain structural characteristics. On the basis of an equivalence relation defined on A, the set A is partitioned into disjoint subsets A_i of order n_i ($i = 1,2, ... h$; $\sum n_i = n$). A probability distribution is then assigned to the set of equivalence classes

$$A_1, A_2, ..., A_h$$

$$p_1, p_2, ..., p_h$$

where $p_i = n_i/n$ is the probability that a randomly selected element of A will occur in the ith subset.

The mean information content of an element of A is defined by Shannon's[12] relation

$$\mathrm{IC} = -\sum_{i=1}^{h} p_i \log_2 p_i \qquad (2)$$

The logarithm is taken at base 2 for measuring the information content in bits. The total information content of the set A is then n times IC.

Rashevsky[13] was the first to calculate the information content of graphs where "topologically equivalent" vertices are placed in the same equivalence class. In Rashevsky's approach, two vertices u and v of a graph are said to be topologically equivalent if and only if for each neighboring vertex u_i ($i = 1, 2, ..., k$) of the vertex u, there is a distinct neighboring vertex v_i of the same degree for the vertex v. Subsequently, Trucco[14] defined topological information of graphs on the basis of graph orbits. In this method, vertices which belong to the same orbit of the automorphism group are considered topologically equivalent. While Rashevsky[13] used simple linear graphs with indistinguishable vertices to symbolize molecular structure, weighted linear graphs or multigraphs are better models for conjugated or aromatic molecules because they more properly reflect the actual bonding patterns, *i.e.*, electron distribution.

To account for the chemical nature of vertices as well as their bonding pattern, Sarkar *et al.*[15] calculated the information content of chemical graphs on the basis of an equivalence relation where two atoms of the same element are considered equivalent if they possess an identical first-order topological neighborhood. Since properties of atoms or reaction centers are often modulated by physicochemical characteristics of distant neighbors, *i.e.,* neighbors of neighbors, it was deemed essential to extend this approach to account for higher-order neighbors of vertices. This can be accomplished by defining open spheres for all vertices of a chemical graph. If r is any non-negative real number and v is a vertex of the graph **G**. then the open sphere $S(v,r)$ is defined as the set consisting of all vertices v_i in **G** such that $d(v,vi) < r$. Then, $S(v,0) = \phi$, $S(v,r) = v$ for $0 < r < 1$, and $S(v,r)$ is the set consisting of v and all vertices v_i of **G** situated at unit distance from v for $1 < r < 2$.

One can construct such open spheres for higher integral values of r. For a particular value of r, the collection of all such open spheres $S(v,r)$, where v runs over the whole vertex set V, forms a neighborhood system of the vertices of **G**. A suitably defined equivalence relation can then partition V into disjoint subsets consisting of topological neighborhoods of vertices of up to rth order neighbors. Such an approach has already been initiated, and the information-theoretic indices calculated are called indices of neighborhood symmetry.[10]

In this method, chemical species are symbolized by weighted linear graphs. Two vertices u_o and v_o of a molecular graph are said to be equivalent with respect to the rth order neighborhood if, and only if, corresponding to each path $u_o, u_1, ..., u_r$ of length r, there is a distinct path $v_o, v_1, ..., v_r$ of the same length, such that the paths have similar edge weights, and both u_o and v_o are connected to the same number and type of atoms up to the rth order bonded neighbors. The detailed equivalence relation is described in our earlier studies.

Once partitioning of the vertex set for a particular order of neighborhood is completed, $\mathrm{IC_r}$ is calculated from eq 2. Basak, Roy, and Ghosh[9] defined another information-theoretic measure, structural information content ($\mathrm{SIC_r}$), which is calculated as

$$\mathrm{SIC_r} = \mathrm{IC_r}/\log_2 n \qquad (3)$$

Table 1. Topological Indexes: 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
$\underline{H}^{D}$	graph distance complexity
$\overline{\mathrm{IC}}$	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
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 *r*th ($r = 0-6$) order neighborhood of vertices in a hydrogen-filled graph
SIC_r	structural information content for *r*th ($r = 0-6$) 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 $h = 5-6$
$^{h}\chi_{PC}$	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

where IC_r is calculated from eq 2 and *n* is the total number of vertices of the graph.

Another information-theoretic invariant, complementary information content (CIC_r),[11] is defined as

$$CIC_r = \log_2 n - IC_r \tag{4}$$

CIC_r represents the difference between the maximum possible complexity of a graph (where each vertex belongs to a separate equivalence class) and the realized topological information of a chemical species as defined by IC_r.

The information-theoretic index on graph distance, I_D^W, is calculated from the distance matrix **D**(**G**) of a chemical graph **G** by the method of Bonchev and Trinajstić:[7]

$$I_D^W = W \log_2 W - \sum_h g_h \cdot h \log_2 h \tag{5}$$

The mean information index, $\bar{I}_D^W$ is found by dividing the information index I_D^W by *W*. IC_r, SIC_r, CIC_r, I_D^W, and $\bar{I}_D^W$ were calculated by Polly.[6] The information theoretic parameters defined on the distance matrix, H^D and H^v were calculated by the method of Raychaudhury et *al.* Sixty TIs were calculated for each of the 38 molecular graphs in Figure 2.

3. STATISTICAL ANALYSIS

3.1. Data Reduction. The TIs used in this paper are shown in Table 1. Initially, all TIs were transformed by the natural logarithm of the value of the index plus one. This was done because the scale of some TIs may be several orders of magnitude greater than others.

3.2. Principal Components Analysis. The data for the isospectral graphs analyzed in this paper may be viewed as *n* (number of isospectral graphs) vectors in *p* (number of calculated parameters) dimensions. The data for each set can be represented by a matrix **X** which has *n* rows and *p* columns. For each of the graphs, the number of calculated parameters was 60 (TIs of Table 1). Each graph is therefore represented by a point in R^{60}, where *R* is the field of real numbers. If each graph *s* was represented in R^2, then one could plot and investigate the extent of relationship between individual parameters. In R^{60} such a simple analysis is not possible. However, since many of the TIs are highly intercorrelated, the points in R^{60} can likely be represented by a subspace of fewer dimensions. The method of PCA or the Karhunen-Loeve transformation is a standard method for reduction of dimensionality.[29] The first principal component (PC) is the line which comes closest to the points in the sense of minimizing the sum of the squared Euclidean distances from the points to the line. The second PC is given by projections onto the basis vector orthogonal to the first PC. For points in R^p, the first *r* PCs give the subspace which comes closest to approximating the *n* points. The first PC is the first axis of the points. Successive axes are major directions orthogonal to previous axes. The PCs are the closest approximating hyperplane, and because they are calculated from eigenvectors of a $p \times p$ matrix, the computations are relatively accessible. But there are important scaling choices, because PCs are scale dependent. To control this dependence, the most commonly used convention is to rescale the variables so that each variable has a mean of zero and a standard deviation of one. The covariance matrix for these rescaled variables is the correlation matrix. The PCA on the TIs for isospectral graphs has been carried out using SAS software.[30]

Table 2. Summary of Principal Components Analysis

	eigenvalue	% cumulative varnce explnd		eigenvalue	% cumulative varnce explnd
PC_1	34.1	56.8	PC_4	2.8	93.2
PC_2	14.4	80.8	PC_5	1.3	95.3
PC_3	4.6	88.5	PC_6	1.1	97. 1

4. RESULTS

The summary of PCA using 60 calculated TIs is shown in Table 2. The first three PCs explain nearly 90% of the variance in the data and the first six PCs with eigenvalue greater than 1.0 explain about 97% of the variance in the original data.

In Table 3 we give the values for PC_1-PC_6 for the 38 graphs analyzed in this paper. It is interesting to note that almost all PCs have distinct values for pairs (*e.g.*, 1.1 and 1.2; 2.1 and 2.2, etc.) of isospectral graphs.

Table 4 presents the values of connectivity indices $^0\chi-^2\chi$ and neighborhood complexity indices IC_0-IC_2 for the graphs.

Table 3. First Six PCs for the Set of 38 Isospectral Graphs (Figure 2)

graph	PC_1	PC_2	PC_3	PC_4	PC_5	PC_6
1.1	−10.6828	−1.5214	0.0283	−2.3056	−0.2901	−0.7411
1.2	−11.2419	−0.7289	0.6454	−2.4077	−1.3287	−1.2562
2.1	−7.5623	−2.8765	0.4809	0.4976	−1.4914	1.6824
2.2	−7.6856	1.4163	−1.0238	−0.9141	−0.5908	−0.4823
3.1	1.6223	−1.7614	−4.5762	−0.4737	1.3809	0.1826
3.2	1.4956	−3.7201	−2.6087	0.5261	0.4641	2.0115
4.1.1	−2.1141	0.3656	−2.2068	0.2315	−0.1458	−0.3351
4.1.2	−2.5286	2.2386	−1.0309	−0.4577	−0.5908	−0.4608
4.2.1	−2.5555	−3.2923	3.9820	1.9017	−0.0220	0.5264
4.2.2	−2.4859	0.7047	−0.5478	0.1951	−1.4380	0.5363
5.1	−7.4612	−0.3102	−0.9097	−0.3816	0.8077	0.1601
5.2	−7.7603	0.9300	−0.9975	−1.7106	−0.3964	−1.3015
6.1	−5.8986	−0.5274	−0.4014	−1.1701	−0.3234	−0.3493
6.2	−5.8739	−5.7170	1.7090	0.1934	−0.2359	1.5281
7.1.1	4.1610	2.2536	0.1775	1.0976	−2.6734	0.1861
7.1.2	4.2882	4.4784	−1.1182	0.2768	0.0386	−2.4036
7.2.1	4.3117	3.0509	−0.2194	0.9898	0.1809	−1.9833
7.2.2	4.3284	3.2733	−0.9286	0.7415	−1.0757	−1.0430
8.1	−8.8239	5.4954	1.2720	4.7684	−0.1428	1.0801
8.2	−8.0694	4.3139	−1.5231	5.6667	3.1130	0.5582
9.1.1	0.6468	4.4113	3.3448	−2.6882	1.9146	−0.3797
9.1.2	1.2862	5.5270	1.7360	−3.5416	2.8117	3.1329
9.2.1	0.1561	0.2784	−0.9364	−0.8100	−0.5981	0.0892
9.2.2	−0.1287	0.9325	1.8555	−0.6934	0.5959	−1.1643
9.3.1	−0.3873	−0.1603	3.0373	−0.3006	−3.1157	0.9897
9.3.2	−0.2827	−0.6395	2.8592	−0.3025	−0.7054	−0.9675
10.1.1	7.5296	−3.0998	3.9813	1.5925	0.9975	−2.0763
10.1.2	7.6726	1.3574	−1.8310	−0.5161	−0.4564	−0.9028
10.2.1	8.3168	0.2849	4.3189	−0.5465	1.2660	0.6830
10.2.2	8.8218	3.3376	0.1809	−1.8163	0.8456	2.0753
10.3.1	7.9681	0.0713	−2.0128	0.5599	−1.5890	0.9229
10.3.2	7.5192	−2.1439	2.0070	1.3297	−1.9649	0.1591
10.4.1	7.9848	−1.1899	−1.6830	0.4285	−1.7291	1.5518
10.4.2	8.0537	−1.6182	−2.0384	0.4788	0.1030	−0.6392
11.1.1	1.2742	−2.3342	−2.4558	−1.0802	1.2188	−0.2514
11.1.2	1.2530	−7.5213	2.1144	0.9705	3.0859	−1.1629
11.2.1	1.5423	−3.1237	−3.2945	−0.4177	1.3260	0.1898
11.2.2	1.3098	−3.7138	−1.3866	0.0878	0.7538	−0.3457

Table 4. Selected Topological Indices for 38 Isospectral Graphs (Figure 2)

graph	${}^0\chi$	${}^1\chi$	${}^2\chi$	IC_0	IC_1	IC_2
1.1	8.690	5.219	3.859	0.898	1.368	2.665
1.2	8.690	5.240	3.812	0.898	1.368	2.701
2.1	8.975	5.812	4.424	0.918	1.418	2.675
2.2	8.975	5.791	4.502	0.918	1.418	2.828
3.1	11.380	7.847	6.318	0.932	1.384	2.726
3.2	11.380	7.826	6.396	0.932	1.384	2.664
4.1.1	9.966	6.847	5.610	0.934	1.417	2.784
4.1.2	9.966	6.826	5.689	0.934	1.417	2.765
4.2.1	9.966	6.864	5.526	0.934	1.417	2.684
4.2.2	9.966	6.864	5.526	0.934	1.417	2.684
5.1	8.975	5.753	4.643	0.918	1.418	2.807
5.2	8.975	5.774	4.575	0.918	1.418	2.717
6.1	9.682	6.291	4.856	0.918	1.404	2.789
6.2	9.682	6.312	4.766	0.918	1.404	2.565
7.1.1	11.121	7.809	6.906	0.946	1.457	2.794
7.1.2	11.121	7.809	6.908	0.946	1.457	2.982
7.2.1	11.121	7.809	6.896	0.946	1.457	2.856
7.2.2	11.121	7.809	6.896	0.946	1.457	2.856
8.1	7.845	5.326	4.628	0.938	1.469	2.802
8.2	7.845	5.326	4.618	0.938	1.469	2.995
9.1.1	10.889	7.232	6.134	0.933	1.517	2.978
9.1.2	10.889	7.220	6.193	0.933	1.517	2.885
9.2.1	10.836	7.258	6.116	0.933	1.458	2.928
9.2.2	10.836	7.236	6.194	0.933	1.458	2.928
9.3.1	10.836	7.274	6.041	0.933	1.458	2.864
9.3.2	10.836	7.274	6.004	0.933	1.458	2.974
10.1.1	12.535	8.847	7.431	0.943	1.429	2.664
10.1.2	12.535	8.809	7.594	0.943	1.429	2.729
10.2.1	12.588	8.805	7.518	0.943	1.483	2.764
10.2.2	12.588	8.815	7.482	0.943	1.483	2.764
10.3.1	12.535	8.847	7.443	0.943	1.429	2.760
10.3.2	12.535	8.847	7.441	0.943	1.429	2.729
10.4.1	12.535	8.847	7.431	0.943	1.429	2.664
10.4.2	12.535	8.830	7.516	0.943	1.429	2.769
11.1.1	11.380	7.809	6.458	0.932	1.384	2.589
11.1.2	11.380	7.830	6.378	0.932	1.384	2.438
11.2.1	11.380	7.847	6.306	0.932	1.384	2.622
11.2.2	11.380	7.847	6.308	0.932	1.384	2.595

For most of the isospectral pairs, ${}^0\chi$, ${}^1\chi$, IC_0, and IC_1 could not discriminate between the isospectral pairs, whereas ${}^2\chi$ as well as complexity parameter IC_2 could discriminate the isospectral pairs reasonably well in most cases.

We retained the first six PCs with eigenvalues > 1.0. This is a substantial reduction in the number of parameters or the dimensionality of the parameter space as compared to the 60-dimensional space corresponding to the 60 TIs calculated originally. Our earlier work on PCA using large and diverse sets of molecular graphs show that a few first PCs explain a large fraction of the variance.[16−20]

In some of their earlier papers, Basak *et al.*[16−20] used the Euclidean distance (ED) in the *n*-dimensional PC-space in characterizing structural similarity/dissimilarity of molecules. In Table 5 we give the ED between 19 isospectral pairs of graphs. For all pairs of graphs considered in this paper, the value of ED was nonzero which shows the discriminating ability of the six-dimensional PC-space generated out of the calculated PCs.

Table 5. Euclidean Distance in 7-Dimensional Principal Component Space for 19 Isospectral Graph Pairs

isospectral pairs		Euclidean distance	isospectral pairs		Euclidean distance
1.1	1.2	0.2142	9.1.1	9.1.2	0.6877
2.1	2.2	0.5781	9.2.1	9.2.2	0.4352
3.1	3.2	0.4627	9.3.1	9.3.2	0.5281
4.1.1	4.1.2	0.2230	10.1.1	10.1.2	0.8340
4.2.1	4.2.2	0.6627	10.2.1	10.2.2	0.5988
5.1	5.2	0.3705	10.3.1	10.3.2	0.5130
6.1	6.2	0.5929	10.4.1	10.4.2	0.4958
7.1.1	7.1.2	0.6831	11.1.1	11.1.2	0.7672
7.2.1	7.2.2	0.2773	11.2.1	11.2.2	0.2627
8.1	8.2	0.6324			

Results and Discussion. We have considered a series of pairs of isospectral graphs shown in Figure 2. In this figure we have used the numbering convention *i.j.k*, where *i* is the same for two isospectral graphs. Based on the relation between the isospectral graphs, the index *j* will be kept the same if the two are closely related; in this case only the index *k* would differ. Thus we have isospectral graphs 9.1.1., 9.1.2, 9.2.1, 9.2.2, 9.3.1, and 9.3.2. As seen from Figure 2, 9.3.1 and 9.3.2 are more closely related compared to 9.1.1 and 9.3.1. Recall that the isospectral graphs have the same characteristic polynomials and spectra. Furthermore, many parameters computed based on the adjacency matrices of two isospectral graphs are identical. Commonly used topological indices such as the Wiener index, Randić's connectivity index, spectral index, indices based on path numbers, etc., become identical for such graphs. Consequently, many ordinary graph-theoretically based indices fail to discriminate isospectral graphs.

We have computed the connectivity indices ${}^0\chi$, ${}^1\chi$, and ${}^2\chi$ as well as the neighborhood complexity indices IC_0, IC_1, and IC_2 that are defined in the previous section for these

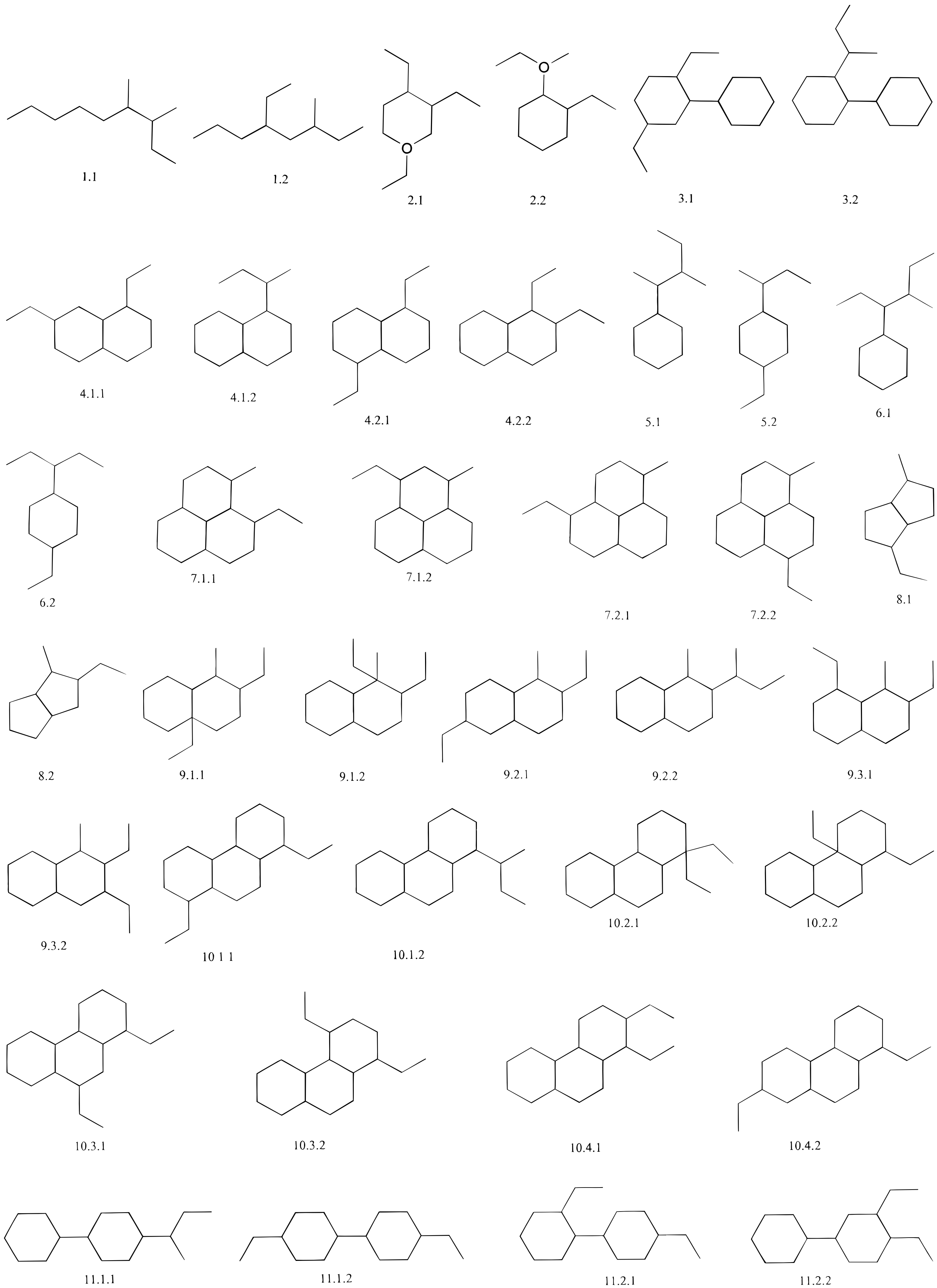


Figure 2. Structures of 38 isospectral graphs.

isospectral graphs shown in Figure 2. The isospectral graphs in Figure 2 are generated by attaching the same fragment at vertices called the isospectral vertices. As discussed before in the literature, certain vertices in some graphs are called isospectral vertices. For example, consider the graphs 2.1 and 2.2 in Figure 2. These two graphs are generated by attaching a fragment containing two vertices connected by a bond to either the para position of the six-membered ring, as in the graph 2.1 in Figure 2 (where the para position is defined as the fourth vertex in 2.1) or by attaching the same fragment to the other circled vertex of the pending fragment which results in the graph 2.2 in Figure 2. All of the isospectral graphs in Figure 2 are constructed in this manner by attaching an identical fragment to one of the isospectral vertices.

Table 4 shows the computed values for the indices ${}^0\chi$, ${}^1\chi$, and ${}^2\chi$ as well as the neighborhood complexity indices IC_0, IC_1, and IC_2. First let us discuss the discriminating powers of these indices before proceeding to the PCA. As seen from Table 4, the index ${}^0\chi$ is the least discriminating while ${}^2\chi$ is somewhat more discriminating. For all isospectral pairs of graphs the ${}^0\chi$ indices are identical as expected since the ${}^0\chi$ index is based on simple topological connectivity.

It is seen from Table 4 that although the ${}^2\chi$ index is relatively more discriminating compared to the ${}^0\chi$ index, the actual ${}^2\chi$ values are numerically too close for some of the isospectral graphs to consider these values to be truly discriminating. This is particularly exemplified by the graphs 11.2.1 and 11.2.2 whose ${}^2\chi$ values are 6.306 and 6.308, respectively (see Table 4). Likewise the ${}^2\chi$ values for the graphs 10.3.1 and 10.3.2 are 7.443 and 7.441, respectively. The ${}^2\chi$ values for the graphs 7.2.1 and 7.2.2. are identical (6.896). Likewise the ${}^2\chi$ values for the graphs 4.2.1 and 4.2.2 are the same (5.526). However, for other graphs considered here the ${}^2\chi$ values are more discriminating. Consequently, it is concluded that although the ${}^2\chi$ values are more discriminating than the zeroth-order index, these values are still not sufficiently discriminating for more complex isospectral graphs, although these indices work well for simpler isospectral graphs, as seen from Table 4.

As evidenced from Table 4, the neighborhood complexity indices IC_0, IC_1, and IC_2 have some similarity to the χ indices in that the higher-order indices are slightly more discriminating compared to the lower-order indices. Thus the IC_0 and the IC_1 indices do not discriminate isospectral graphs at all (see, Table 4). When ${}^2\chi$ is identical, IC_2 is as well. When ${}^2\chi$ is nearly identical, IC_2 is slightly more discriminating.

Since neither the ${}^n\chi$ indexes nor the IC_r indexes seem to be fully satisfactory in terms of discriminating complex isospectral graphs, it was decided to carry out the PCA on these graphs using the indices computed thus far. The philosophy behind the PCA technique and the algorithms derived from the technique have been illustrated in the previous section. The procedure uses an *n*-dimensional space of these indices and computes the Euclidian distances.

Table 3 shows the numerical values for the first six PCs which are labeled PC_1 through PC_6 in Table 4 for the isospectral graphs that are considered in this study. In this analysis we retained only the first six PCs with eigenvalues > 1.0 which leads to a substantial reduction in the number of parameters or the dimensionality of the parameter space as compared to the original 60-dimensional parameter space that we begin with. Earlier work on PCA using large and diverse sets of molecular graphs show that the first few PCs explain a large fraction of the variance.[17−20]

As seen from Table 3, the PC indices are far more powerful and discriminating compared to the simple topological indices considered in Table 4. Let us consider graphs 11.2.1 and 11.2.2 which are considered to be "pathological" from numerical and similarity standpoints in that the χ values and IC_r values are virtually the same. However, as seen from Table 3, the PC_1 and PC_2 values are very different (PC_1: 1.5423, 1.3098; PC_2: −3.1237, −3.7138). As a matter of fact all of the PC_1 through PC_6 values are sufficiently different to discriminate these isospectral graphs.

Let us consider graphs 7.2.1 and 7.2.2 that are not discriminated by their ${}^2\chi$ values. As seen from Table 3, while the PC_1 values for these two graphs are somewhat close (4.3117 and 4.3284) their PC_2 values are 3.0509 and 3.2733. Other higher order PC values differ even more thereby providing a sound and powerful basis for discriminating isospectral graphs.

Next we consider the pairs 4.2.1 and 4.2.2. These two graphs have identical ${}^2\chi$ values and IC_2 values. However, as seen from Table 4 these graphs have very different PC values for all *n*. Thus PCA seems to be a powerful technique to discriminate even isospectral graphs that are not so easily contrasted by topologically based techniques.

It should be pointed out that for a few isospectral graphs the first principal component value, PC_1 is not as discriminating as the higher-order PCs values. For example, the PC_1 values for the isospectral graphs 2.1 and 2.2 are −7.5623 and −7.6856, respectively. However, the PC_2 values are −2.8764 and 1.4163 for the same graphs. Likewise the graphs 7.2.1 and 7.2.2 have the PC_1 values of 4.3117 and 4.3284. However their PC_2 values are 3.0509 and 3.2733. We thus conclude that one needs more than the PC_1 value to discriminate complex isospectral graphs, but often the PC_2 values for those graphs are sufficiently different to contrast them.

ACKNOWLEDGMENT

The research effort of Subhash C. Basak reported in this paper was supported by cooperative agreement CR 819621 from the United States Environmental Protection Agency, Grant F49620-94-1-0401 from the United States Air Force and Exxon Biomedical, Inc., through the structure−activity relationship consortium (SARCON) of the Natural Resources Research Institute. This is contribution number 222 from the Center of Water and the Environment of the Natural Resources Research Institute.

REFERENCES AND NOTES

(1) Basak, S. C.; Bertelsen, S.; Grunwald, G. D. Application of Graph Theoretical Parameters in Quantifying Molecular Similarity and Structure-activity Studies. *J. Chem. Ink Comput. Sci.* **1994**, *34*, 270−276.

(2) Basak, S. C.; Harriss, D. K.; Magnuson, V. R. POLLY 2.3: Copyright of the University of Minnesota.

(3) Wiener, H. Structural Determination of Paraffin Boiling Points. *J. Am. Chem. Soc.* **1947**, *69*, 17−20.

(4) Randić, M. On Characterization of Molecular Branching. *J. Am. Chem. Soc.* **1975**, *97*, 6609−6615.

(5) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure-Activity Analysis;* Research Studies Press: Letchworth, Hertfordshire, England, 1986.

(6) Raychaudhury, C.; Ray, S. K.; Ghosh, J. J.; Roy, A. B.; Basak, S. C. Discrimination of Isomeric Structures Using Information Theoretic Topological Indices. *J. Comput. Chem.* **1984**, *5*, 581−588.
(7) Bonchev, D.; Trinajstić, N. Information Theory, Distance Matrix and Molecular Branching. J. *Chem. Phys.* 1977, 67, 4517−4533.
(8) Basak, S. C. Use of Molecular Complexity Indices in Predictive Pharmacology and Toxicology: A QSAR Approach. *Med. Sci. Res.* **1987**, *15*, 605−609.
(9) Basak, S. C.; Roy, A. B.; Ghosh, J. J. Study of the Structure-function Relationship of Pharmacological and Toxicological Agents Using Information Theory. In *Proceedings of the 2nd International Conference on Mathematical Modelling;* Avula, X. J. R., Bellman, R., Luke, Y. L.,Rigler, A. K., Eds.; University of Missouri-Rolla: Rolla, Missouri, 1980; pp 851−856.
(10) Roy, A. B.; Basak, S. C.; Harriss, D. K.; Magnuson, V. R. Neighborhood Complexities and Symmetry of Chemical Graphs and Their Biological Applications. In *Mathematical Modelling in Science and Technology;* Avula, X . J . R. Kalman, R .E., Lipais, A. I., Rodin, E. Y., Eds.; Pergamon Press 1984; pp 745−750.
(11) Basak, S. C.; Magnuson, V. R. Molecular Topology and Narcosis: a Quantitative Structure-activity Relationship Qsar (study of alcohols using complementary information content CIC). *Arzneim. Forsch./ Drug Res.* **1983**, *33*, 501−503.
(12) Shannon, C. E. A Mathematical Theory of Communication. *Bell Sys. Tech. J.* **1948**, *27*, 379−423.
(13) Rashevsky, N. Life, Information Theory and Topology. *Bull. Math. Biophys.* **1955**, *17*, 229−235.
(14) Trucco, E. A Note on Rashevsky's Theorem about Point Bases in Topological Biology. *Bull. Math. Biophys.* **1956**, *18*, 65−85.
(15) Sarkar, R.; Roy, A. B.; Sarkar, P. K. Topological Information Content of Genetic Molecules-I. *Math. Biosci.* **1978**, *39*, 299−312.
(16) 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.
(17) Basak, S. C.; Grunwald, G. D. Use of Topological Space and Property Space in Selecting Structural Analogs. *Mathl. Model. Sci. Comput.* **1994**, in press.
(18) Basak, S. C.; Grunwald, G. D. Estimation of Lipophilicity from Structural Similarity. *New* J. *Chem.* **1995**, *19*, 231−237.
(19) Basak, S. C.; Grunwald, G. D. Molecular Similarity and Risk Assessment: Analog Selection and Property Estimation Using Graph Invariants. *SAR QSAR Environ. Res.* **1994**, *2,* 289−307.
(20) 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.
(21) Balasubramanian, K. *Chem. Phys Lett.* **1990**, *169,* 224; **1995**, *232,* 415.
(22) Balasubramanian, K. *J. Chem. Inf. Comput. Sci.* **1995**, *35,* 243.
(23) Balasubramanian, K. *J. Chem. Inf. Comput. Sci.* **1995**, *35,* 761.
(24) Liu, X.; Balasubramanian, K.; Munk, M. E. *J. Chem. Inf. Comput. Sci.* **1990**, *30,* 263.
(25) Herndon, W. C. *Inorg. Chem.* **1983**, *22,* 654.
(26) Herndon, W. C. In *Chemical Applications of Topology & Graph Theory;* King, R. B., Ed.; Elsevier: Amsterdam, 1983.
(27) Razinger, M.; Balasubramanian, K.; Munk, M. C. *J. Chem. Inf. Comput. Sci.* **1993**, *33,* 197.
(28) Balaban, A. T.; Liu, X.; Klein, O. J.; Babic, D.; Schmalz, T. G.; Seitz, W. A.; Ravic, M. Graph invariants for Fullerenes. *J. Chem. Inf. Comput. Sci,* **1995**, *35*, 396.
(29) Gnanadesikan, R. Methods for Statistical Analysis of Multivariate Observations; J Wiley: New York, 1977.
(30) SAS/STAT User's Guide. Version 6, 4th ed.; SAS Institute Inc.: Cary, NC, 1989; Vol. 2, p 846.

CI970052G