Journal of Mathematical Chemistry 8(1991)121–134

STUDY OF COMPUTATION, RELATEDNESS AND ACTIVITY PREDICTION OF TOPOLOGICAL INDICES

I. FABIČ-PETRAČ and B. JERMAN-BLAŽIČ
Institut Jožef Stefan, Ljubljana, Yugoslavia

V. BATAGELJ
Department of Mathematics, University of Ljubljana, Yugoslavia

Abstract

A large number of numerical graph invariants (topological indices) have been define and used in many different fields of chemistry. Some of them are used as chemical structure descriptors in QSAR (quantitative structure–activity relationship) studies. The paper describes the development and implementation of a computer program for computation of the most often used topological indices: n, $n \ln n$, A, M_1, M_2, X_R, F, x_1, W, p, J, $D^{(2)}$, GDI, r, $\overline{I}_D^E$, $\overline{I}_w^E$. As these indices reflect different aspects of molecular topology, the intercorrelation among them is investigated by applying hierarchical clustering methods. A method based on string comparison techniques is developed for the determination of indices correlated to biological activity for a studied set of compounds. The biological activity prediction on the basis of a subset of topological indices least-correlated amongst themselves is done by applying the nearest neighbourhood approach.

1. Introduction

In the near past, graph-theoretic methods have been largely applied in many fields of chemistry. Recently, we are witnessing the growth of their use as methods for characterization of chemical structures and use in the field of structure–activity correlations. Graph models of chemical structures retain their full topology which largely determines important characteristics of molecules [1]. The topology of the molecule represents the mutual relations of the atoms, i.e. the information of the connectivity between them in the structure [2]. Mathematicians and chemists have undertaken a long search to find suitable parameters derived from the molecular graph which will reflect the molecular topology. During the last two decades, a large number of numerical graph invariants have been defined and applied in different fields of theoretical chemistry, pharmacology, toxicology, and environmental chemistry. Such quantitative measures which reflect the structural features of the molecule are called topological indices. A topological index of a compound under consideration is a graph invariant derived from the chemical graph of that compound. Topological indices express in numerical form the topology of the chemical compound represented. Topological indices were developed for the purpose of obtaining correlations with the physicochemical properties of chemical substances [3] and to express the molecular

similarity or dissimilarity. They were used for classification and for prediction of chemical and biological properties of the compounds. One of the most important developments in recent years has been the increasing use of topological indices in the design of drugs and other biologically active substances. One important question in drug design is to what extent the structure of a compound and its biological activity are correlated. This is also the main task for QSAR (quantitative structure–activity relationship) analysis. The basic assumption is that the molecular properties are determined primarily by structure and that compounds with similar structure lead to similar biological action or other physicochemical properties. Once a relationship between the molecular descriptors and biological activity is known, it is used to predict the activity of yet untested or even unsynthesized compounds [3,4]. In the text that follows, a study of the relatedness of the topological indices is made. Special methods were developed for determination of less correlated indices. In the process of determination, the ordering of the compounds represented by their topological indices and the corresponding biological response value were used. Some property predictions were done by applying the nearest neighbour techniques appropriately accommodated for the studied set of compounds.

2. Topological indices

The graph representation of a molecule consists of points representing the molecule's atoms and bonds linking them as straight lines. In graph theory, points are usually referred to as vertices and lines are referred to as edges. In chemical graphs, the hydrogen atoms are often omitted because they normally do not play a major role in determining the structure of a molecule. The topological indices are derived from such graphs, called hydrogen-suppressed graphs or, sometimes, skeletal structures. Graphs with multiple edges (double, triple bond), weighted vertices (heteroatoms), and weighted edges can also be used, although usually they are not considered. The length of any line representing a chemical bond and the angles between the lines are not considered in this approach.

A topological index of a compound under consideration is a graph invariant derived from the chemical graph of the compound. This number characterizes the molecule and it does not depend on how the vertices are enumerated. Indices are designed by transforming a chemical graph into a number and the means by which this was accomplished varies from index to index.

A graph invariant may be a polynomial (characteristic polynomial), a set of numbers (spectrum of a graph), or a numerical value. In particular, the molecules may be described by several quantitative descriptors reflecting different structural aspects. The purpose of defining a topological index is sometimes to represent each chemical structure with a numerical value, keeping it at the same time as discriminatory as possible. This means assigning to every chemical graph a numerical invariant such that two graphs have the same value of that index if and only if they are

isomorphic. The coincidence of the invariants of two graphs is a prerequisite for isomorphism, but it is not a sufficient condition. Until now, such a sufficient condition and an efficiently computable graph invariant have not been discovered [6].

2.1. ADJACENCY MATRIX AND DISTANCE MATRIX

The fundamental mathematical structures which map the molecular graph G into mathematical terms suitable for derivation of the topological indices are the adjacency matrix $A(G)$ and the distance matrix $D(G)$. We denote the molecular graph as $G = (V, E)$, $V(G)$ is the *vertex set* and $E(G)$ is the *edge set* of G. We assume that $V = \{1, 2, \ldots, n\}$. For a graph with n vertices, the corresponding *adjacency matrix* $A(G)$ is a square $n \times n$ matrix. Its entries a_{ij} have either value 1 or 0, i.e. vertices i and j being connected or not:

$$A(G) = [a_{ij}]_{n \times n} \quad a_{ij} = \begin{cases} 1 & i \text{ and } j \text{ are adjacent,} \\ 0 & \text{otherwise.} \end{cases}$$

For the derivation of a majority of topological indices, we assume that G is a simple graph, i.e. that G possesses no loops and multiple edges. In this case, we are only interested in whether a connection between two atoms exists and we assume that the edge has a weight equal to unity. The derivation of some topological indices requires information about the nature of the chemical bonds. It is supposed [7] that two atoms connected with more than one edge (i.e. double or triple bond) are closer to each other compared to the atoms connected with only one edge. A multiple edge is thus represented in the following way [2]:

(a) multiple edge between two atoms: if the bond order between vertices i and j is b, then $a_{ij}^w = 1/b$;

(b) aromatic systems: if n atoms are connected by m bonds, then every edge in this cycle has value n/m, e.g. for benzene $a_{ij}^w = 6/9 = 2/3$.

The weighted adjacency matrix with entries a_{ij}^w is denoted as

$$A^w(G) = [a_{ij}^w]_{n \times n} \quad a_{ij}^w = \begin{cases} w_{ij} & i \text{ and } j \text{ are adjacent,} \\ 0 & \text{otherwise.} \end{cases}$$

w_{ij} is the weight of the edge between vertices i and j.

The *distance matrix* $D(G)$ of a graph G is a square $n \times n$ matrix with entries d_{ij} indicating the distance between all pairs of vertices. The distance means the length of the shortest path connecting vertices i and j. In our approach, the distance matrix is calculated from the adjacency matrix according to the well-known algorithm [8] for computing the shortest paths in a graph. The length of the shortest

path between two vertices in a graph is equal to the minimum number of edges connecting them. In the case of a graph with weighted edges, the weighted distance matrix $D^w(G)$ is computed from the weighted adjacency matrix $A^w(G)$. Its entries d_{ij}^w are equal to the minimum sum of weights of edges along the path between vertices i and j.

2.2. TOPOLOGICAL INDICES DERIVED FROM THE ADJACENCY MATRIX

The following topological indices may be calculated from the adjacency matrix:

(a) *Total adjacency index*

$$A = \left(\sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} \right) /2$$

is equal to the number of edges. If the total adjacency index is known, then a cyclomatic number (number of independent cycles) can be derived as: $\lambda = A - n + 1$.

(b) *The Zagreb group indices*

$$M_1 = \sum_{i=1}^{n} v_i^2 ,$$

where v_i is the degree of vertex i which is equal to the sum of all entries of the ith row in the $A(G)$,

$$v_i = \sum_{j=1}^{n} a_{ij} ,$$

$$M_2 = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} v_i v_j .$$

(c) *The Randić connectivity index*

$$X_R = \sum_{a_{ij} \neq 0} \frac{1}{\sqrt{v_i v_j}} .$$

(d) *The Platt index*

$$F = \sum_{i=1}^{A} e_i ,$$

where e_i is the degree of the ith edge in the graph. The Platt index can be calculated through the equation $F = M_1 - 2A$ when the values of M_1 and A are already known.

(e) *The largest eigenvalue index*

$$x_1 = \max_{x_k} \{\det(A - x_k I) = 0\}$$

is the largest eigenvalue of the adjacency matrix.

2.3. TOPOLOGICAL INDICES DERIVED FROM THE DISTANCE MATRIX

The following indices may be derived from the distance matrix:

(a) *The Wiener index*

$$w = \left(\sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij} \right) /2.$$

w is also called the path number of a graph.

(b) *The polarity number*

$$p = \left(\sum_{i} d_i^{(3)} \right) /2.$$

p is the number of distances of length three in a graph (by $d_i^{(3)}$, we denote entries of length three in the distance matrix D).

(c) *Average distance sum connectivity index or the Balaban index*

$$J = \frac{q}{\lambda + 1} \sum_{a_{ij} \neq 0} \frac{1}{\sqrt{v_i^w v_j^w}},$$

where λ is the cyclomatic number, calculated from the adjacency matrix, and q is the number of edges; v_i^w is also called the distance sum index. The Balaban index can be presented in a form which is more suitable for computation:

$$J = \frac{A}{A - n + 2} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{a_{ij}}{\sqrt{v_i^w v_j^w}}.$$

(d) *Mean square distance topological index*

$$D^{(2)} = \sqrt{\frac{\sum_{i=1}^{n} g_i i^2}{\sum_{i=1}^{n} g_i}},$$

where g_i is the number of occurrences of distance with length i in the graph.

(e) *Graph distance index*

$$GDI = \sum_{i=1}^{n} g_i^2.$$

2.4. CENTRIC TOPOLOGICAL INDICES

(a) *Radius of a graph*

$$r = \min_i (\max_j d_{ij}).$$

2.5. INDICES BASED ON INFORMATION THEORY

Information theory is a convenient basis for quantitative characterization of structures. It introduces simple structural indices called information content. Information-theoretic topological indices are derived by the application of Shannon's formalism to chemical graphs [9]. A set A of K elements is derived from the molecular graph on the basis of certain structural characteristics, i.e. the shortest paths (distances) in a graph. The set A is partitioned into k disjoint subsets A_i of order n_i, $\sum_{i=1}^{k} n_i = K$. This partition can be denoted as $P = K(n_1, n_2, \ldots, n_k)$. Subsets A_i represent classes of an equivalence relation defined on A. Then, $p_i = n_i/K$ is the probability that a randomly selected element of A will lie in the ith subset and a probability distribution $Pd = K(p_1, p_2, \ldots, p_k)$ is constructed. The entropy of this distribution is determined by the Shannon formula [10]:

$$\bar{I} = -\sum_{i=1}^{k} p_i \log_2 p_i$$

and is called mean information content or the information index of the structure [9].

The set of the shortest paths (distances) in a graph can be distributed into subsets in two ways [11]:

- The total number of distances is partitioned into classes of distances according to their equality or inequality in the following way:

$$P_D^E = \frac{n(n-1)}{2} (g_1, g_2, \ldots, g_{d_{\max}})$$

and then *the information index for the equality of distances* is given by:

$$\bar{I}_D^E = -\sum_{i=1}^{d_{max}} \frac{2g_i}{n(n-1)} \log_2 \frac{2g_i}{n(n-1)} .$$

- The total distance (the Wiener index) is partitioned into different individual distances d_i:

$$P_D^w = w(g_1 d_1, g_2 d_2, \ldots, g_{d_{max}} d_{max})$$

and the respective *information index for the magnitude of distances* is calculated by the equation:

$$\bar{I}_D^w = -\sum_{i=1}^{d_{max}} g_i \frac{d_i}{w} \log_2 g_i \frac{d_i}{w} .$$

3. Relatedness of the topological indices

The large number of existing topological indices raises the question as to what extent they are correlated. That is, to what extent do they express the same type of structural information. Topological indices have been demonstrated [2, 12] to reflect in general the shape and the size of the compounds or the degree and nature of branching [5] of the molecular structures they represent. It has been investigated and found [13] that there exists strong correlation between X_R, w, $\bar{I}_D^w$, M_1, and x_1. These indices reflect predominantly the Van der Waals volume of the molecule. It was also found [14] that indices can be classified into two separate classes. Indices belonging to one of these classes express the same structural information:

(a) w, X_R, x_1, M_1, $\bar{I}_D^E$, $\bar{I}_D^w$;
(b) centric indices of Balaban.

The indices w, M_2, $\bar{I}_D^E$, $\bar{I}_D^w$ and M_1 reflect primarily the size of a molecule, i.e. the Van der Waals volume. The centric indices of Balaban express mainly the amount of branching, while the largest eigenvalue of the adjacency matrix x_1 contains both: components of shape and size. Principal component analysis (PCA) was carried out on the 90×90 variable matrix [15] corresponding to 90 topological indices calculated for a large data set of chemical structures. It was found that the first principal axis corresponds to the parameters expressing size and shape of the molecular graph and the second axis to parameters representing the neighbourhood indices. The third axis was found to represent the degree of branching in the molecule.

These results suggest that a selection of the least correlated topological indices may be done and later used as an efficient description of the chemical structure. The selected indices should contain information about different structural characteristics. Because every topological index has its own specific contribution to the information about the molecular structure, it must be noticed that the best description of the molecular structure will be obtained by the use of all the different topological indices; but this is impractical, especially if indices which are not efficiently computable are used.

If topological indices are used for biological activity prediction for a given set of compounds, then all subsets of topological indices should be examined. This is a computationally very demanding task, so we try to reduce the number of indices by eliminating indices expressing similar structural information. The purpose of reducing the number of topological indices is to obtain a subset of them such that it will be as good in property prediction as if the property prediction is done by application of a whole set of topological indices. Hence, the task is to reduce the number of topological indices with a small loss of the activity prediction accuracy.

In our approach, the study of the correlations between different sets of topological indices and biological response was made for the database consisting of 81 benzamide and benzamidine derivatives with anticonvulsant activity (compound nos. 1–17), dopamine receptor affinity (nos. 18–28) and with inhibition effects on the antibody–antigen interactions [16]. Their biological activity was represented by the logarithm of the octanol/water partition coefficient.

3.1. REDUCTION OF THE NUMBER OF TOPOLOGICAL INDICES BY HIERARCHICAL CLUSTERING METHODS

The selection of indices may be done by application of hierarchical clustering methods. Hierarchical clustering methods are widely used in chemical classifications [17,18]. Ward's method was found to be the most useful [17,19] for these purposes.

Sixteen topological indices (n, $n \ln n$, A, M_1, M_2, X_R, F, x_1, W, p, J, $D^{(2)}$, GDI, r, $\overline{I}_D^E$, $\overline{I}_w^E$) were calculated for 81 compounds. Pearson's correlation coefficients r_{ij} between standardised data vectors were used as a measure of similarity. The dissimilarity matrix obtained by transformation $\sqrt{1-r_{ij}}$ was used as the input data for the computer program CLUSE [20]. The clusterization of indices and the corresponding dendrogram are shown in fig. 1. The following grouping of the topological indices was obtained through the inspection of the resulting dendrogram:

(1) M_1, M_2, F, p, n, A, X_R, $n \ln n$;

(2) GDI;

(3) $D^{(2)}$, $\overline{I}_D^E$;

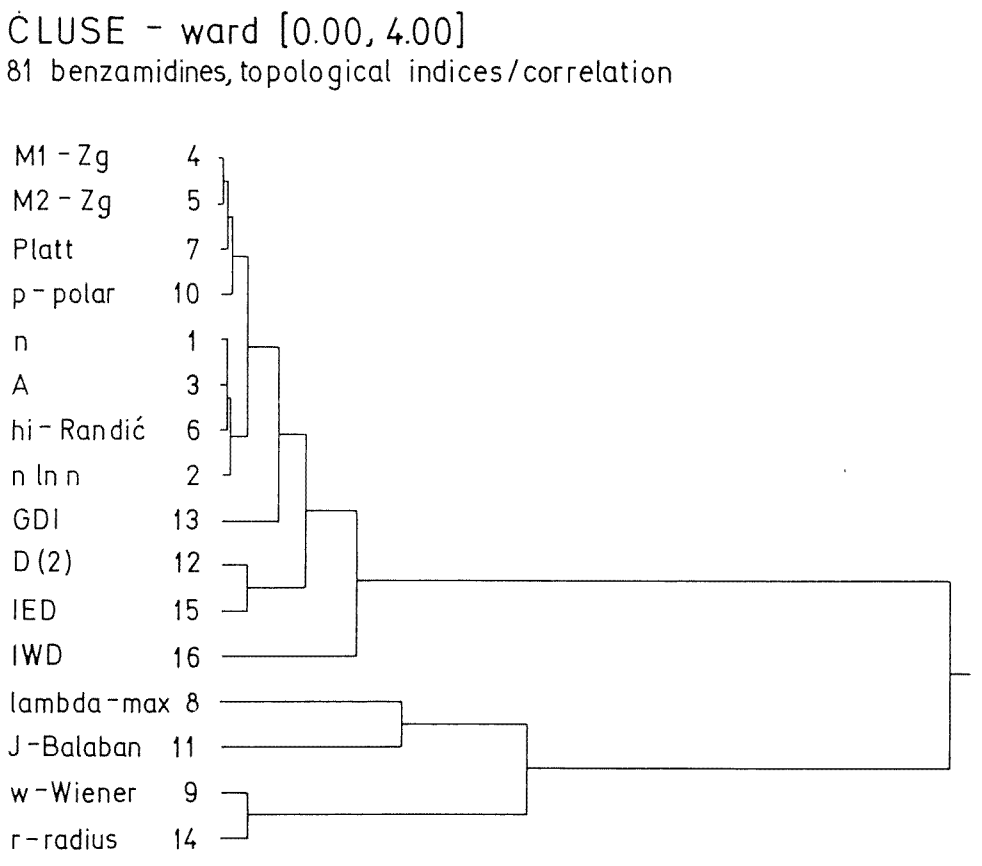

Fig. 1. Hierarchical clustering of topological indices using Ward's method.

(4) $\overline{I}_D^w$;

(5) x_1;

(6) J;

(7) w, r.

The next step was choosing an index from each group for a representation of the group. The index with the lowest sum of correlation coefficients to the others in the same group was selected. If the group contained only two indices, then the sum of correlation coefficients between a particular index and all indices belonging to all other groups was calculated and the index with the greater sum was marked as the representative. The subset of selected indices corresponding to the groups presented in fig. 1 was $\{M_1, GDI, \overline{I}_D^E, \overline{I}_D^w, x_1, J, w\}$.

3.2. SELECTION OF TOPOLOGICAL INDICES ON THE BASIS OF THEIR CORRELATION TO BIOLOGICAL ACTIVITY

The purpose of a selection is to obtain a composite index (i.e. a subset of indices) with the best predictive power. The selection is done by considering the correlation of the indices with the parameters of biological activity. At first, we have ordered the compounds according to ascending values of their biological response. Then the compounds were ordered according to the values of a particular topological index. If this index is correlated to biological activity, then it is expected that the ordering of compounds is similar to the ordering of compounds ordered according to the values of biological responses. Among all strings of compounds which have been ordered according to different topological indices we wanted to find those with similar ordering to the string obtained by ordering according to the biological activity. A method and procedure from the area of string comparison techniques was used for a determination of the similarity between orderings. The measure of similarity between two strings was expressed as the sum of all different identities found in all traces [21] between the first string and all substrings of the second string obtained by successively eliminating the left-most element. The procedure is explained in detail elsewhere [22]. The obtained correlations are displayed in table 1. The indices with values greater than 0.8 were selected as the members of the composite index. These are $\{r, D^{(2)}, \bar{I}_D^E, n, A, F\}$. We can see that the selected indices are from the groups of indices which reflect the shape and the size of the molecules.

Table 1

Correlations between topological indices and biological activity for benzamidines

Topological index	Measure of similarity between orderings
r	0.8522
$D^{(2)}$	0.8272
$\bar{I}_D^E$	0.8250
n	0.8130
$n \ln n$	0.8130
A	0.8127
F	0.8043
$\bar{I}_D^w$	0.7997
p	0.7985
M_1	0.7981
w	0.7972
X_R	0.7951
M_2	0.7932
GDI	0.7790
x_1	0.6031
J	0.1985

4. Activity prediction

The prediction of biological activity was done according to three different subsets of topological indices. First the subset of indices obtained by hierarchical clustering methods was considered, then the subset of indices derived on the basis of correlation to biological activity was used. In the last trial, all sixteen topological indices n, $n \ln n$, A, M_1, M_2, X_R, F, x_1, w, p, J, $D^{(2)}$, GDI, r, $\overline{I}_D^E$, $\overline{I}_D^w$ were considered for comparison of the activity prediction accuracy.

The prediction procedure assumes that the predicted value of biological activity depends on the values of some elements in the neighbourhood. By the term neighbourhood, we mean the compounds which are similar enough to the compound with unknown biological activity. A general equation describing this relation is:

$$b_{\text{predicted}}(x) = \sum_{y \in N} b(y) f(s(x,y)) a(x), \tag{1}$$

where N is a selected neighbourhood, $f(s(x, y))$ represents a function of a similarity $s(x, y)$ between compounds from the neighbourhood and the compound with unknown biological activity, while $a(x)$ is a normalizing factor. The measure of similarity between compounds x and y was obtained as $1/(1 + e(x, y))$, where $e(x, y)$ is the Euclidean distance between vectors of topological indices for compounds x and y.

The most often used procedure [17, 18, 23] for activity prediction computes the predicted value as the average of biological responses of the compounds from selected neighbourhoods. Let us assume that the considered neighbourhood N consists of m compounds with known biological responses $b(i)$, $i = 1, \ldots, m$. In this case, the predicted value is:

$$b_{\text{predicted}}(x) = \frac{1}{m} \sum_{y \in N} b(y) = \frac{1}{m} \sum_{i=1}^{m} b(i). \tag{2}$$

We will try to improve the results of prediction obtained by this procedure by weighting the compounds which are structurally more similar to the compound with unknown biological activity. This is done by application of the following equation:

$$b_{\text{predicted}}(x) = \bar{b} + \frac{1}{m} \sum_{i=1}^{m} (b(i) - \bar{b})(s(x,i) - \bar{s}); \tag{3}$$

$$\bar{b} = \frac{1}{m} \sum_{i=1}^{m} b(i), \quad \bar{s} = \frac{1}{m} \sum_{i=1}^{m} s(x,i).$$

In both cases, the prediction error is evaluated by the expression:

$$\text{prediction error}(x) = \frac{|b_{\text{predicted}}(x) - b_{\text{measured}}(x)|}{b_{\text{measured}}}. \tag{4}$$

Two main ideas concerning the selected neighbourhood were tested. In the first, the biological activity was predicted on the basis of the values of activity for all compounds whose similarity measure between them and the compound with unknown activity was greater than a chosen threshold value. The whole prediction error was calculated as the average of prediction errors for all compounds for which the prediction was possible. The results of this approach are shown in table 2.

Table 2

Property prediction considering the threshold value of the measure of similarity

Threshold value	$\{M_1, GDI, \overline{I}_D^E, \overline{I}_D^w, x_1, J, w\}$	$\{r, D^{(2)}, \overline{I}_D^E, n, A, F\}$	All top. indices
0.8	4%	5%	
	7% (53)	8% (79)	
0.7	6%	6%	3%
	7% (65)	7% (79)	7% (54)
0.65	6%		
	7% (69)		
0.6	6%		6%
	7% (70)		7% (66)
0.5	7%	6%	6%
	8% (75)	7% (81)	7% (74)
0.4			8%
			9% (77)

Underlined numbers represent the predictions obtained by considering similarity between the compounds from selected neighbourhoods and the unknown one (eq. (3)), while the others represent just the average of their biological responses (eq. (2)). The numbers in brackets show the number of compounds for which the prediction was possible (this means that there were some compounds with degree of similarity greater than a chosen threshold value). In this way, very good prediction values were obtained because only the compounds with high structural similarity were considered. This method does not allow a prediction for all compounds, i.e. for those with lower similarity from the threshold value.

Therefore, another approach was tested. For each compound, the neighbourhood consisting of a fixed number of nearest neighbours was considered. The predictions were computed according to eq. (3). The results are shown in table 3. We can see that with a smaller number of nearest neighbours (3–3), the biological activity for all compounds is much better predicted.

From table 3, it could be concluded that the subset of indices obtained by hierarchical clustering methods contains different aspects of the molecular structure, while the second subset contains mainly the factors which influence the biological

Table 3

Property prediction based on a fixed number of nearest neighbours

Number of NN	$\{M_1, GDI, \overline{I}_D^E, \overline{I}_D^w, x_1, J, w\}$	$\{r, D^{(2)}, \overline{I}_D^E, n, A, F\}$	All top. indices
2	8.3%	9%	7%
3	5.1%	5.8%	4.9%
4	5.3%	5.4%	5.1%
5	5.6%	5.8%	5.6%
6	6%	5.9%	5.7%
7	6.3%	6%	5.9%
8	6.4%	6%	6%
9	6.4%	6.1%	6.2%
10	6.5%	6.2%	6.2%
11	6.6%	6.2%	6.3%
12	6.8%	6.3%	6.4%
13	7%	6.5%	6.6%
14	7%	6.8%	6.8%
15	7%	6.9%	6.9%

activity. The number of topological indices can be reduced to 5 or 6 indices, which is quite convenient especially when the method is applied for elimination of topological indices with time-consuming computation (typically, their computation represents an NP-hard problem) [2].

5. Conclusion

The graph-theoretic approach is of great importance in the field of chemistry, and graph invariants are powerful tools in chemical applications. The complexity of the algorithm for the computation of a series of topological indices is polynomial because we selected efficiently computable indices. The indices can be easily computed and applied for property prediction for a large data set. One important area where indices are likely to have a major impact in the future is in the design of drugs. Topological indices might substantially shorten the length of the drug design process by predicting the activity of compounds directly from their molecular graphs. For the best biological response prediction for a given set of compounds, we have to examine all subsets of a set of topological indices. Because this is computationally too complex, an empirical method was developed. We first tried to eliminate indices containing similar structural information by application of hierarchical clustering methods and then on the basis of their correlation to biological activity for a set of biologically active compounds. In this way, we succeeded in reducing substantially the number of indices to be computed without loss of the activity prediction accuracy.

References

[1] A.T. Balaban, J. Chem. Inf. Comput. Sci. 25(1985)334–343.
[2] A.T. Balaban, I. Motoc, O. Mekenyan and D. Bonchev, *Topics in Current Chemistry*, Vol. 114, No. 21 (Springer, Berlin/Heidelberg, 1983).
[3] Y.C. Martin, *Quantitative Drug Design, A Critical Introduction*, Medicinal Research Series, Vol. 8 (Marcel Dekker, 1978).
[4] A.J. Stuper, W.E. Brugger and P.C. Jurs, *Computer-Assisted Studies of Chemical Structure and Biological Function* (Wiley–Interscience, New York, 1979).
[5] L. Kier and L. Hall, *Molecular Connectivity in Chemistry and Drug Research* (Academic Press, New York, 1976).
[6] J.D. Benstock, D.J. Berndt and K.K. Aggarwal, Discr. Appl. Math. 19(1988)45–63.
[7] A.T. Balaban, *Chemical Applications of Graph Theory* (Academic Press, New York, 1976).
[8] Aho, Hopcroft and Ulmann, *The Design and Analysis of Computer Algorithms* (Addison–Wesley Series in Computer Science and Information Processing, Reading, MA, 1974).
[9] D. Bonchev, *Information Theoretic Indices for Characterization of Chemical Structures* (Research Studies Press, Chichester, 1983).
[10] C. Shannon and W. Weaver, *Mathematical Theory of Communication* (University Illinois Press, Urbana, 1949).
[11] D. Bonchev and N. Trinajstić, J. Chem. Phys. 67, No. 10 (1977).
[12] M. Barysz, D. Plavšić and N. Trinajstić, Commun. Math. Chem. 19(1986)89–116.
[13] I. Motoc, A.T. Balaban, O. Mekenyan and D. Bonchev, Commun. Math. Chem. 13(1982)369–404.
[14] D.H. Rouvray, in: *Chemical Application of Topology and Graph Theory*, ed. R.B. King, Studies in Physical and Theoretical Chemistry, Vol. 28 (Elsevier, Amsterdam, 1983), pp. 159–177.
[15] S.C. Basak, G.J. Niemi, R.R. Regal and G.D. Veith, *Proc. 5th Int. Conf. on Mathematical Modelling*, San Francisco (1985).
[16] C. Hansh and M. Yashimoto, J. Med. Chem. 17(1974)1160.
[17] P. Willett, Anal. Chim. Acta 136(1982)29–37.
[18] P. Willett, J. Chem. Inf. Comput. Sci. 24(1984)29–33.
[19] P. Willett, J. Chem. Inf. Comput. Sci. 25(1985)78–80.
[20] V. Batagelj, CLUSE – Clustering Programs, Manual, Ljubljana (1989).
[21] J. Kruskal, SIAM Rev. 25(1983)201–237.
[22] B. Jerman-Blažič, I. Fabič and M. Randić, J. Comp. Chem. 7(1986)176–188.
[23] G.W. Adamson and J.A. Bush, J. Chem. Inf. Comput. Sci. 15(1975)55–58.