On the Basis of Invariants of Labeled Molecular Graphs

Igor I. Baskin and Mariya I. Skvortsova

Institute of Organic Chemistry, Leninsky pr. 47, Moscow 117913, Russia

Ivan V. Stankevich*

Institute of Organoelement Compounds, Vavilov str. 28, Moscow 117813, Russia

Nikolai S. Zefirov

Department of Chemistry, Moscow State University, Moscow 119899, Russia

Received October 28, 1994[⊗]

It is proved that any molecular graph invariant (that is any topological index) can be uniquely represented as (1) a linear combination of occurrence numbers of some substructures (fragments), both connected and disconnected, or (2) a polynomial on occurrence numbers of connected substructures of corresponding molecular graph. Besides, any (0,1)-valued molecular graph invariant can be uniquely represented as a linear combination (in the terms of logic operations) of some basic (0,1)-valued invariants indicating the presence of some substructures in the chemical structure. Thus, the occurrence numbers of substructures in a structure (or numbers indicating the presence or absence of substructures in a structure for the case of (0,1)-valued invariants) are shown to constitute the basis of invariants of labeled molecular graphs. A possibility to use these results for the mathematical justification of substructures-based methods in the "structure−property" problem is also discussed.

INTRODUCTION

The search for structure−property relationships is an important problem of contemporary chemistry, and the methods of molecular description play an essential role in these investigations. One of the most popular approaches to the solution of this problem is based on representing molecular structure as a weighted molecular graph and using graph invariants (called also topological indices, TIs[1−6]) for its characterization.

It should be noted that there exists an infinite number of ways, by which a molecule could be described in terms of TIs. The use of various molecular graphs or graph invariants for the same structure makes it possible to design different sets of TIs. Therefore, in the search for "structure−property" correlations the problem of justified choice of TIs and type of functional dependence of a property on TIs in relationship "TIs−property" usually appears.

However, the justified choice of TIs is not always possible. The main reason for that is that TIs are usually constructed using refined mathematical operations with graphs, and, therefore, it is difficult to interpret them in a framework of some physical or chemical theory and to relate unambiguously with the property under consideration.

The following question appears: does there exist a finite set of basic graph invariants, such that any invariant could be uniquely expressed as a linear combination of these basic invariants? If such a set exists then its elements form a finite basis in the algebra of these invariants (the set of graph invariants with operations of addition, multiplication, and multiplication on real number forms an algebra). Therefore, one can choose TIs from this basis and use only a linear kind of functional dependence in the search for "TIs−property" relationships.

The problem of finding a set of graph invariants and basic subgraphs was considered by Randić[7] in 1992. It was shown that its solution would make it possible to represent chemical structures unambiguously and to get criteria for similarity and dissimilarity of chemical structures. It was suggested to use path graphs and their subgraphs as basic subgraphs, while their occurrence numbers in a structure could be used to codify it. However, it was shown by some examples that different structures could contain the same set of such basic subgraphs. Therefore, such basic subgraphs could not be considered as basic in common sense of the word.

However, we have to point out that a rigorous solution of the problem of finding a set of graph invariants was obtained in 1983 for the case of simple graphs[8] but, being published in Russian, seems to remain unknown. Let $\Gamma^{(n)}$ be the set of all simple (that is, with nonweighted vertices and edges), both connected and disconnected, graphs with n vertices. It was proven by methods of commutative algebra[8] that any invariant $f(G)$ of a graph $G \in \Gamma^{(n)}$ is uniquely represented in the form

$$f(G) = \sum_j c_j g_j(G)$$

where c_j denotes some constants, independent on G, $g_j(G)$ is the occurrence number of a graph $G_j \in \Gamma^{(n)}$ in G (that is the number of different subgraphs of G which are isomorphic to G_j), and the sum runs over all graphs $G_j \in \Gamma^{(n)}$. This means, that the set $\{g_j\}$ is the basis in algebra of invariants of graphs from $\Gamma^{(n)}$. Besides, any invariant of graph $G \in \Gamma^{(n)}$ is determined by the numbers of subgraphs in G constructed from G by deleting edges in different possible nonequivalent ways.

However, invariants of vertex- and edge-weighted graphs are of great importance for different problems of chemistry. The weights of vertices and edges of such graphs are

[⊗] Abstract published in *Advance ACS Abstracts,* February 15, 1995.

0095-2338/95/1635-0527$09.00/0

determined by the types of corresponding atoms and bonds. A weighted graph reflects the features of molecular structure more completely than a simple one.

Partitioning atoms and bonds in molecule into some classes and ascribing to each class a label (some symbol), one can get a labeled molecular graph. Ascribing to labels considered as parameters some numerical values (weights), one can get a weighted graph from a labeled one. Thus, it is natural to consider a weighted graph as a particular case of a labeled one.

Thus, the investigation of algebra of invariants of labeled graphs is of importance for the development of both graph theory and mathematical chemistry.

There are a number of approaches in QSAR called logic-structural or logic-combinatorial.[9,10] These or related approaches are mainly used for carrying out structure−activity studies for the case of such biological activity that may take only two values: "1" (active compounds) and "0" (nonactive compounds). Thus, in these approaches, only graph invariants $f(G)$ taking values (0, 1) are considered. For the molecular description, a set of subgraphs $\{G_j\}$ is chosen, and some simple invariants called indicator variables

$$g_j(G) = \begin{cases} 1, & G_j \in G \\ 0, & G_j \notin G \end{cases}$$

are considered. The property can be expressed as some logic function $f(G)$ on $\{g_j(G)\}$, in terms of logic operations called conjunction, disjunction, and negation. Thus, each approach is defined by the set $\{G_j\}$ and the kind of logic function. There are evident analogies between the above described methods and classical ones, based on TIs, when a set of TIs and the kind of function on these TIs are also specified. For the case of logic-structural (logic-combinatorial) approaches the problem of searching for a basis of invariants (so any (0, 1)-valued invariant is uniquely expressed by basis (0,1)-valued invariants in terms of logic operations) also appears.

In this paper, a basis of invariants for labeled graphs both for algebra of arbitrary invariants and for an algebra of the (0, 1)-valued invariants is found. Some examples are presented. A general model of structure−property relationship is also constructed.

PRINCIPAL RESULTS: THREE THEOREMS ON THE BASIS OF GRAPH INVARIANTS

Construct the following set of labeled graphs.

Consider the set of simple graphs $\Gamma^{(n)}$ and two finite sets of arbitrary labels (symbols), $V = \{v_1,...,v_{p_1}\}$, $E = \{e_1,...,e_{p_2}\}$, $v_i \neq v_j$, $e_i \neq e_j$, for $i \neq j$. Place labels on vertices (from V) and edges (from E) of graphs of $\Gamma^{(n)}$ using all nonequivalent ways. Denote by $H_{V,E}^{(n)}$, the set of constructed nonisomorphic vertex- and edge-labeled graphs, and by N, the number of elements in $H_{V,E}^{(n)}$. It is also possible that in graphs of $\Gamma^{(n)}$ only vertices ($E = \varnothing$ is an empty set) or only edges ($V = \varnothing$ is an empty set) are labeled. Denote sets of such graphs by $H_V^{(n)}$ and $H_E^{(n)}$, respectively.

Let us consider the labels as variables which can take real values. Then any graph $H \in H_{V,E}^{(n)}$ is represented by symmetric matrix $\mathbf{A} = (a_{ij})$, where element a_{ii} is equal to the label of vertex i; a_{ij} ($i \neq j$) is equal to zero for nonadjacent vertices i and j and equal to the label of the edge (i,j) for adjacent vertices i and j.

Definition. An invariant of labeled graph $H \in H_{V,E}^{(n)}$ is a scalar function on elements of matrix $\mathbf{A}$ independent on the way of numbering the graph vertices.

The following theorem 1 is true.

Theorem 1. Any invariant $f(H)$ ($H \in H_{V,E}^{(n)}$) is uniquely represented in the form

$$f(H) = \sum_{j=1}^{N} c_j g_j(H) \qquad (1)$$

where c_j are some constants independent on H and dependent on f, $g_j(H)$ are the occurrence numbers of a graph $H_j \in H_{V,E}^{(n)}$ in the graph H (that is the number of different subgraphs of H which are isomorphic to H_j). Thus, the set $\{g_j\}$ is the basis in algebra of invariants of graphs from $H_{V,E}^{(n)}$. Besides, the value of any invariant $f(H)$ for a graph H is determined by the numbers of subgraphs in H constructed by deleting edges in H in all nonequivalent ways.

Proof. Order the graphs from $H_{V,E}^{(n)}$ in the following way. Firstly, enumerate arbitrarily all graphs with $n(n-1)/2$ edges; secondly, all graphs with $([n(n-1)/2] - 1)$ edges, etc., until the graphs consisting of isolated vertices. Denote by B the square matrix with elements $b_{ij} = g_j(H_i)$, $(i,j = \overline{1,N})$. Evidently, (1) if graphs H_i and H_j have the same number of edges, then $b_{ij} = g_j(H_i) = b_{ji} = g_i(H_j) = 0$, and $b_{jj} = g_j(H_j) = 1$ and (2) if graphs H_i and H_j have different number of edges and $j < i$, then $g_j(H_i) = 0$. Thus, the matrix B is a triangular matrix; its diagonal elements are equal to units; under them are placed zeroes. Therefore, there exists an inverse matrix B^{-1}. Write the system of equations

$$f(H_i) = \sum_{j=1}^{N} c_j g_j(H_i) = \sum_{j=1}^{N} b_{ij} c_j \qquad (i = \overline{1,N}) \qquad (2)$$

or, in matrix form, $\bar{f} = B\bar{c}$, where $\bar{f} = (f(H_1),...,f(H_N))$, $\bar{c} = (c_1,...,c_N)$ are column vectors. This system 2 always has a unique solution, $\bar{c} = B^{-1}\bar{f}$. Therefore, there exists a unique decomposition 1 of an invariant $f(H)$ for the given numbering of graphs H_j.

Show that expansion 1 does not depend on the numbering of graphs H_j. Suppose that some other numbering leads to vectors $\bar{f}'$, $\bar{c}'$, and matrix B' (not necessarily triangular). The transition from the first numbering to the second one is achieved using the permutation π: $j \to \pi(j)$ ($j = \overline{1,N}$) or the corresponding square $N \times N$ permutation matrix X, $\det X \neq 0$. Evidently $X\bar{f} = \bar{f}'$, $X\bar{c} = \bar{c}'$, and $XBX^{-1} = B'$. As we have proven, for the special numbering described above expansion 1 $\bar{f} = B\bar{c}$ is true. Multiplying both parts of this equation by matrix X, we get

$$\bar{f}' = X\bar{f} = (XBX^{-1})(X\bar{c}) = B'\bar{c}'$$

Therefore, expansion 1 is true for any numbering.

Theorem 1 is proven.

Theorem 2. Any graph invariant $f(H)$ ($H \in H_{V,E}^{(n)}$) is represented as a polynomial on variables which are equal to the occurrence numbers of some connected subgraphs of H. The numbers of vertices in these subgraphs and the degree of the polynomial are less or equal to n.

Proof. Firstly, show that the occurrence number of any nonconnected subgraph C in a graph H is expressed by the occurrence numbers of some connected subgraphs of H.

Suppose that C consists of k components of connectedness, that is $C = \cup_{i=1}^{k} C_i$, where $\{C_i\}$ are connected subgraphs and $C_i \cap C_j = \varnothing$, $i \neq j$. In the general case, it is possible that some $\{C_i\}$ are isomorphic subgraphs. Suppose that $\{C_i\}$ are subdivided into p groups Ω_i ($i = \overline{1,p}$), so subgraphs in each group are isomorphic one to another, but subgraphs of different groups are nonisomorphic; m_i are the numbers of elements in Ω_i, $m_i \geq 1$, $\sum m_i = k$, and $i = \overline{1,p}$. Enumerate $\{C_i\}$ in the following way: firstly $\{C_i\}$ of Ω_1 are enumerated; secondly, $\{C_i\}$ of Ω_2, etc. Let M_i be the set of all subgraphs of graph H, which are isomorphic to subgraphs of group Ω_i; l_i are the numbers of elements in M_i ($i = \overline{1,p}$). Evidently, $l_i \geq m_i$.

Construct a new subgraph of graph H, choosing in any possible way m_i different elements of M_i simultaneously for $i = \overline{1,p}$. The number of such subgraphs is equal to $\prod_{i=1}^{p} C_{l_i}^{m_i}$, $C_{l_i}^{m_i} = l_i!/[m_i!(l_i - m_i)!]$. Subgraphs constructed from $\{M_i\}$ may be of two kinds, in which the initial subgraphs belonging to $\{M_i\}$, are (1) nonintersecting and (2) intersecting. Denote by t_1 and t_2 the numbers of subgraphs of the first and the second kind, correspondingly. Evidently, $t_1 + t_2 = \prod_{i=1}^{p} C_{l_i}^{m_i}$. Note that t_1 is equal to the occurrence number of subgraph C in H and coincides, according to the definition, with the number of subgraphs in H which are isomorphic to C. Besides, subgraphs of the second kind have less than k components of connectedness, and the sum $t_1 + t_2 = \prod_{i=1}^{p} C_{l_i}^{m_i}$ is a polynomial of degree $k = \sum m_i$ on variables $l_i (i = \overline{1,p})$.

Thus, the occurrence number t_1 of nonconnected subgraph C with k components of connectedness is expressed by the occurrence numbers of its connected components and some subgraphs with less than k components of connectedness. Applying many times this result to all disconnected subgraphs in theorem 1, we obtain the statement of theorem 2.

Theorem 2 is proven.

Now let us turn to the logic-structural (logic-combinatorial) approaches in structure–property relationship studies and remember some definitions and statements in mathematical logic.[11]

The set of functions $\{f(x_1,...,x_n)\}$, where $x_i \in (0,1)$ ($i = \overline{1,n}$), $f(x_1,...,x_n) \in (0,1)$, is called an algebra of logic (Boolean algebra). Denote by A this algebra. The basic functions of A are $\neg$ (negation), $\cdot$ (or $\wedge$, conjunction), $\vee$ (disjunction), $+$ (sum), and 1 (identity):

$$\neg x = \begin{cases} 0, & x = 1 \\ 1, & x = 0 \end{cases}$$

$$x \cdot y = x \wedge y = \begin{cases} 1, & x = y = 1 \\ 0, & \text{in other cases} \end{cases}$$

$$x \vee y = \begin{cases} 0, & x = y = 0 \\ 1, & \text{in other cases} \end{cases}$$

$$x + y = \begin{cases} 0, & x = y \\ 1, & x \neq y \end{cases}$$

$$1(x) = x$$

The system of function, B, is called a complete one, if any function $f \in A$ is expressed as superposition of functions from B. It is known that the systems of functions $\{\wedge,\vee,\neg\}$ of $(+,\cdot,1\}$ are complete ones. Besides, any $f \in A$ is uniquely represented as

$$\sum_{\substack{k \geq 0 \\ 1 \leq i_1 < ... < i_k \leq n}} c_{i_1,...,i_k} x_{i_1} \cdot ... \cdot x_{i_k}, \qquad c_{i_1,...,i_k} \in (0,1)$$

In the logic-structural approaches the system $\{\wedge,\vee,\neg\}$ is usually used. However, it is possible (and, in our opinion, it is more convenient) to use the system $\{+,\cdot,1\}$, as it provides the analogy with the case of arbitrary graph invariants with standard mathematical operations of addition and multiplication.

Consider now the set of (0,1)-valued invariants $\{f(H)\}$ ($H \in H_{V,E}^{(n)}$) with operations of addition ($+$), multiplication ($\cdot$) and multiplication by numbers from the field ($\times$), similar to the logic operations described above. Then the set $\{f(H)\}$ is also an algebra. Denote by

$$x_j = g_j(H) = \begin{cases} 1, & H_j \in H \\ 0, & H_j \notin H \end{cases}$$

Identify graph H with the vector $\{g_j(H)\}_1^N$. Then $f(H) \equiv f(x_1,...,x_N)$, and the set of such functions will be a subset in Boolean algebra A.

Theorem 3. Any (0,1)-valued invariant $f(H)$ ($H \in H_{V,E}^{(n)}$) is uniquely represented as

$$f(H) = \sum_{j=1}^{N} c_j g_j(H)$$

where $c_j \in (0,1)$ are some constants depending on f only. Therefore, $\{g_i(H)\}$ is a basis in the set of described above invariants.

Proof. The proof of this theorem is similar to that of theorem 1. In this case for matrix $B = (b_{ij})$: $b_{ij} \in (0,1)$. The proof is also based on the fact that there exists an inverse matrix B^{-1}: $BB^{-1} = B^{-1}B = E$ (E is identity matrix); the addition and multiplication of elements B and B^{-1} is carried out as in Boolean algebra. For the proof of the existence of B^{-1}, it is necessary to proove that the system of equations $B\bar{x} = \bar{y}$ ($\bar{x},\bar{y}$ are vectors with components $x_k, y_k \in (0,1)$, $k = \overline{1,N}$) has a unique solution $\bar{x}$ for any given $\bar{y}$. However, the system with triangular matrix B always has a unique solution, which is defined by the method of sequential elemination of unknown variables. Indeed, on each step of this procedure, an equation of the type

$$x_k + a = y_k$$

is being solved, for some k ($1 \leq k \leq N$) and a constant a. Evidently, for a given a and $y_k \in (0,1)$ this equation has the unique solution, $x_k \in (0,1)$.

Theorem 3 is proven.

EXAMPLES

Example 1. Let $n = 3$, $V = (v_1,v_2)$. The set of vertex-labeled graphs $H_V^{(3)}$ is given in Figure 1.

Each graph $H_k (k = \overline{1,20})$ corresponds to a square symmetric matrix $A^{(k)} = (a_{ij}^{(k)})$; a_{ii} is equal to the label of vertex i; $a_{ij} = 0$ or $a_{ij} = 1$ for nonadjacent and adjacent vertices i

and $j(i \neq j)$, respectively. Consider the graph invariant

$$f(H) = \sum_{i \leq j} a_{ij}\sqrt{a_{ii}a_{jj}}, \quad H \in H_V^{(3)}$$

Evidently, this invariant is a generalization of the Randic[12] index $\chi = \sum_{edges(i,j)}(d_i d_j)^{-1/2}$ defined for simple graphs; it can be turned to χ if we take $a_{ii} = d_i^{-1}$ (d_i is degree of vertex i. Calculate the occurrence numbers of H_j in H_i ($i,j = \overline{1,20}$) and form the following system of eq 3

$$f(H_i) = \sum_{j=1}^{20} c_j g_j(H_i), \quad (i = \overline{1,20}) \tag{3}$$

System 3 consists of 20 equations with 20 unknown variables:

$$f(H_1) = 3v_1 = c_1 + 3c_5 + 3c_{11} + c_{17}$$

$$f(H_2) = 3v_2 = c_2 + 3c_9 + 3c_{15} + c_{18}$$

$$f(H_3) = v_2 + 2\sqrt{v_1 v_2} = c_3 + c_6 + 2c_{10} + c_{12} + 2c_{16} + c_{19}$$

$$f(H_4) = v_1 + 2\sqrt{v_1 v_2} = c_4 + 2c_7 + c_8 + 2c_{13} + c_{14} + c_{20}$$

$$f(H_5) = 2v_1 = c_5 + 2c_{11} + c_{17}$$

$$f(H_6) = 2\sqrt{v_1 v_2} = c_6 + 2c_{16} + c_{19}$$

$$f(H_7) = v_1 + \sqrt{v_1 v_2} = c_7 + c_{13} + c_{14} + c_{20}$$

$$f(H_8) = 2\sqrt{v_1 v_2} = c_8 + 2c_{13} + c_{20}$$

$$f(H_9) = 2v_2 = c_9 + 2c_{15} + c_{18}$$

$$f(H_{10}) = v_2 + \sqrt{v_1 v_2} = c_{10} + c_{12} + c_{16} + c_{19}$$

$$f(H_{11}) = v_1 = c_{11} + c_{17}$$

$$f(H_{12}) = v_2 = c_{12} + c_{19}$$

$$f(H_{13}) = \sqrt{v_1 v_2} = c_{13} + c_{20}$$

$$f(H_{14}) = v_1 = c_{14} + c_{20}$$

$$f(H_{15}) = v_2 = c_{15} + c_{18}$$

$$f(H_{16}) = \sqrt{v_1 v_2} = c_{16} + c_{19}$$

$$f(H_{17}) = f(H_{18}) = f(H_{19}) = f(H_{20}) = 0 = c_{17} = c_{18} = c_{19} = c_{20}$$

Solving this system, we obtain $c_i = 0$, $i = \overline{1,10}$, $c_{11} = c_{14} = v_1$, $c_{12} = c_{15} = v_2$, and $c_{13} = c_{16} = \sqrt{v_1 v_2}$.

Thus,

$$f(H) = v_1 g_{11}(H) + v_2 g_{12}(H) + \sqrt{v_1 v_2} g_{13}(H) + v_1 g_{14}(H) + v_2 g_{15}(H) + \sqrt{v_1 v_2} g_{16}(H)$$

This means that the invariant $f(H)$ is represented as a linear combination of six parameters which are equal to the occurrence numbers of $H_{11} - H_{16}$ in the initial graph H; the coefficients in this expansion depend on parameters v_1 and v_2.

Example 2. Calculate the occurrence number of subgraph C consisting of two components of connectedness, in graph H, given in Figure 2.

In this case $m_1 = 1$, $m_2 = 1$, $p = 2$, $l_1 = 2$, $l_2 = 2$, $t_1 = 2$, and $t_2 = 2$; sets M_1, M_2, and subgraphs of the first and second kind are shown in Figure 3.

So, $t_1 + t_2 = \prod_{i=1}^{2} C_{l_i}^{m_i} = (C_2^1)^2 = 4$, and the occurrence number of nonconnected subgraph C in graph H is expressed by the occurrence numbers of connected subgraphs presented in Figure 4.

A GENERAL MODEL OF STRUCTURE−PROPERTY RELATIONSHIP

Suppose that the training set of chemical compounds represented as labeled graphs $\{H_i\}$ ($i = \overline{1,N_1}$) with property values $\{y_i\}$ are given. Let n_i be the number of vertices in H_i ($i = \overline{1,N_1}$), $n = \max n_i$. Add to each H_i ($n - n_i$) isolated vertices, so the resulted graph (denote it again by H_i) will have n vertices. Suppose that all isolated vertices have any label not used for labeling vertices in initial chemical graphs. Suppose that graphs $\{H_i\}$ are enumerated in the way described in the proof of the theorem 1. Then matrix $\tilde{B} = (g_i(H_j))(i,j = \overline{1,N_1})$ will have inverse matrix $\tilde{B}^{-1}$. Denote by $\{H_i\}(i = \overline{N_1 + 1,N})$ all subgraphs (on n vertices) of graphs H_i ($i = \overline{1,N_1}$) constructed from these graphs by deleting edges in all nonequivalent ways.

In structure−property relationship studies, it is postulated that a property "y" is a function of a chemical structure. If a molecule is represented as a labeled graph, then "y" is some graph invariant, $y = f(H)$. According to theorem 1

$$f(H) = \sum_{i=1}^{N} c_i g_i(H) \tag{4}$$

Form the following system of equations using initial data, $y_j = f(H_j)(j = \overline{1,N_1})$

$$y_j = \sum_{i=1}^{N} c_i g_i(H_j) \tag{5}$$

and try to solve it, that is to find $\{c_i\}$. For the solution of this system (where the number N_1 of equations is always less than the number N of unknown variables $\{c_i\}$) it is necessary to choose the principal and free unknown variables. The first variables are $c_1,\ldots,c_{N_1}$, as $\det \tilde{B} \neq 0$, and the second ones are $c_{N_1+1},\ldots,c_N$. Denote by $\bar{y}$, $\bar{c}$, and $\bar{a}$ the column-vectors with components y_j, c_j, $\sum_{i=N_1+1}^{N} c_i g_i(H_j)$ ($j = \overline{1,N_1}$), respectively. Then system 5 can be written in the following form: $\bar{y} = \tilde{B}\bar{c} + \bar{a}$. Its unique solution is $\bar{c} = \tilde{B}^{-1}\bar{y} - \tilde{B}^{-1}\bar{a}$. Substituting these parameters $c_1,\ldots,c_{N_1}$, which are expressed by $y_1,\ldots,y_{N_1}$ and $c_{N_1+1},\ldots,c_N$, in (4), we obtain a general mathematical model of structure−property relationship constructed on some training set of compounds. This model depends on parameters $c_{N_1+1},\ldots,c_N$, which cannot be determined from the initial data. Any other model is a particular case of the general one, with some values of parameters

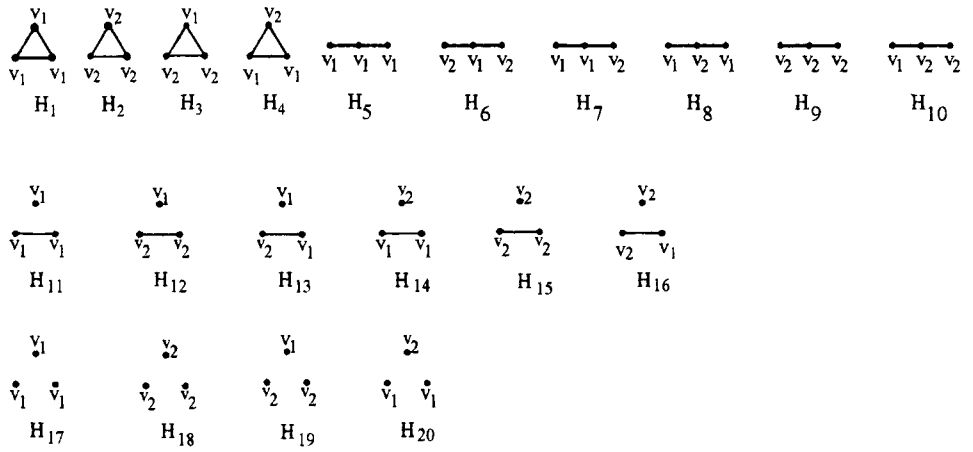

Figure 1. The set of vertex-labeled graphs $H_V^{(3)}$.

Figure 2. The graph H and its subgraph C.

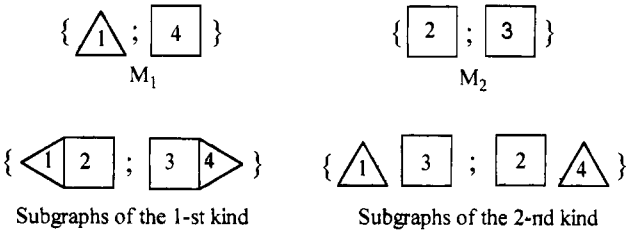

Figure 3. The sets M_1 and M_2 and subgraphs of the first and the second kind.

Figure 4. Connected subgraphs used for the calculation of the occurrence number of subgraph C in graph H.

$c_{N_1+1},\ldots,c_N$. So, we construct *all theoretically possible* models that *exactly* describe the structure−property relationship.

It should be noted that for any structure−property model a question about its predictive power arises. However, before studying the predictive performance of a model, it is necessary to solve the principle theoretical problem of definition of its area of application: predictions should be made only for compounds taken from such an area. In this paper we do not touch upon these questions; they will be thoroughly considered in a future publication.

CONCLUSION

In the present paper it is proved that any graph invariant (that is, any physical or chemical property or quantitatively defined biological activity) can be uniquely represented as (1) a linear combination of the occurrence numbers of some substructures (fragments), both connected and disconnected, or (2) a polynomial on occurrence numbers of connected substructures. Besides, any (0,1)-valued graph invariant may be uniquely represented as a linear combination (in the sense of logic operations) of some basic (0,1)-valued invariants. In all cases, the set of some subgraphs is used for the complete description of graph structure (that is, molecular structure).

It also follows from the proven theorems that different graph invariants (that is, TIs) differ one from another by choosing the coefficients $\{c_i\}$ in expansion of TIs on the basis.

Besides, one can consider the results discussed as the strict mathematical justification of the use of additive methods for calculating different physical−chemical properties and biological activity of organic compounds.

REFERENCES AND NOTES

(1) Stankevich, M. I.; Stankevich, I. V.; Zefirov, N. S. Topological Indexes in Organic Chemistry. *Russ. Chem. Rev.* **1988**, *57*, 191−208.
(2) Rouvray, D. H. Should We Have Designs on Topological Indexes? In *Chemical Applications of Topology and Graph Theory*; King, R. B., Ed., Elsevier: Amsterdam, 1983; pp 159−177.
(3) Balaban, A. Chemical Graphs. XXXIV. Five New Topological Indices for the Branching of Tree-like Graphs. *Theor. Chim. Acta* **1979**, *53*, 355−375.
(4) Seybold, P. G.; May, M.; Bagal, U. A. Molecular Structure-Property Relationships. *J. Chem. Educ.* **1987**, *64*, 575−581.
(5) Randić, M. Generalized Molecular Descriptors. *J. Math. Chem.* **1991**, *7*, 155−168.
(6) Rouvray, D. H. Predicting Chemistry from Topology. *Sci. Am.* **1986**, *254*, 40−47.
(7) Randić, M. Representation of Molecular Graphs by Basic Graphs. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 57−69.
(8) Mnukhin, V. B. Basis of Algebra of Graph Invariants. In: *Mathematical Analysis and its Applications*, Rostov-na-Donu, 1983; pp 55−60 (in Russian).
(9) Kadyrov, Ch. Sh.; Tyurina, L. A.; Simonov, V. D.; Semenov, V. A. *Computer Search for Chemical Compounds with Predefined Properties*; Fan: Tashkent, 1989 (in Russian).
(10) Rosenblith, A. V.; Golender, V. E. *Logic-Combinatorial Methods in Drug Design*; Zinatne: Riga, 1983 (in Russian).
(11) Lavrov, I. A.; Maximova, L. L. *Tasks on the set theory, mathematical logic and algorithm theory*; Nauka: Moscow, 1975 (in Russian).
(12) Randić, M. On characterization of molecular branching. *J. Am. Chem. Soc.* **1975**, *37*, 6609−6615.

CI940119P