Molecular Similarity. 1. Analytical Description of the Set of Graph Similarity Measures

Mariya I. Skvortsova and Igor I. Baskin

N. D. Zelinsky Institute of Organic Chemistry, Leninsky Prosp., 47, Moscow 117813, Russia

Ivan V. Stankevich

A. N. Nesmeyanov Institute of Organoelement Compounds, Vavilov Str. 28, Moscow, 117813, Russia

Vladimir A. Palyulin and Nikolai S. Zefirov*

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

Received May 21, 1997

The elaboration of methods for defining molecular similarity measures is one of the important fields of modern theoretical chemistry. These measures are used for solving a number of problems of theoretical and computer chemistry, in particular the prediction of properties of chemical compounds. For the construction of any molecular similarity measures molecules are represented as some mathematical objects $\{M\}$, on which quantitative similarity measures $d(M_1, M_2)$ $(M_1, M_2 \in \{M\})$ are introduced. The most widely used way of molecular representation is based on picturing molecules as labeled graphs, labels of which encode types of atoms and bonds. There are many different similarity measures defined for graphs, expressed in terms of vectors of graph invariant, sequences, and sets derived from graphs or in terms of maximal common subgraph, etc. In general, there exists an infinite number of graph similarity measures. In the present paper an analytical description of the set of symmetric similarity measures defined for arbitrary labeled graphs is given. The found general formula for the measure depends on a number of parameters satisfying some conditions. Any particular graph similarity measure may be obtained from this formula at definite values of parameters.

INTRODUCTION

The elaboration of methods for defining molecular similarity measures is one of the important fields of modern theoretical chemistry. These measures are used for solving a number of problems of theoretical and computer chemistry, in particular, the prediction of properties of chemical compounds, in reactivity theory, in chemical data bases, etc.[1,2]

For the construction of any molecular similarity measure molecules are represented as some mathematical objects. There exist many ways of such representation: for example, one can describe a molecule by means of a finite set, finite sequence, vector, labeled graph, scalar function, matrix of interatomic distances, etc.[1] Further, on a constructed set $\{M\}$ of mathematical objects quantitative similarity measures $d(M_1, M_2)$ for elements M_1 and M_2 are introduced which are interpreted as quantitative similarity measures of corresponding molecules. As a rule, $d(M_1, M_2)$ is a symmetric function, *i.e.*, $d(M_1, M_2) = d(M_2, M_1)$ (recently asymmetric similarity measures were suggested[3]). The chemical structures are usually represented by means of vertex and edge labeled graphs; the labels encode atoms and bonds of different chemical nature. More simple molecular representation can be derived from the above mentioned graph representation, and measures defined for corresponding mathematical objects may be considered as measures on graphs. Note, that using different labels on graphs, one can obtain more sophisticated molecular representations reflecting some features of the spatial and electronic structure of a molecule.

There are known many graph similarity measures.[1−8] According to the review,[1] we shall describe some of them.

Let graphs G_1 and G_2 be represented by vectors $(x_1, ..., x_m)$ and $(y_1, ..., y_m)$; x_i, y_i being any graph invariants. In this case one can define the following similarity measures

$$d_1(G_1, G_2) = \sum_i |x_i - y_i|$$

$$d_2(G_1, G_2) = (\sum_i (x_i - y_i)^2)^{1/2} \text{ (“Euclidean distance”)}$$

$$d_3(G_1, G_2) = \frac{\sum_i (x_\mathrm{i} - \bar{x})(y_\mathrm{i} - \bar{y})}{(\sum_i (x_\mathrm{i} - \bar{x})^2)^{1/2} (\sum_i (y_\mathrm{i} - \bar{y})^2)^{1/2}}$$

where

$$\bar{x} = \frac{1}{m}\sum_i x_i, \quad \bar{y} = \frac{1}{m}\sum_i y_i$$

$$d_4(G_1, G_2) = \sum_i (x_i - y_i)\log_2(x_i/y_i)$$

Published on Web 07/23/1998

$$d_5(G_1, G_2) = \frac{\sum_i x_i y_i}{\sum_i x_i^2 + \sum_i y_i^2 - \sum_i x_i y_i}$$

(it is supposed for $d_4(G_1, G_2)$ that $x_i > 0$, $y_i > 0$, $\sum_i x_i = 1$, $\sum_i y_i = 1$).

Let graphs G_1, G_2 be represented by sets D_1 and D_2 (for example, D_1 and D_2 are sets of definite structural fragments). In this case, one can define the following similarity measures

$$d_6(G_1, G_2) = |D_1| + |D_2| - 2|D_1 \cap D_2|$$

$$d_7(G_1, G_2) = |D_1 \cap D_2|^2/|D_1| \cdot |D_2|$$

where the symbol $|D|$ denotes the total number of elements in D.

A number of similarity measures can be defined in terms of the maximal common subgraph of graphs G_1 and G_2, MCS(G_1, G_2)

$$d_8(G_1, G_2) = |G_1| + |G_2| - 2|\mathrm{MCS}(G_1, G_2)|$$

$$d_9(G_1, G_2) = (|\mathrm{MCS}(G_1, G_2)|)/(|G_1| \cdot |G_2|)$$

where the symbol $|G|$ denotes the total number of vertices and edges in a graph G. One can also consider the minimal common supergraph and construct the analogous measures.

Let graphs G_1 and G_2 be represented by sequences S_1 and S_2, elements of which are symbols or numbers. In this case, one can define the following measures

$$d_{10}(G_1, G_2) = |S_1| + |S_2| - 2|\mathrm{MCS}(S_1, S_2)|$$

$$d_{11}(G_1, G_2) = (|\mathrm{MCS}(S_1, S_2)|)^2/(|S_l| \cdot |S_2|)$$

$$d_{12}(G_1, G_2) = (1 - d_{11}(G_1, G_2))/(|S_1| + |S_2|)$$

where MCS(S_1, S_2) is the maximal common subsequence of S_1 and S_2; the symbol $|S|$ denotes the length of the sequence S. Evidently, there exists an infinite number of graph similarity measures. For example, one can construct new measures taking products of any two earlier suggested measures $d_i(G_1, G_2)$ and $d_j(G_1, G_2)$ or taking $d(G_1, G_2) = f(d_j(G_1, G_2))$ where $f(x)$ is some function on variable x. Thus, novel similarity measures can be easily constructed.

In our opinion, *the first problem* in these studies consists in establishing a general analytical formula for any graph similarity measure so that any particular measure could be obtained from it. *The second problem* is to select an appropriate measure for a particular task using the above mentioned general formula.

In the present paper, an analytical description of the set of symmetric similarity measures defined for labeled graphs is given. The general formula established for the measure depends on a number of parameters satisfying some conditions. Any particular graph similarity measure may be obtained from this formula at definite values of parameters. The second problem dealing with chemical application of these theoretical results will be considered in our next paper of this series.

MAIN RESULTS

Let G_i ($i = 1,..., N$) be some set of vertex and edge labeled graphs with arbitrary labels, and H_j ($j = 1,..., N$)—some set of subgraphs of these graphs. Denote by $b_{ij} = g_j(G_i)$ the occurrence numbers of H_j into G_i, by g_j—the corresponding graph invariants, and by $B = (b_{ij})$ ($i,j = 1,..., N$)—the corresponding matrix.

Theorem 1. (1) There exists such a set of subgraphs Hj ($j = 1,..., N$) of the given graphs G_i ($i = 1,..., N$), so that

$$\det B \neq 0 \tag{1}$$

(2) The corresponding invariants g_j ($j = 1,..., N$) form a finite basis of invariants of graphs G_i ($i = 1,..., N$), that is any invariant f of these graphs is uniquely represented in the following form

$$f(G) = \sum_{j=1}^{N} a_j g_j(G) \tag{2}$$

where a_j are some constants independent on G and dependent on f;

(3) In formula (2) one can take any basis $\bar{f} = (f_1, ..., f_N)$ obtained from the vector-column $\bar{g} = (g_1, ..., g_N)$ by means of transformation $\bar{f} = A\bar{g}$, where A is any square ($N \times N$) matrix, so that $\det A \neq 0$;

(4) There exists such a basis $\bar{f} = (f_1, ..., f_N)$ (that is, matrix A), so that

$$f_N(G_1) = f_N(G_2) = ... = f_N(G_N) = 1 \tag{3}$$

$$f_1(G_N) = f_2(G_N) = ... = f_{N-1}(G_N) = 0 \tag{4}$$

In this basis

$$f(G) = \sum_{j=1}^{N\text{-}1} a_j f_j(G) + a_N$$

Theorem 2. Let $d(G_k, G_l)$ be any symmetric similarity measure on the graphs G_i ($i = 1,..., N$; k, $l = 1,..., N$; $d(G_k, G_k) = 0$), $\bar{f} = (f_1, ..., f_N)$—a basis of invariants of these graphs satisfying (3) and (4), $\bar{f}_k = (f_1(G_k), ..., f_{N-1}(G_k))$, $\bar{f}_l = (f_1(G_l), ..., f_{N-1}(G_l))$ are vector-columns.

Then there exists a unique symmetric ($N-1$) × ($N-1$) matrix $M = (m_{ij})$ ($i, j = 1,..., N$-1), so *that the measure* $d(G_k, G_l)$ is represented in the following form

$$d(G_k, G_l) = \sum_{i,j}^{N\text{-}1} (f_i(G_k) - f_i(G_l))(f_j(G_k) - f_j(G_l))m_{ij} = (M(\bar{f}_k - \bar{f}_l), \bar{f}_k - \bar{f}_l) \tag{5}$$

where the right part is a scalar product of vectors $M(\bar{f}_k - \bar{f}_l)$ and $\bar{f}_k - \bar{f}_l$, and $M(\bar{f}_k - \bar{f}_l)$ is a product of matrix M and vector $\bar{f}_k - \bar{f}_l$.

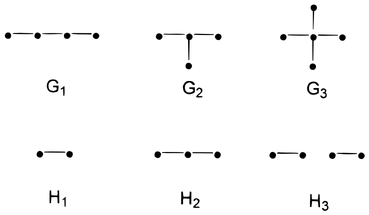
Remarks. (1) Theorem 1 (sections (1), (2), and (3)) was proven earlier[9,10] by us for the particular case when graphs G_i ($i = 1,..., N$) are all graphs with the same number of vertices n, and $H_i = G_i$ ($i = 1,..., N$).

(2) The proofs of Theorem 1 and 2 and algorithms for searching for vector $\bar{a} = (a_1, ..., a_N)$ and matrix M are given in Appendix.

(3) It is not difficult to show that the measure $d(G_k, G_l)$ is a metric if and only if M is a positive definite matrix.[11] The proof of this statement is based on the well-known fact that any measure $d(\bar{x},\bar{y})$ ($\bar{x},\bar{y} \in R^n$ − n-dimensional Euclidean space) is a metric if and only if $d(\bar{x},\bar{y})$ can be represented in the following form: $d(\bar{x},\bar{y}) = (M(\bar{x} - \bar{y}), \bar{x} - \bar{y})$ with some positive definite $n \times n$ matrix M.

EXAMPLES

Example 1 (to Theorem 1). Consider the graphs G_1, G_2, G_3 and their subgraphs H_1, H_2, H_3:



Then

$$g_1(G_1) = 3, \quad g_1(G_2) = 3, \quad g_1(G_3) = 4$$

$$g_2(G_1) = 2, \quad g_2(G_2) = 3, \quad g_2(G_3) = 6$$

$$g_3(G_1) = l, \quad g_3(G_2) = 0, \quad g_3(G_3) = 0$$

$$B = \begin{pmatrix} 3 & 2 & 1 \\ 3 & 3 & 0 \\ 4 & 6 & 0 \end{pmatrix}, \quad \det B = 6 \neq 0$$

Formula (2) for some graph invariant f in this case is

$$f(G) = \sum_{j=1}^{3} a_j g_j(G)$$

Write the system of equations for searching for a_j ($j = 1, 2, 3$)

$$\begin{cases} f(G_1) = a_1 g_1(G_1) + a_2 g_2(G_1) + a_3 g_3(G_1) = 3a_1 + 2a_2 + a_3 \\ f(G_2) = a_1 g_1(G_2) + a_2 g_2(G_2) + a_3 g_3(G_2) = 3a_1 + 3a_2 \\ f(G_3) = a_1 g_1(G_3) + a_2 g_2(G_3) + a_3 g_3(G_3) = 4a_1 + 6a_2 \end{cases}$$

Solve this system and get

$$a_1 = f(G_2) - \frac{1}{2}f(G_3), \quad a_2 = \frac{1}{2}f(G_3) - \frac{2}{3}f(G_2),$$
$$a_3 = f(G_1) + \frac{1}{2}f(G_3) - \frac{5}{3}f(G_2)$$

So, taking values $f(G_i)$ ($i = 1, 2, 3$) for an arbitrary invariant f, we expand it on the basis g_1, g_2, and g_3. For example, take

$$f(G) = \sum_{i=1}^{n} v_i^2$$

where v_i is the degree of ith vertex ($i = 1, ..., n$). Then $f(G_1) = 10$, $f(G_2) = 12$, $f(G_3) = 20$. Thus, $a_1 = 2$, $a_2 = 2$, $a_3 = 0$ and $f(G) = 2g_1(G) + 2g_2(G)$, and we express invariant f in terms of occurrence numbers of some fragments.

Let us consider another basis, f_i ($i = 1, 2, 3$), obtained with the help of matrix

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 0 & 4 & 5 \\ 0 & 0 & 6 \end{pmatrix} \quad (\det A \neq 0)$$

by the following way

$$\begin{pmatrix} f_1 \\ f_2 \\ f_3 \end{pmatrix} = A \begin{pmatrix} g_1 \\ g_2 \\ g_3 \end{pmatrix}$$

that is

$$\begin{cases} f_1 = g_1 + 2g_2 + 3g_3 \\ f_2 = 4g_2 + 5g_3 \\ f_3 = 6g_3 \end{cases}$$

In this case

$$f_1(G_1) = 10, \quad f_1(G_2) = 9, \quad f_1(G_3) = 16$$

$$f_2(G_1) = 13, \quad f_2(G_2) = 12, \quad f_2(G_3) = 24$$

$$f_3(G_1) = 1, \quad f_3(G_2) = 0, \quad f_3(G_3) = 0$$

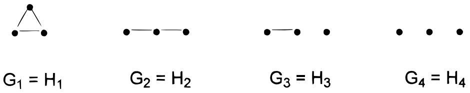
Analogously, any graph invariant f may be expanded on basis f_i ($i = 1, 2, 3$) with coefficients

$$a_1 = f(G_2) - \frac{f(G_3)}{2}, \quad a_2 = \frac{3}{8}f(G_3) - \frac{2}{3}f(G_2),$$
$$a_3 = f(G_1) - \frac{4}{3}f(G_2) + \frac{1}{8}f(G_3)$$

Example 2 (to Theorem 2). Consider the following symmetric measure:

$$d(G_k, G_l) = |G_k| + |G_l| - 2|MCS(G_k, G_l)|$$

where MCS(G_k, G_l) is the maximal common subgraph of graphs G_k and G_l. The symbol $|G|$ denotes the total sum of numbers of vertices and edges in a graph G. Let G_i ($i = 1, ..., 4$) be the set of all nonlabeled graphs with $n = 3$ vertices, and $H_i = G_i$ ($i = 1, ..., 4$):



Then

$$B = (b_{ij}) = (g_j(G_i)) = \begin{pmatrix} 1 & 3 & 3 & 1 \\ 0 & 1 & 2 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Take $A = E$ (E is a unit matrix). Then $f_i = g_i$ ($i = 1, ..., 4$), det $B \neq 0$, and satisfy conditions (3) and (4). Calculate values $d(G_k, G_l)$ ($k, l = 1,..., 4$; $k \neq l$, $k < l$):

$$d(G_1, G_2) = 1, \quad d(G_1, G_3) = 2, \quad d(G_1, G_4) = 3,$$
$$d(G_2, G_3) = 1, \quad d(G_2, G_4) = 2, \quad d(G_3, G_4) = 1$$

Obtain formula (5) for this measure. According to the algorithm of constructing matrix M (see proof of Theorem 2), it is first necessary to find two matrices F_1 and C, from which $M = (m_{ij})$ can easily be obtained: $M = F_1^{-1}C(F_1^{-1})^*$. We obtain matrix F_1 from $F = AB = B$ by "deleting" the fourth column and fourth line in F. So

$$F_1 = \begin{pmatrix} 1 & 3 & 3 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{pmatrix}, \quad F_1^{-1} = \begin{pmatrix} 1 & -3 & 3 \\ 0 & 1 & -2 \\ 0 & 0 & 1 \end{pmatrix},$$

$$(F_1^{-1})^* = \begin{pmatrix} 1 & 0 & 0 \\ -3 & 1 & 0 \\ 3 & -2 & 1 \end{pmatrix}$$

We determine elements c_{kl} ($k, l = 1, ..., 3$; $k \leq l$) of symmetrical matrix C by formula (10): $c_{11} = 3$, $c_{22} = 2$, $c_{33} = 1$, $c_{12} = 2$, $c_{13} = 1$, $c_{23} = 1$. So

$$C = \begin{pmatrix} 3 & 2 & 1 \\ 2 & 2 & 1 \\ 1 & 1 & 1 \end{pmatrix}$$

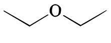
$$M = F_1^{-1}C(F_1^{-1})^* = \begin{pmatrix} 1 & -3 & 3 \\ 0 & 1 & -2 \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} 3 & 2 & 1 \\ 2 & 2 & 1 \\ 1 & 1 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 & 0 \\ -3 & 1 & 0 \\ 3 & -2 & 1 \end{pmatrix} =$$
$$\begin{pmatrix} 6 & -3 & 1 \\ -3 & 2 & -1 \\ 1 & -1 & 1 \end{pmatrix}$$

It is easy to verify that for the matrix M indeed

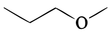
$$d(G_k, G_l) = \sum_{i,j=1}^{3} (f_i(G_k) - f_i(G_l))(f_j(G_k) - f_j(G_l))m_{ij} =$$
$$(M(\bar{f}_k - \bar{f}_l), \bar{f}_k - \bar{f}_l)$$

where $\bar{f}_1 = (1,3,3)$, $\bar{f}_2 = (0,l,2)$, $\bar{f}_3 = (0,0,1)$, $\bar{f}_4 = (0,0,0)$

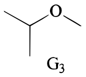
Example 3 (to Theorem 2). Consider graphs G_i ($i = 1, ..., 8$):



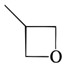
G_1



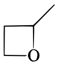
G_2



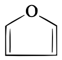
G_3



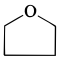
G_4



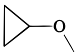
G_5



G_6

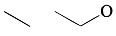


G_7

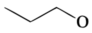


G_8

Select their subgraphs H_i ($i = 1, ..., 8$):

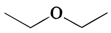


H_1



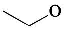
H_2

H_3

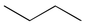


H_4

H_5



H_6



H_7

H_8

Consequently,

$$B = (b_{ij}) = (g_j(G_i)) = \begin{pmatrix} 2 & 0 & 0 & 1 & 2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 1 & 0 & 0 \\ 0 & 4 & 1 & 2 & 3 & 0 & 0 & 0 \\ 2 & 2 & 1 & 2 & 3 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 2 & 2 & 0 & 1 & 3 & 0 & 1 & 0 \\ 0 & 2 & 0 & 0 & 3 & 1 & 0 & 1 \end{pmatrix}$$

Consider

$$A = \begin{pmatrix} 0.5 & 8 & -4.5 & -2.5 & 4.5 & -0.5 & -4.5 & 1 \\ -1 & -11 & 4 & 2 & -2 & 5 & 3 & 1 \\ -3 & -4 & -3 & -1 & 2 & 13 & 1 & 1 \\ -1 & -8 & 2 & 1 & -1 & 6 & 2 & 1 \\ -3 & -8 & -1 & 0 & 0 & 13 & 3 & 1 \\ -2 & -6 & 0 & 0 & 0 & 7 & 2 & 1 \\ -1 & -4 & -1 & 0 & 0 & 6 & 1 & 1 \\ 0 & -2 & -1 & 0 & 0 & 3 & 0 & 1 \end{pmatrix}$$

Then

$$F = AB = \begin{pmatrix} 1 & 0 & 2 & 0 & 3 & 1 & 0 & 1 \\ 0 & 1 & 0 & 2 & 1 & 3 & 1 & 1 \\ 0 & 0 & 1 & 0 & 2 & 0 & 3 & 1 \\ 0 & 0 & 0 & 1 & 1 & 2 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 3 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 2 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}, F_1 = \begin{pmatrix} 1 & 0 & 2 & 0 & 3 & 1 & 0 \\ 0 & 1 & 0 & 2 & 1 & 3 & 1 \\ 0 & 0 & 1 & 0 & 2 & 0 & 3 \\ 0 & 0 & 0 & 1 & 1 & 2 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 3 \\ 0 & 0 & 0 & 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Vectors $\bar{f}_k = (f_1(G_k), ..., f_7(G_k))$ (k=1, ..., 7) are rows in matrix F_1. Let $d(G_k, G_l) = |G_k| + |G_l| - 2|MCS(G_k, G_l)|$, where $|G_k|$ is the overall number of vertices and edges in G_k, and MCS is the maximum common subgraph of graphs G_k and G_l. Denote for brevity: $d(G_k, G_l) = d_{kl}$ ($k < l$). Then

$$\begin{array}{lllllll} d_{12} = 2 & d_{23} = 2 & d_{34} = 3 & d_{45} = 2 & d_{56} = 6 & d_{67} = 4 & d_{78} = 2 \\ d_{13} = 2 & d_{24} = 1 & d_{35} = 1 & d_{46} = 4 & d_{57} = 2 & d_{68} = 4 & \\ d_{14} = 3 & d_{25} = 3 & d_{36} = 5 & d_{47} = 2 & d_{58} = 2 & & \\ d_{15} = 1 & d_{26} = 3 & d_{37} = 3 & d_{48} = 2 & & & \\ d_{16} = 5 & d_{27} = 1 & d_{38} = 1 & & & & \\ d_{17} = 1 & d_{28} = 1 & & & & & \\ d_{18} = 3 & & & & & & \end{array}$$

Then find symmetric matrix $C = (c_{kl})$, ($k, l = 1, ..., 7$) using formulas (10) and matrix M:

$$C = \begin{pmatrix} 1 & 1 & 1 & 1 & 2 & 1 & 2 \\ 1 & 1 & 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 3 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 2 & 1 & 1 & 1 \\ 2 & 0 & 1 & 1 & 2 & 0 & 1 \\ 1 & 1 & 0 & 1 & 0 & 4 & 1 \\ 2 & 1 & 0 & 1 & 1 & 1 & 2 \end{pmatrix}$$

Thus, for the graphs belonging to this set, formula (5) is

$$M = (F_1^{-1})C(F_1^{-1})^* = \begin{pmatrix} 83 & -46 & 18 & 67 & -29 & -22 & 12 \\ -46 & 29 & -12 & -41 & 17 & 13 & -7 \\ 18 & -12 & 13 & 19 & -12 & -5 & 4 \\ 67 & -41 & 19 & 62 & -25 & -21 & 10 \\ -29 & 17 & -12 & -25 & 14 & 7 & -5 \\ -22 & 13 & -5 & -21 & 7 & 8 & -3 \\ 12 & -7 & 4 & 10 & -5 & -3 & 2 \end{pmatrix}$$

true with this matrix M.

CONCLUSIONS

1. Any graph similarity symmetric measure $d(G_k, G_l)$ for graphs from any finite set is defined uniquely by some nonsingular matrix A and symmetric matrix M. All measures from literature may be represented in the form (5) obtained by us with definite values of above mentioned parameters.

2. The obtained general analytical formula (5) allows one to construct an infinite number of novel similarity measures varying the matrices A and M.

3. When predicting the properties of chemical compounds using methods based on some similarity measure (cluster analysis, method of nearest neighbors, "competitive" neural networks), a similarity measure must be determined (fitted) using the training set of compounds. Such an approach to the measure construction can be considered as a universal one and may be automated in contrast to usually used approaches consisting in guessing the measure for a particular task.

APPENDIX (PROOFS OF THEOREMS)

Proof of Theorem 1. (1) Show that one can take $H_j = G_j$ ($j = 1$,..., N). Order G_j in decreasing numbers of vertices and edges (graphs with the same numbers of vertices and edges may be ordered arbitrarily). Then $b_{ij} = g_j(G_i) = 1$ for $i = j$, and $b_{ij} = g_j(G_i) = 0$ for $i > j$. Hence, B is an uppertriangular matrix with units on its diagonal, and det $B = 1 \neq 0$. Thus, the chosen subgraphs H_j ($j =$ l, ..., N) satisfy to condition (1).

(2) Write (2) for all $G = G_i$:

$$f(G_i) = \sum_{j=1}^{N} a_j g_j(G_i) \quad (i = 1,...,N) \tag{6}$$

To prove the statement from section 2 of Theorem 1, it is necessary and sufficient to prove that the system of eq 6 for unknown parameters $\bar{a} = (a_1, ..., a_N)$ has a unique solution for any numbers $\bar{y} = (f(G_1), ..., f(G_N))$. Write (6) in the matrix form

$$\bar{y} = B\bar{a} \tag{7}$$

Evidently that (7) has a unique solution of kind $\bar{a} = B^{-1}\bar{y}$ (since det $B \neq 0$ and therefore there exists an inverse matrix B^{-1}).

(3) Substitute g_j for f_j in (2) ($j = 1$,..., N). Repeating the mathematical treatment from section 2, we obtain the system of equations in matrix form

$$\bar{y} = F\bar{a} \tag{8}$$

where $F = (f_{ij}) = (f_j(G_i))$, $i, j = 1$,..., N. However, $F = AB$, det $F =$ det A•det $B \neq 0$. Hence, eq 8 has a unique solution of kind $\bar{a} = F^{-1}\bar{y}$.

(4) It is evident that a matrix F for a basis satisfying conditions (3) and (4) has the following structure

$$F = \begin{pmatrix} f_1(G_1)...\ f_{\mathrm{N}}(G_1) \\ \dotfill \\ f_1(G_{\mathrm{N}})...\ f_{\mathrm{N}}(G_{\mathrm{N}}) \end{pmatrix} = \begin{pmatrix} & & & & & 1 \\ & & & & & . \\ & & F_1 & & & . \\ & & & & & 1 \\ 0 & . & . & . & 0 & 1 \end{pmatrix}$$

where F_1 is any $(N-1) \times (N-1)$ square matrix, det $F_1 =$ det $F \neq 0$. Since $F = AB$, then $A = FB^{-1}$. Considering different nonsingular matrix F_1, we obtain different matrices A with required property.

Theorem 1 is proven.

Proof of Theorem 2. Introduce the following designations:

$$F_1 = \begin{pmatrix} f_1(G_1)...\ f_{\mathrm{N}-1}(G_1) \\ \dotfill \\ f_1(G_{\mathrm{N}-1})...\ f_{\mathrm{N}-1}(G_{\mathrm{N}-1}) \end{pmatrix}$$

F_1*—matrix transpose of F_1; $\bar{e}_k$ ($k = 1$,..., $N-1$) are vector-columns of length $(N-1)$, the kth component of $\bar{e}_k$ is equal to 1, while other components of $\bar{e}_k$ are equal to 0; $C = F_1MF_1^* = (c_{kl})$—a square $(N-1) \times (N-1)$ matrix. Evidently $\bar{f}_k = F_1^*\bar{e}_k$, $\bar{f}_l = F_1^*\bar{e}_l$, det $F_1 \neq 0$ (see proof of Theorem 1), and C is a symmetric matrix (if M is symmetric one).

Rewrite the right part of (5), using the introduced designations. For $k, l = 1$,..., $N - 1$:

$$\begin{aligned}(M(\bar{f}_k - \bar{f}_l), \bar{f}_k - \bar{f}_l) = (MF_1^*(\bar{e}_k - \bar{e}_l),F_1^*(\bar{e}_k - \bar{e}_l)) = \\ (F_1MF_1^*(\bar{e}_k - \bar{e}_l),(\bar{e}_k - \bar{e}_l)) = (C(\bar{e}_k - \bar{e}_l),(\bar{e}_k - \bar{e}_l)) = \\ (C\bar{e}_k,\bar{e}_k) + (C\bar{e}_l,\bar{e}_l) - 2(C\bar{e}_k,\bar{e}_l) = c_{kk} + c_{ll} - 2c_{kl}\end{aligned}$$

Analogously, for $k = 1$,..., $N-1$ and $l = N$ using the condition $\bar{f}_N = (0,0,...,0)$

$$\begin{aligned}(M(\bar{f}_k - \bar{f}_l),(\bar{f}_k - \bar{f}_l)) = (M\bar{f}_k, \bar{f}_k) + (M\bar{f}_l, \bar{f}_l) - \\ 2(M\bar{f}_k, \bar{f}_l) = (M\bar{f}_k, \bar{f}_k) + (M\bar{f}_N,\bar{f}_N) - 2(M\bar{f}_k, \bar{f}_N) = \\ (M\bar{f}_k, \bar{f}_k) = (C\bar{e}_k, \bar{e}_k) = c_{kk}\end{aligned}$$

So, eq 5 can be written in the following form:

$$\begin{cases} d(G_k, G_l) = c_{kk} + c_{ll} - 2c_{kl} & (k, l = 1,..., N - 1;\ k < l) \\ d(G_k, G_N) = c_{kk} & (k = 1, ..., N - 1) \end{cases} \tag{9}$$

It is evident that the system of eq 9 for elements of matrix C has the unique solution

$$\begin{cases} c_{kk} = d(G_k, G_N)\ (k = 1, ..., N - 1); \\ c_{kl} = 0.5(d(G_k, G_N) + d(G_l, G_N) - d(G_k, G_l)) \end{cases} \tag{10}$$

$$(k,l = 1, ..., N - 1;\ k < l)$$

Thus, $C = (c_{kl})$ *is* uniquely determined by $d(G_k, G_l)$, and there exists a unique matrix $M = (F_1)^{-1}C(F_1^*)^{-1} = F_1^{-1}$• $C(F_1^{-1})^*$ with required property.

Theorem 2 is proven.

REFERENCES AND NOTES

(1) Johnson, M. A. A review and examination of mathematical spaces underlying molecular similarity analysis. *J. Math. Chem.* **1989**, *3*, 117−145.

(2) *Concepts and Applications of Molecular Similarity;* Johnson, M. A.; Maggiorra, G. M., Eds.; Wiley: New York, 1990.

(3) Maggiora, G. M.; Mestres, J.; Hagadone, T. R.; Lajiness, M. S. Computer-Aided Drug Discovery, Pharmacia and Upjohn Book of Abstracts, 213th ACS National Meeting, San Francisco, April 13−17, 1997. AN 1997:160175.

(4) Ponec, R.; Strand, M. A novel approach to the characterization of molecular similarity. The 2nd order similarity index. *Coll. Czech. Chem. Commun.* **1990**, *55*, 896−902.

(5) Takahashi, Y.; Sukekawa, M.; Sasaki, S.-i. Automatic identification molecular similarity using reduced-graph representation of chemical structure. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 639−643.

(6) Basak, S. C.; Bertelsen, S.; Grunwald, G.D. Application of graph theoretical parameters in quantifying molecular similarity and structure-activity relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 270−276.

(7) Hall, L. H.; Kier, L. B.; Brown, B. B. Molecular similarity based on novel atom type electrotopological state indices. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1074−1080.

(8) Cheng, C.; Maggiora, G.; Lajiness, M.; Johnson, M. Four association constants for relating molecular similarity measures. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 909−915.

(9) Baskin, I. I.; Skvortsova, M. I.; Stankevich, I. V.; Zefirov, N. S. On the basis of invariants of labeled molecular graphs. *Dokl. Akad. Nauk* **1994**, *339*, 346−350.

(10) Baskin, I. I.; Skvortsova, M. I.; Stankevich, I. V.; Zefirov, N. S. On the basis of invariants of labeled molecular graphs. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 527−531.

(11) Gantmakher, F. R. *The Theory of Matrices*; GRFML: Moscow, 1967; pp 224−225.

CI970037B