communications in mathematical
match
and in computer chemistry

no. 49, September 2003
MATCDY (49) 171- 178 (2003)

ISSN 0340 - 6253

TAXONOMY ALGORITHM FOR MOLECULAR GRAPHS

Lev I. Makarov

Sobolev Institute of Mathematics, Siberian Branch of the
Russian Academy of Sciences, Novosibirsk 630090, Russia

(Received November 19, 2002)

Abstract. A taxonomy algorithm for vertices of a weighted graph based on the notion of separability of vertices of an edge is presented. The algorithm has found application in predicting the properties of a compound by the data about its structure, and in elucidating a fragment composition of a compound by its infra-red spectrum.

INTRODUCTION

Computer processing of data about molecular structures of a sample of chemical compounds often employs graph models of both the compound structures and the sample. Such models are used in the investigations into the problem of quantitative structure-property relationship of chemical compounds (QSPR), studies of the structure of compounds by their infra-red (IR) spectra [1-5], etc.

Molecular graph whose vertices correspond to atoms or groups of atoms of a compound, and edges — to chemical bonds between atoms serves as a model of the compound structure.

To determine a quantitative measure of structural similarity (closeness) of a pair of compounds, the value of the distance between their molecular graphs is used that depends on the size of the largest common subgraph (fragment). The larger the common fragment, the greater structural similarity of the compounds, and the smaller the distance between them. Thus the sample of the compounds under study may be associated with the weighted graph whose vertices correspond to compounds, and the length of edges is defined by the distances between their molecular graphs.

In studies of the sample compounds the problem of the taxonomy (partition) of vertices of a weighted graph of the sample naturally arises. The problem consists in partitioning of

the set of the graph vertices into taxons - subsets of vertices corresponding to structurally similar compounds. The paper presents fast greedy algorithm [6] for the taxonomy of vertices of a weighted graph based on the notion of the separability of vertices of an edge.

SIMILARITY OF MOLECULAR GRAPHS

A graph $G(V, X)$ is called labeled if some labels are assigned to its elements — vertices of the set V and edges of the set X. Molecular graphs are labeled ones. Two labeled graphs are called isomorphic if between their sets of vertices there exists a one-to-one correspondence preserving the adjacency of vertices and labels of elements. The graph $G'(V', X')$ is called the subgraph of the graph G if $V' \subseteq V$, $X' \subseteq X$.

A graph F is called a common subgraph of labeled graphs G and H if they contain subgraphs G' and H' isomorphic to the graph F.

In finding the common subgraph F of the graphs G and H, one can specify a particular type of subgraphs G' and H': connected or disconnected subgraphs, induced subgraphs, etc. A type of subgraphs is determined by specific features of the problems. Two molecular graphs contain no common subgraph if their vertices have no common labels. For example, assume that for a sample of compounds it is necessary to find such common subgraphs of molecular graphs which contain solely vertices involved in the cycles. Then, labels different from those of the vertices of acyclic subgraphs are assigned to the vertices of cycles of graphs. In this case, some molecular graphs of the sample compounds can have no common subgraphs (for instance, benzene and hexane).

A common subgraph of two graphs is called the largest subgraph if a real function depending on the number of vertices and edges assumes its maximum value at this subgraph.

The measure of structural similarity of graphs $G_i(V_i, X_i)$ and $G_j(V_j, X_j)$ may be estimated by quantitative characteristics depending on the number of vertices and edges of the largest common subgraph. Denote the largest common subgraph of graphs G_i and G_j by $G_{ij}(V_{ij}, X_{ij})$, $P_i = |V_i| + |X_i| > 0$, $P_j = |V_j| + |X_j| > 0$, $P_{ij} = |V_{ij}| + |X_{ij}| \geq 0$. Then the values of distances between the graphs, for example, the distance $r_{ij} = P_i + P_j - 2P_{ij}$ [7] or the relative distance $R_{ij} = r_{ij}/p(i,j)$, $p(i,j) = (P_i + P_j - P_{ij}) > 0$, $0 \leq R_{ij} \leq 1$, may be used as the estimates of the similarity of the graphs.

It is not always reasonable to use the distance r_{ij} to estimate structural similarity of graphs. For example, for structurally different graphs G_i and G_j that do not have common subgraph ($P_{ij} = 0$) the distance r_{ij} is less than the distance r_{ik} between structurally similar graphs G_i and G_k that has common subgraph ($P_{ik} > 0$) if $P_j < P_k - 2P_{ik}$. For the given example, $R_{ij} = 1 > R_{ik} = 1 - P_{ik}/p(i,k)$ holds. So, to estimate structural similarity of graphs, the relative distance R_{ij} is used.

Thus, for a given sample of compounds the complete weighted graph $G(V, X)$ may be constructed the vertices of which correspond to the sample compounds, and the edge lengths are equal to the distances between molecular graphs of compounds.

SEPARABILITY OF VERTICES OF A WEIGHTED GRAPH

Let $G(V, X)$ be an undirected graph without loops and multiple edges which has the set of vertices $V = \{v, u, ...\}$, and the set of edges X where each element $x \in X$ is a pair vu of vertices of V. The graph G is called weighted, if a positive real number $l(x) > 0$ — the edge length (weight) — is assigned to each of its edges $x = vu$. The subgraph G' where the set of edges X' consists of all edges connecting vertices of V' in G is called the subgraph induced in the graph G by the set of vertices V', and is denoted by $\langle V' \rangle$.

The shortest spanning tree $S(V, X_S)$ of connected graph G is its connected subgraph without cycles (tree) that contains all vertices of the graph and has the minimum sum of the edge lengths.

For the shortest spanning trees of the weighted graph the validity of the following statements is easily established.

STATEMENT 1: Each shortest spanning tree $S(V, X_S)$ of the graph G for each vertex v contains at least one edge x^v of the minimum length $l(x^v) = l^v = \min l(x)$, where x belongs to the set of all edges incident to the vertex v in the graph G.

STATEMENT 2: Removal of the edge $x = vu$ from the shortest spanning tree $S(V, X_S)$ results in two trees $S'(V', X_{S'})$ and $S''(V'', X_{S''})$, where $v \in V'$, $u \in V''$. These trees are the shortest spanning trees for induced subgraphs $\langle V' \rangle$ and $\langle V'' \rangle$.

Define the *separability* of vertices of the edge $x = vu$ in G as

$f_x(G) = \max(l(x)/l^v, l(x)/l^u)$.

In other words, separability of vertices of the weighted graph edge is defined as the

ratio of its length to the length of the shortest adjacent edge. From Statement 1 follows

STATEMENT 3: For any edge $x = vu$ of the shortest spanning tree $S(V, X_S)$ of the weighted graph $G(V, X)$ the equality $f_x(G) = f_x(S)$, $x \in X_S$, holds.

Further we denote $f_x(G)$ as f_x.

TAXONOMY OF VERTICES OF A WEIGHTED GRAPH

The partition $\tilde{V}_G = \{V_k \mid k = 1, \ldots, T\}$ of the set V of vertices of connected weighted graph G into subsets V_k such that each graph $G_k(V_k, X_k) = \langle V_k \rangle$ is connected will be called the taxonomy of the graph G, and graphs G_k and their sets of vertices V_k will be called taxons.

If the weighted graph G is a complete graph, then for its subgraph (taxon) $G_k = \langle V_k \rangle$ the quantity $D_k = \max_x l(x), x \in X_k$, is called the diameter, and the vertex v of the value $\min_v \max_u l(x), x = vu \in X_k$, is called the center of G_k.

For complete graph any partition of vertices is a taxonomy. If a graph is a tree, then removal of some edge from it gives rise to its partition into two trees - taxons.

Evidently, the taxonomy $\tilde{V}_S = \{V_k \mid k = 1, \ldots, T\}$ of the shortest spanning tree $S(V, X_S)$ is at the same time the taxonomy $\tilde{V}_G$ of the graph G.

For the taxonomy $\tilde{V}_S = \tilde{V}_G$ of vertices of the shortest spanning tree S of the graph G the separability of taxons V_k and V_t connected by the edge $x = vu \in X_S$, $v \in V_k, u \in V_t$, in the spanning tree S is defined to be equal to the separability f_x of vertices of this edge.

When obtaining the taxonomy of the graph G by the algorithm that takes into account the separability of vertices, we use statements 1-3. This allows one to consider the shortest spanning tree S instead of the whole graph G.

In partitioning of vertices of a weighted graph into taxons of closely-spaced vertices, it is naturally taken that vertices of the edge with small separability belong to one and the same taxon, while those of the edge with large separability — to different taxons.

Greedy taxonomy algorithm removes sequentially the edges ordered according to non-increase either of separabilities of their vertices, or of their lengths from the shortest spanning tree. Each removal of one edge causes the number of taxons to increase by one. It is natural to assess the taxonomy quality by a simple function depending on the separabilities of taxons or lengths of edges joining them. For example, the taxonomy

quality may be estimated by the value of the mean length of the set of edges removed from the spanning tree, or by the value of the mean separability of taxons. The greater the value of the quantity chosen, the higher the taxonomy quality.

Denote the set of edges removed from the shortest spanning tree S by the taxonomy algorithm on the K-th step by $X_S^K, |X_S^K| = K \geq 1$.

Then the quality $Q(\tilde{V})$ of the taxonomy $\tilde{V}_G = \tilde{V}_S = \tilde{V}, |\tilde{V}| = K + 1$, of the graph G may be defined as $Q(\tilde{V}) = \sum_x h_x/K$, where $x \in X_S^K$, h_x is the quantitative characteristic (separability of vertices, or length) of the edge x removed from the shortest spanning tree.

Let the algorithm perform ordering of edges according to non-increase in one of the above variables. If in the function $Q(\tilde{V})$ this variable is used as h_x, then the greatest quality will be inherent in the taxonomy obtained by removing the first edge. This follows from

STATEMENT 4: Let $y_1 \geq y_2 \geq y_3 \geq \cdots$ be a non-increasing sequence of nonnegative real numbers and let $g(n) = \sum_1^n y_i/n$. Then $\max_n g(n) = g(1)$.

For the function $g(n)$ the equality $(n+1)(g(n)-g(n+1)) = g(n)-y_{n+1}$ holds. Besides, for $g(n)$ and the above sequence the inequalities $g(n) \geq y_n \geq y_{n+1}$ hold true for all n. From this it follows that $g(n) - g(n+1) \geq 0$ and $\max_n g(n) = g(1)$.

Thus the algorithm makes use of the following ways of specifying h_x:

if $h_x = l(x)$, then ordering of edges is performed by the values of separability f_x,

if $h_x = f_x$, then ordering of edges is performed by the values of their lengths $l(x)$.

TAXONOMY ALGORITHM

The proposed taxonomy algorithm is intended for the partitioning of vertices of complete weighted graph $G(V, X)$ of the sample of compounds into taxons corresponding to structurally similar compounds.

Since the compounds of the sample are structurally different, the partitioning into such taxons calls for the possibility of restricting their diameter. Otherwise, partitioning of vertices of a graph may result in such a taxon where adjacent vertices correspond to structurally similar compounds, and diametral ones — to structurally different compounds. Thus the taxonomy algorithm makes use of the parameter $D_0 \geq 0$ that specifies the maximum possible value of the taxon diameter.

When applying a taxonomy algorithm, a user often seeks to get a small number of large taxons. In our algorithm taxons of the diameter $D_k \leq D_0$ are not partitioned further whatever the quality of the taxonomy. At $D_0 = 0$ the taxonomy of the highest quality is chosen of all partitions obtained.

The taxonomy algorithm consists of the following steps.

1. The values of parameters h_x and D_0 are specified.

2. For the graph $G(V, X)$ its diameter D is found.

If $D \leq D_0$, then taxonomy is not performed.

3. If $D > D_0$, then in the graph G the shortest spanning tree $S(V, X_S)$ is found, the separability f_x of vertices of each edge $x \in X_S$ is calculated, and all these edges are arranged in order according to non-increase in the value corresponding to the given h_x.

4. In the order obtained the edges are removed from X_S one by one.

Removal of the set of edges X_S^K, $|X_S^K| = K \geq 1$ leads to the partition $\tilde{V} = \{V_k\}$, $k = 1, \ldots, K + 1$, of the spanning tree S into $(K + 1)$ spanning trees S_k. Each spanning tree S_k is the shortest spanning tree of the taxon $G_k = (V_k)$.

If the edge belongs to the spanning tree S_k of the taxon G_k of the diameter $D_k \leq D_0$, then this edge is not removed from the spanning tree S_k. Thus further partitioning of the taxon G_k is not performed.

5. If $D_0 = 0$, then for each of partitions $\tilde{V}$ of vertices obtained in step 4. the taxonomy quality $Q(\tilde{V}) = \sum h_x / K$, $x \in X_S^K$, is calculated, and the taxonomy having the highest quality is chosen.

Of course, final assessment of the quality of the taxonomy obtained can be given solely by an expert-chemist, since formal estimates of its quality can fail to take account of essential aspects of the investigations conducted.

APPLICATIONS OF THE TAXONOMY ALGORITHM

The proposed taxonomy algorithm has been successfully tested in BACC computer system — analysis and classification of chemical compounds — designed in Sobolev Institute of Mathematics (IM), Siberian Branch of the Russian Academy of Sciences (SB RAS), in studies of QSPR problems [2].

Besides, the algorithm has been used in Novosibirsk Institute of Organic Chemistry

(NIOC), SB RAS, in finding fragment composition of the compound by the data on structures of compounds similar to it in IR spectrum [5].

BACC system is intended for the analysis of molecular graphs and prediction of properties of compounds.

Analysis of molecular graphs consists in the following: calculation of their topological indices, finding common fragments of graphs, substructural search for compounds search for the graphs containing a given subgraph, etc.

Prediction of the properties of the compounds under study is performed on the basis of their structural similarity to the compounds of the training sample (TS) divided into property classes. Property classes of compounds may be characterized, for example, by their biological activity — toxicity, medicinal properties, etc.

For each TS property class the construction of the weighted graph of its compounds is performed, and partitioning into taxons of structurally similar compounds is done. In each taxon its center is found that can be considered the vertex corresponding to structurally typical compound of the taxon. The compound under study is assigned to that property class of TS which contains the center of the taxon closest to this compound.

In experiments on compound property prediction conducted using BACC system the relative prediction error was, on the average, 22% for TS with 3 property classes for taxonomy algorithm parameters $D_0 = 0$ and $h_x = l(x)$.

Application of the taxonomy algorithm in finding fragment composition of the compound by its IR spectrum is as follows. The sample of compounds similar to the compound under study in IR spectra is extracted from the Data Base (NIOC SB RAS) containing about 32000 IR spectra and structures of compounds. The sample is partitioned into taxons of structurally similar compounds. In each taxon the largest common fragment of all its molecular graphs is found. More than 150 experiments aimed at revealing fragment composition of compounds have been run with the following algorithm parameters: $h_x = l(x), D_0 = 0, D_0 = l_G, D_0 = l_S$, where l_G is the mean length of the edge of the sample graph, l_S is the mean length of the edge of its shortest spanning tree. The analysis of the taxonomy shows that at $D_0 = l_S$ large structural fragments of compounds of each taxon can be found. The number of vertices of such a fragment is about 2/3 of the mean number of vertices of molecular graphs of taxon compounds. These fragments are used

in computer generation of compounds structurally similar to the compound under study. This makes it possible to reduce the number of compounds generated, thus reducing the time of the elucidation of the compound structure by its IR spectrum.

REFERENCES

[1] A. J. Stuper, W. E. Brugger, P. C. Jurs, *Computer Assisted Studies of Chemical Structure and Biological Function*, Wiley, N. Y., 1979.

[2] L. I. Makarov, Methods and Algorithms for Predicting the Properties of Chemical Compounds by Common Fragments of Molecular Graphs. *J. Struct. Chem.*, **1998**, 39, 93-102.

[3] K. Varmuza, P. N. Penchev, H. Scsibrany, Maximum Common Substructures of Compounds Exhibiting Similar Infrared Spectra. *J. Chem. Inf. Comput. Sci.*, **1998**, 38, 420-427.

[4] V. N. Piottukh-Peletskii, T. F. Bogdanova, B. G. Derendyaev, Complete Sets of Structure Fragments for Interpreting IR Spectra Using IR Databases. *J. Struct. Chem.*, **1996**, 37, 323-331.

[5] B. G. Derendyaev, L. I. Makarov, T. F. Bogdanova, V. N. Piottukh-Peletskii, Taxonomy of Structures Selected from the IR Spectroscopy Database. *J. Struct. Chem.*, **2001**, 42, 271-280.

[6] E. M. Reingold, J. Nievergelt, N. Deo, *Combinatorial Algorithms, Theory and Practice*, Prentice-Hall, Englewood Cliffs, 1977.

[7] M. Johnson, M. Naim, V. Nicholson, C.-C. Tsai, Unique Mathematical Features of the Substructure Metric Approach to Quantitative Molecular Similarity Analysis. In: R. B. King and D. H. Rouvray (Eds), *Graph Theory and Topology in Chemistry*, Elsevier, Amsterdam, 1987, 219-225.