Toward Reconstruction of Trees by Using Graph Invariants

István Lukovits

Central Research Institute of Chemistry, Hungarian Academy of Sciences,
H-1525 Budapest, P.O. Box 17, Hungary

Received May 18, 1993●

The endpoint–endpoint distances of a tree were used together with the endpoint–branching point distances to reconstruct the underlying tree. The concept of triads and triangles (i.e. existing triads), composed of three endpoint–endpoint distances, was introduced. Each triangle defines a skeleton, that is a part of the underlying graph, and each skeleton defines a triangle uniquely. Several rules were derived, which can be used to eliminate irrelevant triads. A numerical example for a tree containing five endpoints has been worked out.

1. INTRODUCTION

There is quite an interest in encoding chemical structures with a single number or with a reasonable set of numbers. This problem could not be solved thus far, since no single graph invariant, or a reasonable set of such invariants, is known, which would define any graph up to isomorphism or, at least, which would be unique for each different structure. Several highly discriminative graph invariants were proposed; the best known examples are Balaban's index J,[1] the ID number and its variants,[2] and the "real vertex invariants".[3] Recently an invariant producing practically unique numbers for each structure has been derived.[4] Another approach starts with eigenvalues of the adjacency matrix or of the distance matrices to obtain unique numbers for each structure.[5]

Several approaches exist which can be used to encode various classes of molecules. These codes can be easily decoded. For trees the N-tuple representation has been proposed by Knop et al.,[6,7] and this code has been extended for cyclic structures[8] and for hexagon lattice trees.[9] For benzenoid hydrocarbons (polyhexes) the boundary code[6] and the DAST code[10] have been recommended, but other procedures have also been suggested.[11] An algorithm has been proposed for the generation of complete sets of isomers based on the molecular composition and the molecular weight.[12] Recently trees could be generated by using local vertex invariants and other topological indices.[13] Smolenskij showed[14] that endpoint–endpoint distances (EED) of a tree ($d_{1,2}$, $d_{1,3}$, ..., $d_{1,N}$, $d_{2,3}$, ..., $d_{2,N}$, ..., $d_{N-1,N}$, where $d_{i,j}$ denotes the distance between vertices i and j and N is the number of endpoints) are defining the structure of the underlying tree unambiguously. On the basis of this result, Zaretskij derived several rules that a matrix composed of EEDs must fulfill in order to correspond to an existing tree.[15]

Smolenskij's result is not sufficient to reconstruct a tree if there is a set of EEDs, but the subscripts are not known. By exchange of two "opposite" EEDs in a tree with four endpoints (Figure 1), a different tree with an identical set of EEDs will be obtained (this statement will be proved in the Appendix). The aim of the present paper is to suggest a set of graph invariants that are sufficient to reconstruct a tree. It has been conjectured that EEDs *plus* endpoint–branching point distances are sufficient to achieve this task. Several rules for the distance matrices were derived, and a numerical example has been worked out.

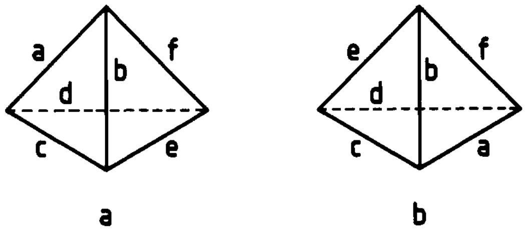

Figure 1. Two possible arrangements of six distances.

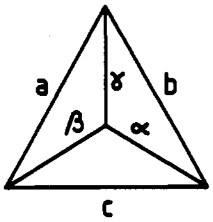

Figure 2. Triangle and its corresponding skeleton.

2. RULES FOR TRIADS

The basic concept of our approach is a collection of three numbers, the triad, which is composed of EEDs. The word "triangle" will denote in this paper a triad that can be found in the actual graph. From a set of $K = N(N-1)/2$ distances $\binom{K}{3}$ triads can be composed, but there will be only $\binom{N}{3}$ triangles. $d_{i,j}$ denotes the shortest path between vertices i and j (in trees it is the only possible path between i and j). The "skeleton" is a part of the underlying graph that is uniquely defined by a triangle (Figure 2). The skeleton may be obtained by using the following formula:[16]

$$\begin{pmatrix} +0.5 & +0.5 & -0.5 \\ +0.5 & -0.5 & +0.5 \\ -0.5 & +0.5 & +0.5 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} \gamma \\ \beta \\ \alpha \end{pmatrix} \qquad (1)$$

where a, b, and c denote the EEDs and α, β, and γ are the branches of the skeleton. Here a, b, and c and α, β, and γ are used to identify these distances, and they also denote the lengths of these paths. The lengths of these branches are the endpoint–branching point distances that are required in our approach. Note that the order of the distances determines the order of the branches in eq 1 (Figure 1). The set of EEDs will be denoted by S_1; the set of endpoint–branching point distances will be denoted by S_2. It will be supposed that S_1 and S_2 are known.

On the basis of the matrix of eq 1, it could be shown that only triangles with three edges of even length (eee-triangles) or triangles composed of two edges of odd length and one edge

● Abstract published in *Advance ACS Abstracts,* February 15, 1994.

0095-2338/94/1634-0287$04.50/0 © 1994 American Chemical Society

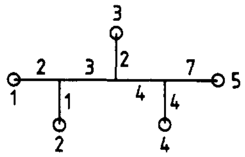

Figure 3. Graph considered in the numerical example. The numbers denote topological distances.

Table 1. Matrix of Distances between the Endpoints of the Graph in Figure 3

0	3	7	13	16
	0	6	12	15
		0	10	13
			0	11
				0

of even length (eoo-triangles) may appear in distance matrices and similarly in matrices.[16] It was also shown that, in a graph containing N endpoints, that is $K = N(N-1)/2$ EEDs, out of the $K + 1$ possibilities to color these distances (i.e., with a letter e or o), very few cases can actually be realized.

The following simple rules should be applied in order to eliminate triads and to obtain triangles (most of them are trivial; if not, the rule will be proved):

(1) Triangles in which the sum of distances is odd cannot exist.[15,16]

(2) Triangles in which $a + b \leq c$ cannot exist. Note that if ordinary distances are considered, equality is allowed, but if EEDs are considered, equality is not allowed.

(3) Triangles cannot produce nonexisting skeletons. A skeleton is nonexisting (a) if it contains a branch that is not contained in S_2 and (b) if the branch appears in S_2, but occurs in S_2 less times than in the skeleton.

(4) If numbers i and j appear only once and number k appears at least twice in S_1, triangle i, j, k can appear only once, and all identical triads must be deleted.

(5) If numbers i and j appear only once in S_1, there is only one i, j, k triangle, and all other i, j, m triads must be deleted.

Proof of Rules 4 and 5. Let us suppose that other i, j, k or i, j, m triangles exist. This means that there are at least two paths connecting vertices (i,k) and (j,k) or (i,m) and (j,m), a contradiction. The rule has been proved.

(6) Triangles cannot involve "opposite" distances (see Figure 1).

(7) Distances appearing once in S_1 will appear in $N - 2$ triangles. Distances appearing two times in S_1 will appear in $2(N - 2)$ triangles if they are opposite and in $2(N - 2) - 1$ triangles if they are adjacent.

3. NUMERICAL EXAMPLE

In order to illustrate the application of our rules a simple example will be considered (Figure 3). The tree has five endpoints, and from the set of ten distances (Table 1) 120 triads could be composed. S_1 = (3, 6, 7, 10, 11, 12, 13, 13, 15, 16), and S_2 = (1, 2, 2, 4, 4, 5, 5, 6, 7, 8, 8, 9, 11, 11, 14). Since S_1 contains four even numbers and six odd numbers, the graph belongs to class eo^3,[16] and there will be one eee-triangle and nine eoo-triangles. Application of rules 1 and 2 leaves us with 46 triads (Table 2). The following technique has been used to reduce the number of prospective triangles: first rule 3 was used; then rule 4 was applied. After these steps were completed, 21 triads remained: 2 of these were eee-triads (no. 8 and 12; triad 31 was deleted in step 12). Since one of them must appear in the triangles, triad 8 was selected together with triad 1 (step 26). Having selected two triangles, rules 5 and 6 could be applied to delete triads, and also rule 7 could be used to pick out triangles. The procedure can be followed in Table 2. After step 36 we were left with seven triangles (3,6,7; 3,12,13; 3,15,16; 6,10,12; 7,10,13; 7,13,16; and 11,13,16) and with seven triads (6,11,13; 6,13,15; 10, 11,13; 10,13,15; 11,12,13; 11,12,15; and 12,13,15), three of which must be triangles. At this stage the following distances in S_2 were available: 6, 10, 11, 11, 12, 13, 13, 15, 15). The only way to distribute these numbers in three triangles is by choosing triads 6,13,15; 10,11,13; and 11,12,15. It is easy to show that the triangles found by this algorithm correspond to the triangles of our graph in Figure 3.

Table 2. Selection of Triangles from Triads (Figure 3)

no.	triad	skeleton	deleted (by rule[a])	step
1	3,6,7	1,2,5		26
2	3,10,11	1,2,9	6	27
3	3,11,12	1,2,10	3	1
4	3,12,13	1,2,11		29
5	3,12,13	1,2,11	4	15
6	3,15,16	1,2,14		30
7	6,7,11	1,5,6	5	22
8	6,10,12	2,4,8		26
9	6,11,13	2,4,9	X	37
10	6,11,13	2,4,9	4	16
11	6,11,15	1,5,10	3	2
12	6,12,16	1,5,11	5	23
13	6,13,13	3,3,10	3	3
14	6,13,15	2,4,11		37
15	6,13,15	2,4,11	4	17
16	7,10,11	3,4,7	3	4
17	7,10,13	2,5,8		31
18	7,10,13	2,5,8	4	18
19	7,10,15	1,6,9	6	28
20	7,11,12	3,4,8	3	5
21	7,11,16	1,6,10	3	6
22	7,12,13	3,4,9	3	7
23	7,12,13	3,4,9	3	8
24	7,12,15	2,5,10	3	9
25	7,13,16	2,5,11		32
26	7,13,16	2,5,11	4	19
27	7,15,16	3,4,12	3	10
28	10,11,13	4,6,7		37
29	10,11,13	4,6,7	4	20
30	10,11,15	3,7,8	3	11
31	10,12,16	3,7,9	3	12
32	10,13,13	5,5,8	6	33
33	10,13,15	4,6,9	X	37
34	10,13,15	4,6,9	4	21
35	11,12,13	5,6,7	X	37
36	11,12,13	5,6,7	4	22
37	11,12,15	4,7,8		37
38	11,13,16	4,7,9		35
39	11,13,16	4,7,9	4	23
40	11,15,16	5,6,10	3	13
41	12,13,13	6,6,7	3	14
42	12,13,15	5,7,8	X	37
43	12,13,15	5,7,8	4	24
44	13,13,16	5,8,8	6	34
45	13,15,16	6,7,9	5	36
46	13,15,16	6,7,9	4	25

[a] Remark: an X denotes that the triad was deleted in the final stage (step 37) as a consequence of several rules (see text).

4. CONCLUSIONS

The present approach should be appropriate for computer programs, and in that way more complicated cases could be investigated. There is no way to select a priori triangles, i.e., triads which *must* be triangles, without using the screening procedure explained above. That means that the final set of triads may depend on how one or two triads are picked by the investigator after rules 1–4 were applied to the data set.

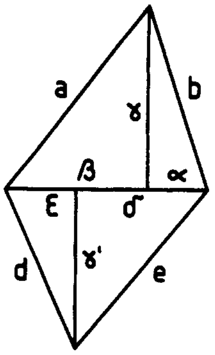

Figure 4. Two triangles with a common edge.

However, if wrong triads are selected, more unresolved cases are expected at the end of the procedure than by selecting correct triangles. It may be expected that further rules could be derived at a later stage, which might improve the efficiency of the screening process. Work in this direction is in progress.

APPENDIX

Consider a tree or a subgraph with four endpoints and six edges corresponding to the EEDs (d-edges). Edges a and e, b and d, and c and f are three pairs of opposite edges. The following statement will be proved:

Theorem: Exchange of any pair of opposite d-edges will not alter S_1, provided this exchange is legitimate (i.e., no ooo or eeo triads result).

Proof: Consider the distances in Figure 4. Two pairs of these distances must obey the Zaretskij rule:[15]

$$a + e = c + f \qquad (2)$$

Then f may be expressed as

$$f = \gamma + \gamma' + |\beta - \epsilon| = 0.5(a + b - 2c + d + e) + 0.5|a - b - d + e| \qquad (3)$$

In this formula letters a and e, b and d are equivalent, meaning that exchanging opposite pairs will not alter S_1. The underlying graph, however may be altered if $b \neq d$ or $a \neq e$. The theorem has been proved.

REFERENCES AND NOTES

(1) Balaban, A., T. Highly Discriminating Distance-Based Topological Index. *Chem. Phys. Lett.* **1982**, *89*, 399–404.
(2) Randić, M. Molecular ID Numbers: By Design. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 134–136.
(3) Balaban, A. T.; Catana, C. Search for Nondegenerate Real Vertex Invariants and Derived Topological Indices. *J. Comput. Chem.* **1993**, *14*, 155–160.
(4) Müller, W. R.; Szymanski, K.; Knop, J. V.; Mihalić, Z.; Trinajstić, N. The Walk ID Number Revisited. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 231–233.
(5) Liu, X.; Klein, D. J. The Graph Isomorphism Problem. *J. Comput. Chem.* **1991**, *12*, 1243–1251.
(6) Knop, J. V.; Müller, W. R.; Szymanski, K.; Nikolić, S.; Trinajstić, N. *Computer Generation of Certain Classes of Molecules*; SKTH/Kemija u Industriji: Zagreb, 1985.
(7) Knop, J. V.; Müller, W. R.; Jeričević, Z.; Trinajstić, N. Computer Enumeration and Generation of Trees and Rooted Trees. *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 91–99.
(8) Randić, M. Compact Molecular Codes. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 136–148.
(9) Kirby, E. C. Coding and Enumeration of Trees That Can Be Laid upon a Hexagon Lattice. *J. Math. Chem.* **1992**, *11*, 187–198.
(10) Knop, J. V.; Müller, W. R.; Szymanski, K.; Trinajstić, N. Enumeration of Planar Polyhex Hydrocarbons. *Rep. Mol. Theor.* **1990**, *1*, 95–98.
(11) Bangov, I. P. Toward the Solution of the Isomorphism Problem in Generation of Chemical Graphs: Generation of Benzenoid Hydrocarbons. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 167–173.
(12) Raznikov, V. V.; Talrose, V. L. Automatic Generation of Complete Set of Structural Isomers with a Given Molecular Composition and Molecular Weight (in Russian). *Zh. Struct. Chim.* **1970**, *11*, 357–360.
(13) Balaban, T. S.; Filip, P. A.; Ivanciuc, O. Computer Generation of Acyclic Graphs Based on Local Vertex Invariants and Topological Indices—Derived Canonical Labelling and Coding of Trees and Alkanes. *J. Math. Chem.* **1992**, *11*, 79–105.
(14) Smolenskij, E. A. About a Linear Denotation of Graphs (in Russian). *Zh. Vichisl. Math. Math. Fiz.* **1962**, *2* (No. 2), 371–372.
(15) Zaretskij, K. A. Construction of Trees Based on the Distances between Endpoints (in Russian). *Usp. Math. Nauk.* **1965**, *20* (No. 6), 90–92.
(16) Lukovits, I. Frequency of Even and Odd Numbers in Distance Matrices of Trees. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 626–629.