J. Chem. Inf. Comput. Sci. **2000,** *40,* 361−366

Isomer Generation: Semantic Rules for Detection of Isomorphism

István Lukovits†

Chemical Research Center, Hungarian Academy of Sciences, H-1525 Budapest, P.O. Box 17, Hungary

Received July 26, 1999

The problem of exhaustive generation of a nonredundant set of structural formulas (graphs) of acyclic alkanes was considered. A technique based on the compressed adjacency matrix (CAM) was proposed. The algorithm generates CAMs, which encode trees numbered according to the Morgan naming algorithm. Out of this set of CAMs those corresponding to the maximal Morgan codes—which in fact are the CAMs of canonically numbered isomers—must be found. CAMs are conceived as "sentences" of a primitive language. Sentences violating the syntactic and semantic rules of this language have to be discarded. In this paper three semantic rules have been proposed. The algorithm devised is efficient, and the decision to retain or to reject the actual structure does not involve any comparison with other structures. It was proved that several subsets of CAMs will not contain any maximal CAM and therefore it is not necessary to generate them. The whole procedure was illustrated by generating CAMs of all acyclic graphs containing nine vertices.

INTRODUCTION

Canonical numbering is a well-defined way of assigning numbers to vertices of graphs or to atoms in structural formulas. One of the first attempts to devise such an algorithm is due to Randić, who proposed a method based on the eigenvalues of the adjacency matrix.[1] Canonical numbering can be used to generate all nonredundant graphs (structural formulas) containing *N* vertices (atoms).[2] Isomer generation is closely related to the question of isomer enumeration. Although the latter problem has already been solved in the 30s,[3] the generation of formulas of isomers is investigated intensely because of its practical importance and has recently been reviewed.[4] Detection of isomorphism and automorphisms[5] is also related to canonical numbering. It was shown that isomorphism problems can be solved in polynomial time.[6,7] The technique used most frequently for isomer generation is the *N*-tuple code method[8,9] (NTCM). Other techniques used to generate structural formulas were reviewed in a previous paper.[10]

To ensure that all generated structures are different (i.e., we have a nonredundant series), most often time-consuming comparisons between various structures are performed.[11] We proposed a procedure by which it can be decided whether the generated code corresponds to a canonically labeled structure or not. Those codes that correspond to noncanonically labeled structures should be deleted. The method is based on the compressed adjacency matrix (CAM) technique.[10] The set of rules introduced previously to accomplish this task are syntactic rules, meaning that inspection of the codes *alone* is enough to determine whether the underlying structure is labeled according to the "lowest degrees first" (LDF) rule[10] and it is not necessary to restore the original adjacency matrix. The aim of the present paper was to propose *semantic* rules. By semantic we mean those procedures in which the adjacency matrix must be restored from the CAM. The algorithms based on semantic rules are efficient.

The structure of this paper is the following: In the next section we define physical trees, Morgan-trees, and maximal Morgan-trees. It will be shown that several classes of Morgan-trees can be excluded from the generation procedure as they will not contain maximal (or LDF) Morgan-trees. In the third section the syntactic rules proposed earlier[10] will be discussed briefly. The fourth section will be devoted to the new semantic rules. In the last section the results will be discussed. The proof of the "second bistar rule" appears in the Appendix. As an example, the generation of all acyclic graphs containing nine vertices will be considered.

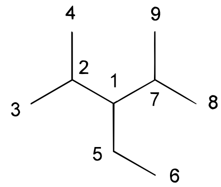
PHYSICAL TREES, MORGAN-TREES, AND MAXIMAL MORGAN TREES

One of the most basic notions of chemical graph theory[12] is the adjacency matrix **A**. **A** is an *N* by *N* array of numbers. The *i*,*j*th entry $A_{i,j} = 1$ if vertices (atoms) *i* and *j* are connected by an edge (being equivalent to a chemical bond), and $A_{i,j} = 0$ otherwise. Vertices connected by an edge are said to be adjacent. **A** is symmetric, i.e., $A_{i,j} = A_{j,i}$, and $A_{i,i} = 0$, for all *i* and *j*. Any structure may be represented by matrix **A**, and from **A** the underlying structure—including the numbering of the vertices—may be unambiguously restored. **A** can represent any structure, but in this paper we shall consider (hydrogen-suppressed) acyclic structures only. An acyclic structure will be referred to—in accordance with the usual practice—as a tree (T).

Physical trees[13] comprise a special subgroup of labeled structures. A physical tree denotes a numbered tree in which each vertex *i* has only one neighbor with a lower ordinal number. The only exception is vertex 1 itself. Matrix **A**, representing a physical tree, will contain a *single* nonzero entry in each column of the upper right-hand portion.[14] Figure 1 depicts a physical tree. The upper right-hand portion of the corresponding adjacency matrix is shown in Table 1. Because of the simple structure of **A**, it is rather easy to

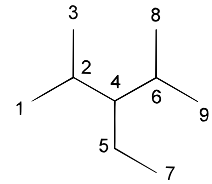
† Phone: +36-1-325-7900. Fax: +36-1-325-7554. E-mail: lukovits@cric.chemres.hu.

10.1021/ci990085r CCC: $19.00
Published on Web 01/19/2000



(1,2,2,1,5,1,7,7)

Figure 1. Physical tree.



(1,2,2,4,4,5,6,6)

Figure 2. Morgan-tree. Application of rule 1se.

Table 1. The Adjacency Matrix of a Physical Tree (Figure 1)[a]

0	1			1		1		
	0	1	1					
		0						
			0					
				0	1			
					0			
						0	1	1
							0	
								0

[a] The upper right-hand portion is shown only. Zeros, except those in the diagonal, were suppressed for the sake of clarity.

Table 2. The Adjacency Matrix of a Morgan-Tree (Figure 2)[a]

0	1							
	0	1	1					
		0						
			0	1	1			
				0		1		
					0		1	1
						0		
							0	
								0

[a] The upper right-hand portion is shown only. Zeros, except those in the diagonal, were suppressed for the sake of clarity.

generate physical trees. Let us consider an adjacency matrix, which contains nonzero entries only in the first row—this special matrix corresponds to a star. (A star is a structure having a single branching vertex and $N - 1$ end points attached to the branching vertex.) From this all numbered structures corresponding to a physical tree can be obtained by systematically changing the structure of **A** until all structures have been created. By examining Table 1 it is easy to realize that the number of physical trees containing N vertices is $(N - 1)!$.

All physical trees can be represented by using the compressed adjacency matrix (CAM) notation.[10,14] CAMs are row vectors containing $N - 1$ entries. Each entry k in the CAM corresponds to the nonzero entry in the $k + 1$ column of **A**; the number indicates the row number of the nonzero entry. The ith entry of a CAM is equal to the serial number of the vertex to which vertex $i + 1$ is attached (i.e., the only adjacent vertex with the serial number lower than $i + 1$). The CAM of a physical tree is shown in Figure 1. The CAM of any graph may be obtained by using the following algorithm (υ_i denotes the degree of vertex i): 1. Start with CAM = (1), indicating the obvious fact that vertex two is attached to vertex one. 2. Enter number two $v_2 - 1$ times. 3. Repeat step 2 in turn for all vertices keeping in mind that number k appears $\upsilon_k - 1$ times in the CAM. (The ordinal numbers of end points, except of vertex 1, therefore will not appear in the CAM.)

A tree fulfilling the restriction CAM(i) ≤ CAM(j) ($i < j$) is a Morgan-tree.[10] A Morgan-tree is obtained if—after assigning number 1 to a vertex—the nonlabeled adjacent vertices of the vertex with *lowest* serial number are labeled first.[2,15] A Morgan-tree and its CAM are shown in Figure 2. The corresponding adjacency matrix is shown in Table 2. Observe the "staircaselike" structure of **A**. In accordance the CAM of any Morgan-tree is a sequence of $N - 1$ nondecreasing integers.

The numbering is still not unique. However, it will be unique if a further rule is added to the Morgan naming algorithm: the vertex of *lowest* degree (valence) should be numbered first. If there are two vertices of the same degree, then the second neighbors determine the order, if these are also of equal degree, then the third neighbors should be considered, etc.[10] As a consequence of this rule (which will be referred to as the lowest degree first—LDF—rule), numbering should start at the end point of the longest side chain. It can be shown that the CAM codes of LDF-trees are maximal.[10,16] Morgan-trees having a maximal code will also be denoted as "maximal Morgan-trees". Maximal codes are unique; therefore, only maximal codes should be retained if a series of all isomers has to be generated and no replications of the same structures are allowed. It is possible to exclude several codes from the generation procedure (see below), but all remaining codes—which will be considered as sentences of a primitive language—have to be examined whether they are maximal or not. This can be achieved by using syntactic and semantic rules. Figure 3 lists all (compressed) CAMs of acyclic nonanes greater to or equal than (1,2,2,4,4,4,4,4), and in addition codes (1,2,2,2,2,2,2,2), (1,2,2,2,5,5,5,5), and (1,2,2,2,5,6,6,6). The last three CAMs denote maximal Morgan-trees (i.e., LDF trees), and it can be shown[10] that no LDF-tree ($N = 9$) with a CAM less than (1,2,2,2,2,2,2,2) exists. There are not maximal Morgan-trees between[10] (1,2,2,2,2,2,2,2) and (1,2,2,2,5,5,5,5). Similarly, there are not maximal Morgan-trees between (1,2,2,2,5,5,5,5) and (1,2,2,2,5,6,6,6) and also between (1,2,2,2,5,6,6,6) and (1,2,2,4,4,4,4,4). There are three rules by which structures can be excluded from the generation procedure; two of these (rules 1g and 2g, where g indicates that it affects the generation algorithm) have been proved in the previous paper,[10] whereas rule 3g will be proved in the Appendix of this paper.

Rule 1g. Do not generate any CAM containing more than a single entry equal to one. (Rule 1g will be referred to as the "end point convention".) After this rule is applied, the first CAM to be considered is $CAM_s = (1,2,...,2)$, where subscript s indicates that the CAM encodes a star.

Rule 2g. There is no maximal CAM_x such that $CAM_s < CAM_x < CAM_b$, where CAM_b denotes the first "bistar",

1 2 2 2 2 2 2 2 2	1 2 2 4 4 6 7 7	1 2 3 3 3 3 5 7 Δ	1 2 3 3 3 5 7 8 Δ
1 2 2 2 5 5 5 5	1 2 2 4 4 6 7 8 Δ	1 2 3 3 3 3 5 8 Δ	**1 2 3 3 3 6 6 6**
1 2 2 2 5 6 6 6	**1 2 2 4 5 5 5 5**	1 2 3 3 3 3 6 6 Δ	1 2 3 3 3 6 6 7 Δ
1 2 2 4 4 4 4 4	1 2 2 4 5 5 5 6 Δ	**1 2 3 3 3 3 6 7**	1 2 3 3 3 6 6 8
1 2 2 4 4 4 4 5 Δ	1 2 2 4 5 5 5 7 Δ	1 2 3 3 3 3 6 8 Δ	**1 2 3 3 3 6 7 7**
1 2 2 4 4 4 4 6 Δ	1 2 2 4 5 5 5 8 Δ	**1 2 3 3 3 3 7 7**	1 2 3 3 3 6 7 8 •
1 2 2 4 4 4 4 7 Δ	1 2 2 4 5 5 6 6 Δ	1 2 3 3 3 3 7 8 •	1 2 3 3 4 4 4 4 Δ
1 2 2 4 4 4 4 8 Δ	1 2 2 4 5 5 6 7 Δ	1 2 3 3 3 4 4 4 Δ	1 2 3 3 4 4 4 5 Δ
1 2 2 4 4 4 5 5 Δ	1 2 2 4 5 5 6 8 Δ	1 2 3 3 3 4 4 5 Δ	1 2 3 3 4 4 4 6 Δ
1 2 2 4 4 4 5 6 Δ	**1 2 2 4 5 5 7 7**	1 2 3 3 3 4 4 6 Δ	1 2 3 3 4 4 4 7 Δ
1 2 2 4 4 4 5 7 Δ	1 2 2 4 5 5 7 8 Δ	1 2 3 3 3 4 4 7 Δ	1 2 3 3 4 4 4 8 Δ
1 2 2 4 4 4 5 8 Δ	**1 2 2 4 5 6 6 6**	1 2 3 3 3 4 4 8 Δ	**1 2 3 3 4 4 5 5**
1 2 2 4 4 4 6 6 Δ	1 2 2 4 5 6 6 7 Δ	1 2 3 3 3 4 5 5 Δ	1 2 3 3 4 4 5 6 Δ
1 2 2 4 4 4 6 7 Δ	1 2 2 4 5 6 6 8 Δ	**1 2 3 3 3 4 5 6**	1 2 3 3 4 4 5 7
1 2 2 4 4 4 6 8 Δ	**1 2 2 4 5 6 7 7**	1 2 3 3 3 4 5 7 Δ	1 2 3 3 4 4 5 8
1 2 2 4 4 4 7 7	1 2 2 4 5 6 7 8 Δ	1 2 3 3 3 4 5 8 Δ	1 2 3 3 4 4 6 6 Δ
1 2 2 4 4 4 7 8 •	**1 2 3 3 3 3 3 3**	1 2 3 3 3 4 6 6 Δ	1 2 3 3 4 4 6 7 Δ
1 2 2 4 4 5 5 5 Δ	1 2 3 3 3 3 3 4 Δ	1 2 3 3 3 4 6 7 Δ	1 2 3 3 4 4 6 8 Δ
1 2 2 4 4 5 5 6 Δ	1 2 3 3 3 3 3 5 Δ	1 2 3 3 3 4 6 8 Δ	1 2 3 3 4 4 7 7 Δ
1 2 2 4 4 5 5 7 Δ	1 2 3 3 3 3 3 6 Δ	1 2 3 3 3 4 7 7 Δ	1 2 3 3 4 4 7 8 •
1 2 2 4 4 5 5 8 Δ	1 2 3 3 3 3 3 7 Δ	1 2 3 3 3 4 7 8 Δ	**1 2 3 3 4 5 5 5**
1 2 2 4 4 5 6 6	**1 2 3 3 3 3 3 8**	1 2 3 3 3 5 5 5 Δ	1 2 3 3 4 5 5 6 Δ
1 2 2 4 4 5 6 7 Δ	1 2 3 3 3 3 4 4 Δ	1 2 3 3 3 5 5 6 Δ	1 2 3 3 4 5 5 7 Δ
1 2 2 4 4 5 6 8 Δ	1 2 3 3 3 3 4 5 Δ	1 2 3 3 3 5 5 7 Δ	1 2 3 3 4 5 5 8
1 2 2 4 4 5 7 7 Δ	1 2 3 3 3 3 4 6 Δ	1 2 3 3 3 5 5 8 Δ	1 2 3 3 4 5 6 6
1 2 2 4 4 5 7 8 Δ	1 2 3 3 3 3 4 7 Δ	**1 2 3 3 3 5 6 6**	1 2 3 3 4 5 6 7
1 2 2 4 4 6 6 6	1 2 3 3 3 3 4 8 Δ	1 2 3 3 3 5 6 7	1 2 3 3 4 5 6 8
1 2 2 4 4 6 6 7 Δ	1 2 3 3 3 3 5 5 Δ	1 2 3 3 3 5 6 8	**1 2 3 3 4 5 7 7**
1 2 2 4 4 6 6 8 Δ	1 2 3 3 3 3 5 6 Δ	1 2 3 3 3 5 7 7 Δ	1 2 3 3 4 5 7 8 •
1 2 3 3 4 6 6 6 Δ	**1 2 3 3 5 6 6 8**	**1 2 3 4 4 4 7 8**	1 2 3 4 5 5 5 6 Δ
1 2 3 3 4 6 6 7 Δ	**1 2 3 3 5 6 7 7**	1 2 3 4 4 5 5 5 Δ	1 2 3 4 5 5 5 7 Δ
1 2 3 3 4 6 6 8 Δ	1 2 3 3 5 6 7 8 •	1 2 3 4 4 5 5 6 Δ	**1 2 3 4 5 5 5 8**
1 2 3 3 4 6 7 7 Δ	**1 2 3 4 4 4 4 4**	1 2 3 4 4 5 5 7 Δ	1 2 3 4 5 5 6 6 Δ
1 2 3 3 4 6 7 8 •	1 2 3 4 4 4 4 5 Δ	1 2 3 4 4 5 5 8 Δ	**1 2 3 4 5 5 6 7**
1 2 3 3 5 5 5 5	1 2 3 4 4 4 4 6 Δ	**1 2 3 4 4 5 6 6**	1 2 3 4 5 5 6 8 Δ
1 2 3 3 5 5 5 6 Δ	1 2 3 4 4 4 4 7 Δ	1 2 3 4 4 5 6 7	**1 2 3 4 5 5 7 7**
1 2 3 3 5 5 5 7 Δ	**1 2 3 4 4 4 4 8**	**1 2 3 4 4 5 6 8**	**1 2 3 4 5 5 7 8**
1 2 3 3 5 5 5 8	1 2 3 4 4 4 5 5 Δ	1 2 3 4 4 5 7 7 Δ	**1 2 3 4 5 6 6 6**
1 2 3 3 5 5 6 6 Δ	1 2 3 4 4 4 5 6 Δ	1 2 3 4 4 5 7 8 Δ	1 2 3 4 5 6 6 7 Δ
1 2 3 3 5 5 6 7	1 2 3 4 4 4 5 7 Δ	**1 2 3 4 4 6 6 6**	**1 2 3 4 5 6 6 8**
1 2 3 3 5 5 6 8 Δ	1 2 3 4 4 4 5 8 Δ	1 2 3 4 4 6 6 7 Δ	**1 2 3 4 5 6 7 7**
1 2 3 3 5 5 7 7	1 2 3 4 4 4 6 6 Δ	**1 2 3 4 4 6 6 8**	**1 2 3 4 5 6 7 8**
1 2 3 3 5 5 7 8 •	**1 2 3 4 4 4 6 7**	**1 2 3 4 4 6 7 7**	
1 2 3 3 5 6 6 6	1 2 3 4 4 4 6 8 Δ	1 2 3 4 4 6 7 8 •	
1 2 3 3 5 6 6 7 Δ	**1 2 3 4 4 4 7 7**	**1 2 3 4 5 5 5 5**	

Figure 3. Compressed adjacency matrixes of 177 acyclic structures containing nine vertices. Commas separating the entries have been omitted. Codes deleted by using syntactic rules 1sy and 2sy have been marked on the right-hand side by using the symbols Δ, and ●, respectively. Codes in **boldface** denote maximal (LDF) codes, whereas remaining plain codes denote structures deleted by using semantic rules 1se, 2se, and 3se.

$CAM_b = (1,2,2,...,2,X,...,X)$ with k digits equal to two and m numbers equal to X, $k = m = (N - 2)/2$ if N is even, and $k + 1 = m = (N - 2)/2$, if N is odd, and $X = N - k$. (Rule 2g will be referred to as the "bistar rule".)

Rule 3g. There is no maximal CAM_x such that $CAM_b < CAM_x < CAM_B$, where CAM_B denotes the "second bistar", if N is even and $CAM_B = (1,2,2,...,2,X',...,X')$, and there are $k - 1$ digits equal to two and $m + 1$ numbers equal to X', and $X' = X - 1$. If N is odd there is a single exception: $CAM = (1,2,2,...,2,X',X,...,X)$ is a maximal Morgan-tree, where k is identical with k in the case N is even. (Rule 3g will be referred to as the "second bistar rule".) The proof is given in the Appendix.

Figure 4 compares the number of labeled trees. The number of Morgan-trees is just a small fraction of physical trees, and this number will be further reduced by the application of rules 1g, 2g, and 3g.

SYNTACTIC RULES

By syntactic rules we shall denote those precepts that apply to CAMs but do not make use of the restored adjacency matrix. The syntactic rules 1sy and 2sy (where "sy" refers to syntactic) used in this paper have been proved earlier.[10] Rules 1sy and 2sy are concise versions of the rules devised in ref 10.

Rule 1sy. Delete any $CAM = (1,2,...,Y,...,Y,...)$ with Y appearing k times if any Y in position k_1 is referred to (i.e., number $k_1 + 1$ appears in the CAM) whereas a Y in position k_2 ($k_2 > k_1$) is not.

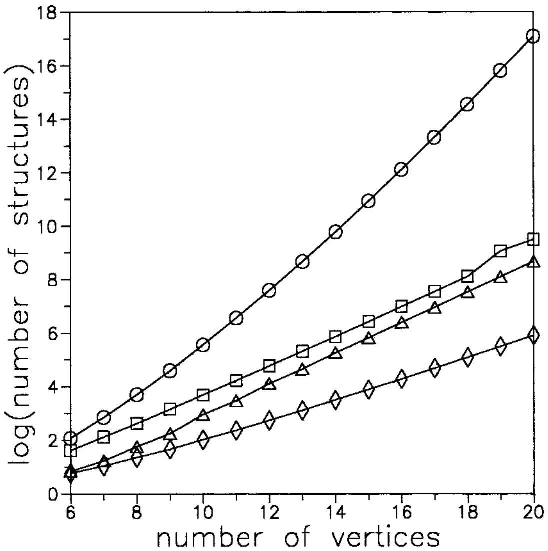


Figure 4. Number of physical trees (○), number of Morgan-trees (□), number of Morgan-trees after applying rules 1g, 2g, and 3g (△), number of LDF-trees − true isomers (◊).

Example: Consider CAM = (1,2,2,4,4,4,5,5). The fourth, fifth, and sixth entries are equal to 4, indicating that three vertices (vertices 5, 6, and 7) are attached to vertex 4 (vertex 4 itself is attached to vertex 2). Vertex 5 appears twice in the CAM, but vertices 6 and 7 do not, indicating that vertex 5 is of degree three whereas vertices 6 and 7 are end points. This means that the LDF rules have been violated and the structure must be deleted.

Rule 2sy. Delete any CAM where the path between first vertex and the first branching vertex is shorter than the path connecting the last vertex with the nearest branching vertex.

Example: Consider CAM = (1,2,2,4,4,4,7,8). The length of the path connecting vertex 1 and the first branching vertex (vertex 2) is equal to one. On the other hand vertex 9 is attached to vertex 8, vertex 8 is attached to vertex 7, and vertex 7 is attached to vertex 4 (which is a branching vertex), so the length of this path is equal to three. Thus the LDF rules have been violated, and the code must be deleted.

After codes are eliminated by rules 1sy and 2sy, the number of remaining codes to be investigated (Figure 3) is 59. These will be examined by using the semantic rules (see next section).

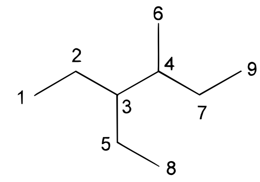
SEMANTIC RULES

The remaining 12 nonmaximal codes are (Figure 3) (1,2,2,4,4,5,6,6), (1,2,2,4,4, 6,7,7), (1,2,3,3,3,5,6,7), (1,2,3,3, 3,5,6,8), (1,2,3,3,3,6,6,8), (1,2,3,3,4,4,5,7), (1,2,3,3,4,4,5,8), (1,2,3,3,4,5,5,8), (1,2,3,3,4,5,6,6), (1,2,3,3,4,5,6,7), (1,2,3,3, 4,5,6,8), and (1,2,3,4,4,5,6,7). The underlined codes are shown in Figures 2, 5, and 6, respectively. The CAMs of the underlying structures do not violate the syntactic rules 1sy and 2sy, and semantic rules 1se, 2se, and 3se will be used to determine which of these structures correspond to a LDF-tree. Extensive use will be made of series of degrees of vertices, and to set apart these from CAMs the new strings will have the form $a-b-c-,...,-z$, where a, b, c, and z denote degrees of vertices. This set of numbers will be referred to as *valence code*.

Rule 1se. Consider paths connecting vertex 1 (an end point) and any other end point x of T in turn, and write down the valences belonging to this path starting with the degree of vertex 1, then with the degree of its first neighbor, then with the degree if its second neighbor, etc. In this way a valence code has been created. Repeat this procedure starting with vertex x of the same path. If the "reversed" valence code is less than the original code, delete the underlying CAM.

As an example consider the structure depicted in Figure 2. The valence code of path 1−7 is 1−3−3−2−1. Starting the list of degrees at vertex 7, the valence code is 1−2−3−3−1. Valence code 1−2−3−3−1 is less than valence code 1−3−3−2−1. Therefore the underlying structure is not an LDF-tree, and CAM = (1,2,2,4,4,5,6,6) cannot be a maximal code. The maximal CAM of this structure is (1,2,3,3,4,4,5,5). Similarly rule 1se may be used to discard codes (1,2,2,4,4,6,7,7), (1,2,3,3,3,5,6,7), (1,2,3,3,3,5,6,8), (1,2,3,3,3,6,6,8), (1,2,3,3,4,4,5,8), (1,2,3,3,4,5,6,7), and (1,2,3,3,4,5,6, 8).

Rule 2se. Compare the valence codes of paths $1-x$ and $1-y$, where x and y denote end points of T. Start the procedure at vertex 2, which must appear in both paths. Compare the next pair of vertices i and j, with i and j positioned on paths $1-x$ and $1-y$, respectively, and $d_{2,i} = d_{2,j} = 1$, where $d_{k,m}$ denotes the (topological) distance between vertices k and m. If $\upsilon_i = \upsilon_j$ and $i \le j$ (the last equality applies only if there is a common portion of paths $1-x$ and $1-y$), then continue the procedure with the next pair of vertices m, n (i.e., $d_{2,m} = d_{2,n} = 2$). If $\upsilon_m > \upsilon_n$ and $m < n$, then check whether m and n "belong" to the same branching vertex or not. (This formulation denotes the nearest branching vertex with a lower ordinal number than m or n.) If vertices m and n belong to the same branching vertex, discard the structure since the LDF rules have been violated. If m and n belong to different pair of branching vertices, proceed with the next pair of vertices. Compare all paths starting from vertex 1 and ending with an end point with one another.

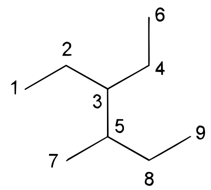


(1,2,3,3,4,4,5,7)

Figure 5. Application of rule 2se.

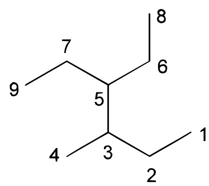
As an example consider the structure depicted in Figure 5. Compare paths 1−6 and 1−8. The degree of vertex 4 is three while the degree of vertex 5 is two, and both belong to branching vertex 3. Therefore the underlying structure is not an LDF-tree, and the corresponding CAM (1,2,3,3,4,4,5,7) should be discarded. The corresponding maximal CAM is (1,2,3,3,5,5,6,7). Rule 2se may also be used to discard codes (1,2,3,3,4,5,6,6) and (1,2,3,4,4,5,6,7).

Rule 3se. Consider any end point K ($K \neq 1$) in T. Renumber the vertices by using the following seven-step algorithm: 1. Assign number 1 to vertex K. 2. Assign number 2 to the only neighbor of K. 3. Set $k = 2$, where k is a running index. 4. Select all vertices X for which $d_{K,X} = k$. (All subscripts refer to the original numbering.) If there are no



(1,2,3,3,4,5,5,8)

Figure 6. Application of rule 3se.



(1,2,3,3,5,5,6,7)

Figure 7. Renumbered tree shown in Figure 6. The tree is now a LDF-tree.

vertices fulfilling condition $d_{K,X} = k$, then go to step 7. 5. Assign the next ordinal numbers to the vertices in the same order as indicated by the "old" ordinal numbers, unless $d_{1,X} = d_{1,K} - d_{K,X}$. In this case X is the last vertex to be numbered. If $d_{1,K} = k$, inspect the first neighbors X' of X. If any $X' = 2$, postpone numbering until all vertices that are not adjacent to vertex 2 have been considered. Then vertex 1 and then the remaining vertices fulfilling condition $d_{1,K} = k$ should be numbered. 6. Increase k by one and go to step 4. 7. Stop the naming procedure.

On the basis of the new numbering, obtain the modified compressed adjacency matrix CAM′. If CAM′ > CAM, discard CAM.

As an example consider the structure depicted in Figure 6. The renumbered tree is shown in Figure 7. Since CAM′ > CAM, the structure must be discarded. CAM′ is a code of an LDF-tree.

CONCLUSIONS

Let us compare the number of labeled acyclic nonanes. The total number of labeled (acyclic) nonanes is $9^7 = 4\,782\,969$, the number of physical trees is $8! = 40\,320$, the number of Morgan-trees is[10] 1430, the number of Morgan-trees after discarding all structures in which vertex 1 is not an end point[10] is 429; the number of Morgan-trees is after applying the "bistar rule", 279, and it is after applying the "second bistar rule" 177. Therefore we have to inspect 177 structures in the case of nonanes (Figure 3). Most of the remaining non-LDF-trees could have been detected by using syntactic rules, and only a small fraction of trees had to be examined by using the semantic rules. However, with increasing N, the portion of trees to be examined by using semantic rules will increase. It should be noted that if the valence of the vertices is restricted to four (i.e., we want to generate the structures of all alkane-isomers containing N atoms), a further syntactic rule—namely, which excludes all CAMs containing more than tree identical entries—is needed.

The inspection of each CAM by using the syntactic rules 1sy and 2sy will take linear time. The time needed to process any CAM by using the semantic rules 1se, 2se, and 3se depends on the number of end points. Both steps are therefore efficient. Whether semantic rules *alone* are sufficient to discard all redundant structures will be considered in a forthcoming publication.[17]

The generation of isomers containing N vertices can be solved in exponential time, as the number of isomers increases exponentially with N. The ratio of Morgan-trees remaining after application of the "second bistar rule" to the number of true isomers (LDF-trees) is a polynomial in terms of N (Figure 4). This means that the procedure proposed in this paper is efficient. On the other hand, by using NTCM, each time it has to be checked, whether the new code denotes a tree (which in most cases is not a Morgan-tree) or not. There are no estimates how many structures have to be generated (and inspected) by using NTCM.

APPENDIX

Proof of Rule 3g. Let N be even and $N \geq 8$. Then $CAM_b = (1,2,...,2,X,...,X)$ with k digits equal to 2 and m numbers equal to X, where $k = m = (N - 2)/2$ and $X = N - m$. Increase one or more (maximally $m - 1$) X by one. Then we have $CAM'_b = (1,2,...,2,X,X',...,X')$, where $X' = X + 1$. But CAM'_b cannot represent a LDF-tree since vertex 1 is adjacent to a vertex of a higher degree than vertex N. Increase the last entry equal to 2 by one. Then we have $CAM'_b = (1,2,...,2,X',X,...,X)$ with $k' = k - 1$ digits equal to 2 and k numbers equal to X, where $X = N - m$ and $X' = 3$. But this CAM'_b does not represent an LDF-tree either as rule 1sy has been violated: X' refers to the first of the digits equal to two, while all Xs refer to X'. The same arguments are valid for all Xs, if $X' < X$. Therefore the next CAM > CAM_b representing a LDF-tree is $CAM_B = (1,2,...,2,X',...,X')$ with k' digits equal to 2 and m' numbers equal to X' in CAM_B, and $k' = k - 1$, $m' = m + 1$, and $X' = N - m - 1$. In the case of octanes $CAM_b = (1,2,2,2,5,5,5)$ and $CAM_B = (1,2,2,4,4,4,4)$, and $k' = 2$, $m' = 4$.

Let N be odd and $N \geq 9$. Then $CAM_b = (1,2,...,2,X,...,X)$ with k digits equal to 2 and m numbers equal to X, where k $= (N - 3)/2$, $m = k + 1$, and $X = N - m = N - k - 1$. Increase the last $m - 1$ numbers by one: $X' = X + 1$. Then we have $CAM^* = (1,2,...,2,X,X',...,X')$, and this CAM* is a LDF-tree. In the case of nonanes, $CAM_b = (1,2,2,2,5,5,5,5)$ and $CAM^* = (1,2,2,2,5,6,6,6)$. But X's cannot be increased further since no CAM = (1,2,2,2,5,7,7,7) exists, so a CAM = (1,2,2,2,5,5,7,7) is possible only. Here we have a tree with vertex 1 adjacent to a vertex of higher degree than vertex N. So we have to increase the last digit equal to 2 by one, and we have $CAM'_b = (1,2,...,2,X',X,...,X)$, where $X' = 3$, and $k - 1$ entries are equal to two, and $k + 1$ entries are equal to X. In this case rule 1sy has been violated. The same arguments apply to all X's if $X' < X - 1$. If $X' = X - 1$, then we have a bistar and CAM_B represents a LDF-tree. In the case of nonanes $CAM_B = (1,2,2,4,4,4,4,4)$. This completes the proof.

REFERENCES AND NOTES

(1) Randić, M. On Unique Numbering of Atoms and Unique Codes for Molecular Graphs. *J. Chem. Inf. Comput. Sci.* **1975**, *15*, 105−108.

(2) Kvasnička, V.; Pospichal, J. Canonical Indexing and Constructive Enumeration in Molecular Graphs. *J. Chem. Inf. Comput. Sci*. **1990**, *30*, 99−105.

(3) Read, R. C. Algorithms in Graph Theory. In *Chemical Applications of Graph Theory*; Balaban, A. Q. T., Ed.; Academic Press: London, 1976; pp 25−61.

(4) Bytautas, L.; Klein, D. J. Chemical Combinatorics for Alkane-Isomer Enumeration and More. *J. Chem. Inf. Comput. Sci*. **1998**, *38*, 1063−1078.

(5) Oujang, Z.; Shengang, S. Y.; Brand, J.; Zheng, C. An Effective Topological Symmetry Perception and Unique Numbering Algorithm. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 299−303.

(6) Faulon, J. L. Isomorphism, Automorphism Partitioning, and Canonical Labeling Can be Solved in Polynomial Time. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 432−444.

(7) Fan, B. T.; Panaye, A.; Douced, J. P. Comment on "Isomorphism, Automorphism Partinioning and Canonical Labeling Can be Solved in Polynomial Time for Molecular Graphs". *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 630−631.

(8) Contreras, M. L.; Trevisiol, G. M.; Alvarez, J.; Arias, G.; Rozas, R. Exhaustive Generation of Organic Isomers. 5. Unsaturated Optical and Geometrical Stereoisomers and a new CIP Subrule. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 475−482.

(9) Knop, J. V.; Müller, W. R.; Jericević, Z.; Trinajstić, N. Computer Enumeration and Generation of Trees and Rooted Trees. *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 91−99.

(10) Lukovits, I. Isomer Generation: Syntactic Rules for Detection of Isomorphism. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 563−568.

(11) Balaban, T. S.; Filip, P. A.; Ivanciuc, O. Computer Generation of Acyclic Graphs Based on Local Vertex Invariants and Topological Indexes. Derived Canonical Labeling and Coding of Trees and Alkanes. *J. Math. Chem*. **1992**, *11*, 79−105.

(12) Trinajstić, N. *Chemical Graph Theory*, 2nd ed.; CRC Press: Boca Raton, FL, 1992; p 1.

(13) 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.

(14) Gutman, I.; Linert, W.; Lukovits, I.; Dobrynin, A. A. Trees with Extremal Hyper-Wiener Index: Mathematical Basis and Chemical Applications. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 349−354.

(15) Morgan, H. L. The Generation of a Unique Description for Chemical Structures.−A Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5*, 107−113.

(16) Let us consider two vectors **A** and **B**, each containing k elements. It will be said that **A** is greater than **B** if for any pair i and j, $1 = i < j = k$ and $A_i = B_i$, but $A_j > B_j$, see also ref 13.

(17) Lukovits, I. To be published.

CI990085R