A Compact Form of the Adjacency Matrix

István Lukovits[†]

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

Received December 26, 1999

It has been shown that the adjacency matrix can be transformed into a row vector and then into a single number. This number can again be decoded to recover the row vector, and this in turn can be decoded to restore the original adjacency matrix. A special, rather efficient coding scheme was devised for acyclic structures.

INTRODUCTION

The adjacency matrix **A** is one of the most important concepts used in chemical graph theory. It is an $N \times N$ array of numbers, such that $\mathbf{A}_{i,j} = 1$ if vertices i and j are connected by an edge—or, in other words, are adjacent,—and $\mathbf{A}_{i,j} = 0$ in all other cases, where N denotes the number of vertices in the graph. Expressions "graph" and "structural formula" as well as "vertex" and "atom" and "edge" and "bond" are considered to be synonyms in this paper. **A** is a symmetric matrix, since $\mathbf{A}_{i,j} = \mathbf{A}_{j,i}$, and $\mathbf{A}_{i,i} = 0$ (i, $j = 1$, ..., N). Once **A** is known, the underlying structure can be reconstructed. Those, who are interested in the mathematical properties of **A**, are referred to the literature.[1,2]

Although **A** contains all the information needed to reconstruct the underlying graph, its two-dimensional structure renders it impractical for computer storage. There are two methods[3] used to replace **A** by simpler designs: 1. the list of bonds (each item contains the ordinal numbers of the incident vertices) and 2. the list of the neighbors of the vertex i ($i = 1$, ..., N). None of these methods has been accepted generally.

Several attempts to represent structures (molecules) by numbers and to avoid **A** at the same time have been reported. These include the DAST code[4] and the boundary code[5] designed to represent polyhexes, the N-tuple code designed to represent trees,[6] and the binary code.[7] Any of these codes may be used to reconstruct the underlying adjacency matrix.

Graph invariants would be ideal to represent structures, but the reconstruction of the underlying structure is complicated in most cases. On the other hand, no graph invariant could be devised so far, that is completely discriminative (meaning that no pairs of structures exist with an identical value of the graph invariant). Recently Hu and Xu, after inspecting more than 400 000 compounds, claimed that their invariant *is* completely discriminative, but trials, no matter how extensive they may be, do not amount to a rigorous mathematical proof.[8,9]

The most direct procedure to "condense" matrix **A** would be to conceive its right-hand upper part as a single, binary number consisting of $N(N-1)/2$ digits. Because **A** is symmetric, this number is sufficient to restore **A**. Clearly, the maximal number would contain $N(N-1)/2$ digits equal to one, and this corresponds to a complete graph (i.e. to a graph in which all vertices are adjacent). The binary number can be transformed into a decimal number. The maximal binary number is therefore equal to $2^{N(N-2)/2} - 1$, which will contain $[N(N-1)/2 \log 2] + 1 \cong [0.30N(N-1)/2] + 1$ digits in the decimal system. (The brackets indicate that the integer part of the expression has to be taken into account, only.) In addition the code must contain the number of vertices. This method of coding is therefore not quite efficient, but its performance could be improved by using instead of the decimal system another number system with a higher base. Each decimal number associated with an adjacency matrix is in fact the ordinal number of respective matrix. The maximal ordinal number is equal to the number of possible $N \times N$ adjacency matrices minus one. This means that there is no other method, which is capable of representing **A** of a general graph with fewer digits in the decimal system. If acyclic graphs (trees) are considered and if special constraints with respect to the numbering of the vertices are introduced (see next section), then a more efficient coding can be devised.

In the next section an alternative scheme (which, however, is not more efficient than the method of binary numbers outlined above) to encode general adjacency matrices will be devised. Then, based on this scheme, a method to represent adjacency matrices of acyclic trees by a single number will be proposed.

COMPACT VERSIONS OF THE ADJACENCY MATRIX

A is symmetric; therefore, it is enough to consider the upper triangle in **A**. Since it contains zeros and ones only, the columns of **A**—i.e., those portions which are above the diagonal—can be conceived as a single number in the binary number system. Consider the structure depicted in Figure 1 (hydrogen atoms were neglected throughout this paper). Each column (Table 1) of matrix **A** (i.e. those portion which lies above the diagonal) may be conceived to represent a binary number. The number of entries in BIN is $N - 1$, and the corresponding decimal numbers are BIN = (0, 1, 2, 0, 29, 6).

It is evident that the information contained in BIN is sufficient to restore **A**. For this each decimal number of **A** is converted into its binary equivalent, and the ith entry of

[†] Corresponding author phone: +36-1-325-7900; fax: +36-1-135- 2148; e-mail: lukovits@cric.chemres.hu.

10.1021/ci990152h CCC: $19.00 © 2000 American Chemical Society
Published on Web 08/06/2000

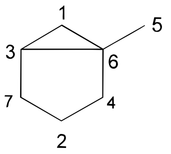

Figure 1. A structure used as an example of the coding procedure. The compact form of its adjacency matrix (Table 2) is equal to $\mathbf{A}_0 = 329542_7$.

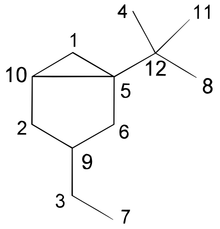

Figure 2. A structure used as an example of the coding procedure. The compact form of its adjacency matrix is equal to $\mathbf{A}_0 = 108227168313541786_{12}$.

BIN becomes the $(i + 1)$th column of $\mathbf{A}$. The concept of BIN presented in this paper may be used for any graph, both acyclic as well as cyclic, whereas a method proposed previously,[10,11] the compressed adjacency matrix (CAM), may be applied for physical trees (see below) only.

The minimal value of any entry of BIN is zero; the maximal value is $2^i - 1$ ($i = 1, ..., N - 1$). Therefore the maximal values of the first three entries of BIN are as follows: 1, 3, and 7, respectively. The (connected) graph with the minimal BIN is the star (A star is a tree containing $N - 1$ endpoints and a single branching vertex of degree $N - 1$.). N labels the branching vertex, and numerals 1 through $N - 1$ label the endpoints. The graph possessing maximal BIN is, as mentioned before, the complete graph.

BIN codes may further be transformed into a single number $\mathbf{A}_0$. If the graph (Figure 1 and Table 1) contains seven vertices, then

$$\mathbf{A}_0 = 64(32|16\{8[4\mathrm{BIN}(1) + \mathrm{BIN}(2)] + \mathrm{BIN}(3)\} + \mathrm{BIN}(4)| + \mathrm{BIN}(5)) + \mathrm{BIN}(6) \quad (1)$$

Therefore

$$\mathbf{A}_0 = 64\times(32\times|16\times\{8\times[4 \times 0 + 1] + 2\} + 0| + 29) + 6 = 329542_7$$

The subscript identifies the number of vertices. A similar formula may be devised for more complicated (i.e. with $N > 7$) cases.

To decode $\mathbf{A}_0$ we proceed in the reverse order. Again the structure depicted in Figure 1 will serve as an example. First, divide 329 542 by 64. (In general the first divisor is 2^{N-1}.) The rest is 6. Therefore BIN(6) = 6. Then divide the integral part of the quotient (5149) by 32; the rest is 29. Therefore BIN(5) = 29. Proceed in this way, the last divisor is four, and the rest is equal to BIN(2). The integral part of the last quotient will be either zero or one, which is stored in BIN(1). Once the row vector of BIN is known, the original adjacency matrix can be restored as outlined above. In addition to this example, Figure 2 shows a second illustration.

The adjacency matrix of any tree contains just $N - 1$ ones in the upper right-hand triangle, and because of this fact a more efficient coding technique could be devised for this kind of structures. As an example consider 3-methylhexane (Figure 3). Figure 3 is an illustration of a "physical tree".[5] In a physical tree each vertex has only a single neighbor with a lower ordinal number. The only exception is vertex 1. As a consequence of this numbering rule, the adjacency matrix (i.e. the upper right-hand triangle) will contain a single 1 in each column, and there is no column containing zeros only. To ensure that the numbering of vertices results in a physical tree the following rules have to be obeyed: 1. Assign number one to any vertex. 2. Allocate the following number ($i = 2, ..., N$) in turn to any nonlabeled vertex, which has one numbered neighbor already. In this way vertex 2 will *always* be a neighbor of vertex 1 and $\mathbf{A}_{1,2} = 1$ for any physical tree. Figure 4 is an illustration of a "nonphysical tree". The coding procedure outlined above yields $\mathbf{A}_0 = 1646688_7$ for the structure depicted in Figure 3, and $\mathbf{A}_0 = 79944_7$ for the tree shown in Figure 4.

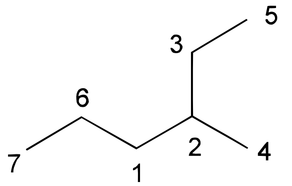

Figure 3. A physical tree: 3-methylhexane. The compact forms of the adjacency matrix are $\mathbf{A}_0 = 1646688_7$ and ${}_0\mathbf{A} = {}_7545$.

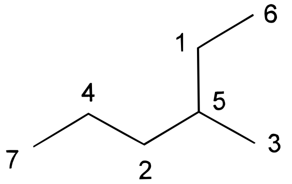

Figure 4. A nonphysical tree: 3-methylhexane. The compact form of the adjacency matrix is $\mathbf{A}_0 = 799447$ and ${}_0\mathbf{A}$ does not exist.

Since adjacency matrices of physical trees contain a single nonzero figure in each column (in the upper right-hand triangle), matrix $\mathbf{A}$ may be replaced by another abbreviated version, the compressed adjacency matrix[10,11] (CAM). CAM is a row vector (like BIN), and CAM(1) is equal to the row-number of the nonzero entry in the *second* column of $\mathbf{A}$. Similarly, CAM(2) is equal to the row-number of the nonzero entry in the third column of $\mathbf{A}$, etc., CAM(i) is equal to the row-number of the nonzero entry in the $(i + 1)$th column of $\mathbf{A}$. Because CAMs are related to physical trees, CAM(1) = 1. It is clear that from any CAM the underlying matrix $\mathbf{A}$ can be restored. The CAM of 3-methylhexane (Figure 3 and Table 2) is CAM = (1, 2, 2, 3, 1, 6), its BIN code is BIN = (1, 2, 2, 4, 1, 32).

Let us introduce the following notation: for any $i = 1,..., N - 1$

$$X_i = \mathrm{CAM}(i+1) \quad (2)$$

Then the following coding equation can be set up, which is valid for heptanes

$${}_0\mathbf{A} = 6|5\{4[3(X_1-1) + (X_2-1)] + (X_3-1)\} + (X_4-1)| + X_5 - 1 \quad (3)$$

Table 1. Upper Triangle of the Adjacency Matrix of the Structure Shown in Figure 1 and the Binary Code (Last, Separate Line)[a]

0		1			1	
	0		1			1
		0			1	1
			0		1	
				0	1	
					0	
						0
-	0	1	2	0	29	6

[a] Zeros, except in the diagonal, were suppressed for the sake of clarity.

Taking structure shown in Figure 3 as an example, we obtain

$$_0\mathbf{A} = 6|5\{4[3(2-1) + (2-1)] + (3-1)\} + (1-1)| + 6-1 = {}_7545$$

The subscripts appear now on the left-hand side of the code to distinguish it from the number obtained by using BIN codes. The decoding proceeds in a reversed fashion; $_0\mathbf{A} = {}_7545$ will be taken as an example. The steps are listed below (the rests of the divisions are given in parentheses)

1. $545/6 = 90\ (5)$
2. $5 + 1 = 6 = X_5$
3. $90/5 = 18\ (0)$
4. $0 + 1 = 1 = X_4$
5. $18/4 = 4\ (2)$
6. $2 + 1 = 3 = X_3$
7. $4/3 = 1\ (1)$
8. $1 + 1 = 2 = X_2$
9. $1/2 = 0\ (1)$
10. $1 + 1 = 2 = X_1$

and we have obtained the entries of the CAM (see eq 2), except CAM(1), which is equal to 1.

DISCUSSION

The value of $\mathbf{A}_0$ and $_0\mathbf{A}$ increase exponentially with N $-$ as the number of isomers does. Therefore special techniques are needed to handle graphs containing more than 15 vertices.[12] Structures containing different numbers of vertices may possess identical values of $\mathbf{A}_0$ or $_0\mathbf{A}$; therefore, the number of vertices must be attached to the respective code.

It has to be emphasized that $\mathbf{A}_0$ and $_0\mathbf{A}$ are *not* graph invariants since their value *does* depend on the numbering of the vertices; therefore, they cannot be used in quantitative structure-activity relationships.

The compact form of the adjacency matrix of 3-methylhexane (Figure 3), if based on the BIN code (see previous section), is $\mathbf{A}_0 = 1646688_7$, whereas if it is based on the CAMs, it is $_0\mathbf{A} = {}_7545$. Because the underlying adjacency matrix (Table 2) is identical, we may write

$$1646688_7 = {}_7545$$

Table 2. Upper Triangle of the Adjacency Matrix of the Structure Shown in Figure 3 and the Compressed Adjacency Matrix (Last, Separate Line)[a]

0	1				1	
	0	1	1			
		0		1		
			0			
				0		
					0	1
						0
-	1	2	2	3	1	6

[a] Zeros, except in the diagonal, were suppressed for the sake of clarity.

As expected, CAMs are more efficient than the technique based on the BIN codes.

Both types of codes are basically ordinal numbers, and the maximal codes correspond to the adjacency matrix of the complete graph. Therefore, for any N, the maximal value of $\mathbf{A}_0$ is $(N-1)! - 1$, the minus 1 factor appears because $_0\mathbf{A} = {}_N0$ is always the first code which represents an adjacency matrix (graph) in which $\mathbf{A}_{1,2} = \mathbf{A}_{1,3} = ... = \mathbf{A}_{1,N} = 1$ (i.e. a star).

In principle the proposed code could also be used to generate CAMs of all isomers of acyclic N-vertex trees. The generation process would start with the generation of numbers 0, 1, 2, ..., $(N-1)! - 1$. After transformation of these codes into CAMs, a special algorithm would be needed to determine whether the actual CAM represents an optimally coded structure[13] or not. Since $_0\mathbf{A}$ is related to physical trees, this procedure would not be efficient. On the other hand, once Morgan-trees (i.e. trees with CAMs containing a sequence of nondecreasing integers)—the number of which is just a small fraction of the number of physical trees[11,14]—could be encoded, the generation of isomers might be feasible. Then comparison with existing schemes, like the N-tuple code,[6] will be possible.

REFERENCES AND NOTES

(1) Harary, F. *Graph Theory*; Addison-Wesley, 1969; p 150.

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

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

(4) Knop, J. V.; Müller, W. R.; Szymanski, K.; Trinajstić, N. Enumeration of Planar Polyhex Hydrocarbons. *Reports Mol. Theor.* **1990**, *1*, 95-98.

(5) 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.

(6) 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.

(7) Randić, M. On Canonical Numbering of Atoms in a Molecule and Graph Isomorphism. *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 171-180.

(8) Hu, C. Y.; Xu, L. On Highly Discriminating Molecular Topological Index. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 82-90. C.Y.

(9) Hu, C. H.; Xu, L. Developing Molecular Identification Numbers by an All-Paths Methodol. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 311-315.

(10) 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.

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

(12) To compute $\mathbf{A}_0 = 108227168313541786_{12}$, representing the structure shown in Figure 2, two registers are needed in the calculator program to handle the digits spilling over.

(13) 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.

(14) Lukovits I. Isomer Generation: Semantic Rules for Detection of Isomorphism. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 361−366.

CI990152H