On Unique Numbering of Atoms and Unique Codes for Molecular Graphs

MILAN RANDIĆ

Department of Chemistry, Tufts University, Medford, Massachusetts 02155

Received October 14, 1974

A procedure is described which assigns unique labels to the graph vertices so that the resulting adjacency matrix is in a prescribed canonical form. The coding procedure is based on the adjacency matrix and its eigenvalue problem. The coefficients of eigenvectors associated with the largest eigenvalue provide the basis for sequencing atoms which are ordered according to the relative magnitudes of the coefficients. The scheme requires diagonalization of the adjacency matrix and a subsequent ordering of *n* numbers; hence, it is simple and practical. As an alternative to other procedures which are briefly discussed here, it provides the basis for the coding of molecular structures and recognition of graphs. It can be used in manipulations, such as storage and retrieval. The numbering appears to have some relationship to structural characteristics of molecular frameworks.

INTRODUCTION

The problem of associating a unique number (code) with a given graph G, which characterizes G up to isomorphism, has been recognized for some time. Interest in the problem has been mainly motivated by the need for developing a method for coding large and complicated chemical compounds.[1] Recent efforts to utilize computers in synthetic organic chemistry make the generation of a unique name for a graph more acute.[2] Here we suggest a device which accomplishes coding by assignment of unique labels to the vertices of G in such a way that the resulting adjacency matrix A takes a certain canonical form. Several other methods have been considered recently. Some of these originated in mathematical studies of graphs[3] and graph isomorphism,[4] while others developed in connection with the classification of chemical compounds[5] and with the development of a retrieval system to manipulate a large number of structures.[6]

In view of the plethora of schemes, it appears prudent to contemplate additional criteria in evaluating individual schemes. In order not to hinder the search for useful alternatives, it would appear to be best to impose upon the individual schemes as few rules as possible. Accordingly, topological schemes based on connectivity matrices would occupy a more favorable position.

DIFFICULTIES AND DEFICIENCIES OF THE CURRENT SCHEMES

We will briefly review the schemes which are concerned with the assignment of numbers to individual vertices. The two representative schemes are: (1) that based on the Morgan algorithm which exploits the concept of extended connectivity[7] and (2) the scheme based on the numbering of vertices associated with the smallest binary code.[4]

A key feature of the Morgan algorithm, used by the Chemical Abstracts Service in determining a set of rules that control the order in which the structure is described, is extended connectivity.[8] The classification of atoms by their connectivity does not permit sufficient differentiation between them. Attempts to make the connectivity concept more discriminating have been made by considering the connectivity of adjacent atoms as well. The extended connectivity values are obtained by adding the initial connectivity values of nearest neighbors and assigning the sum to the vertex considered. As is recognized[8] this method does not *always* allow the maximum possible differentiation, although it generally allows the atoms to be divided into many more than four classes depending on the number of nonhydrogen attachments to each atom. The subsequent order in which the structure is described is determined by sequencing the atoms according to a set of rules. The fact that some atoms have the same extended connectivity values does not make the Morgan algorithm ambiguous. However, these additional rules have limited structural relevance, and the resulting numbering need not reflect topological characteristics of the system.

The other scheme, which provides unique atomic numbering requires that the associated adjacency matrix represents a set of smallest binary codes.[4] However, the procedure for finding the unique numbering may not be perfect, and the suggested procedure consists of exchanging a pair of rows and columns which sometimes leads to a local minima. Thus a check with more complicated permutations, the simplest being a cyclic interchange of 3 rows, appears needed.[9]

Considering the importance of the issue, perhaps a less general scheme aimed primarily at establishing unique vertex numbering may be timely. Such a scheme could be used for testing graph isomorphism and even for developing storage and retrieval procedures, but in contrast to the scheme based on the smallest binary code would not be suitable for generation of all graphs of a prescribed size or cataloging all possible connected graphs.

TOPOLOGICAL EIGENVECTORS AND NUMBERING OF VERTICES

We describe here a simple and practical scheme for deriving sequential numbering of vertices (atoms) in a graph. This scheme is remarkably free of additional conventions except for an agreement for labeling equivalent vertices. The procedure is based on the adjacency matrix and its eigenvalue problem. It is known[10] that the spectrum of the matrix associated with a graph does not uniquely determine a graph, but the eigenvectors provide additional information which can be used for supplementing the test for isomorphism. Even if graphs are isospectral, their eigenvectors will be different unless they are isomorphic, *i.e.*, having identical connectivity.

We consider a simple way of combining the information in eigenvectors with a graph and its adjacency matrix. We shall confine our attention to the eigenvector associated

Table I. Numbering of Atoms for Heptane Isomers

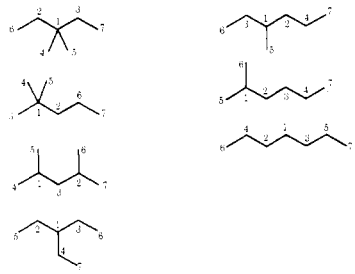


Table II. Numbering of Atoms for Selected Pairs of Isospectral Graphs

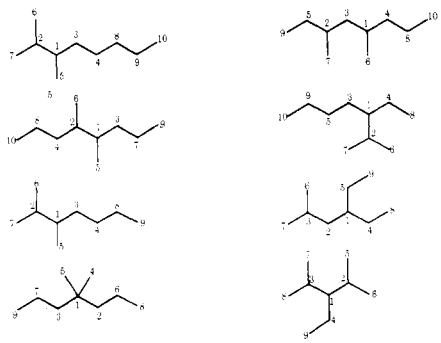
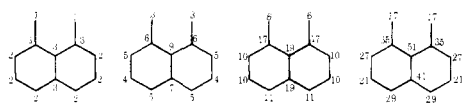
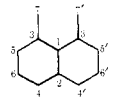


Table III. Connectivity Values for 1,8-Naphthoquinodimethane for Successive Iterative Steps and the Resulting Numbering of Nonequivalent Vertices and Coefficients of the First Eigenvector Leading to the Same Numbering

Extended connectivity:



Associated numbering:



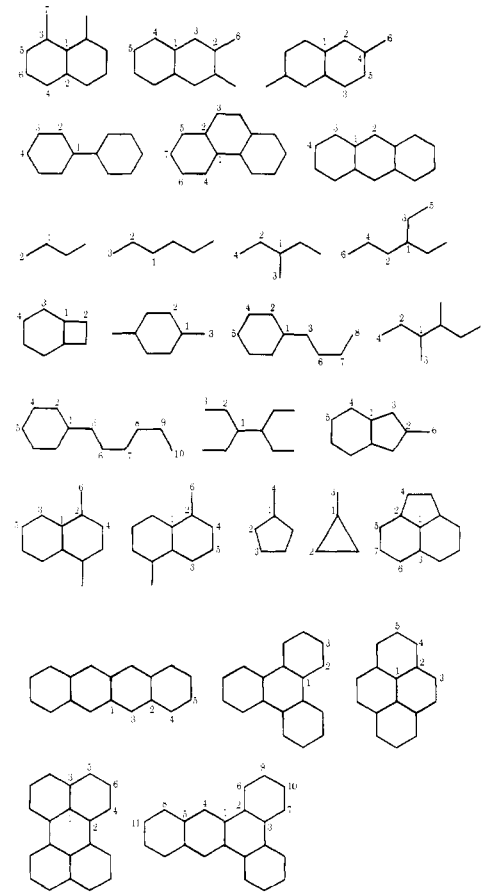
The first eigenvector: (x_1 = 2.3941)

1	2	3	4	5	6	7
0.4658	0.4082	0.3535	0.2557	0.2329	0.2041	0.1476

with the largest eigenvalue and shall refer to such an eigenvector as the first eigenvector.[11] The eigenvectors are determined by a set of coefficients which define the relative participations of basis functions in each eigenfunction. All coefficients of the first eigenvector are taken to be of positive sign. Generally nonequivalent vertices will have coefficients of different magnitudes. Hence the magnitudes of the coefficients of the first eigenvalue can provide the basis for ordering nonequivalent vertices in a sequence. Equivalent vertices can be ordered adopting one of the conventions already used in other schemes; the problem is not unique to the particular procedure. Once we have arrived at a unique vertex numbering, the recognition of identical graphs is trivial; one merely examines adjacency matrices associated with the unique numbering.

In order to approach as close as possible the numbering of atoms given by the Morgan algorithm, one should associate with the largest coefficient the number 1 and with the smallest coefficient the number n, where n is the number of vertices in the structure. Examination of the first eigenvectors reveals some regularities in the relative magnitudes of the coefficients and the valencies of the vertices in that

Table IV. Selected Skeletons of Hydrocarbons and the Numberings Based on the Coefficients of the First Eigenvector[a]



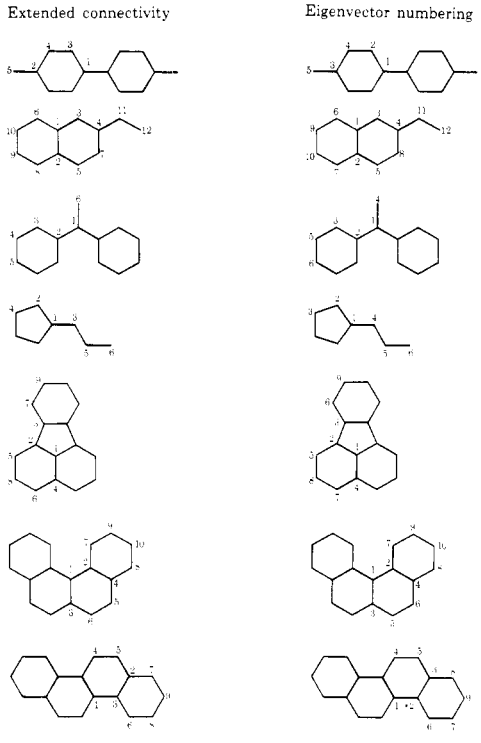
[a] All eigenvectors taken from C. A. Coulson, A. Streitwieser, M. D. Poole, and J. I. Brauman, "Dictionary of pi-Electron Calculations," W. H. Freeman and Co., San Francisco, Calif.

the largest coefficients generally belong to vertices of the highest valency, while the smallest coefficients generally belong to terminal vertices and their neighbors. In Tables I–VI we list the unique numbering for selected graphs. The procedure for establishing the numbering is simple since apart from matrix diagonalization it requires only the ordering of a set of n coefficients. To further test the scheme, we screened some 50 pairs of isospectral graphs. A few representative numberings are listed in Table II. These are cases in which the approach based on examination of the spectrum fails, but the unique numbering easily differentiates between isospectral components.

In cases when several vertices are equivalent, we adopted the convention that the ordering should be such that neighboring numbers are as similar as possible. If initial numbers belong to equivalent vertices, the selection of the first and second vertex is arbitrary, but this in no way affects the form of the adjacency matrix.

One question remains to be considered: Does the proposed scheme fail in special circumstances? The answer would be yes if and only if situations arise such that nonequivalent (from adjacency point of view) vertices have the same coefficient in all of their eigenvectors. One cannot deny that such an occasion might arise, but such a situation must signify a special condition. Accidental degeneracies can be traced to constraints or conditions of some kind as illustrated by the eigenvalues of the hydrogen atom based on the Coulomb potential, or the excessive degeneracies of the simple Hückel method traced to a higher symmetry of associated molecular graphs.[12] Similarly special require-

Table V. Molecular Structures in Which the Schemes Based on Extended Connectivity and on the First Eigenvectors Produce Somewhat Different Labelings



ments allow the appearance of the spectrum of a fragment in the overall spectrum of the system.[13] So one should look for the cause of such an unusual event; its recognition may give a clue to the prescription of the sequential order in such instances.

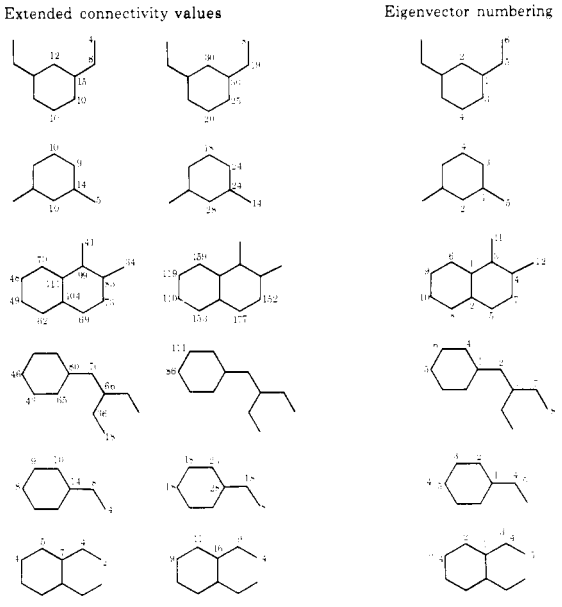
We have found such an unusual graph in the styrene framework. Here the terminal vertex and vertices in the meta to vinyl positions, though nonequivalent, have the same respective coefficients. As we know,[14] this particular graph is unique since it allows the exchange of residuals at specific sites without actually affecting the topological spectra; hence it presents the basis for generation of families of isospectral graphs. Another example is the acyclic graph studied by Schwenk,[15] also a unique generator of isospectral graphs. However, even such exceptional situations do not require additional rules, and a procedure for ordering equivalent vertices will determine the numbering of atoms uniquely.

In conclusion we should point out that the price which must be paid for the simplicity of this procedure is that it lacks some of the properties of other schemes. For example, the resulting numbering does not make the skeleton obvious, nor does it provide a complete basis for generation of acceptable graphs of a given size and specified connectivity. While the former may be revealing and the latter essential for methodical studies, neither is important for coding molecular structures, or for their recognition and manipulation, so the proposed scheme may prove economical for these general purposes.

ON THE RELATIONSHIP BETWEEN NUMBERING AND STRUCTURE

Since numbering is based on the coefficients of the first topological eigenfunction, one may expect it to reflect some structural features as well. By searching for such a relationship we are, in fact, searching for the relationship between the eigenvector of the first eigenvalue and the structural characteristics of the system. Examination of the resulting numberings (Table I–VI) reveals that small numbers are

Table VI. Molecular Skeletons in Which the Extended Connectivity Scheme Results in Oscillatory Behavior, Instability of the Numbering, and Indeterminancy



associated with vertices of the highest valencies, while terminal (monovalent) vertices have the largest numbers. There are, however, exceptions even to this general observation, such as those in *p,p*-biphenoquinodimethane and 2-vinylnaphthalene (Table V). To eliminate arbitrariness associated with conventional labeling of equivalent vertices we give in Tables III–VI only the numbering for nonequivalent vertices and then distinguish between them with a prime superscript. Even for construction of the adjacency matrix we do not need to distinguish between equivalent positions. One can sequence equivalent atoms in the same group since their relative positions do not affect the form of the matrix.

We adopted the rule that the highest coefficient should have associated with it the number 1 in order to parallel the algorithm due to Morgan. This is purely a matter of convenience, and we did not expect the two schemes to be related in any way. It was therefore surprising to find that a comparison for 1,8-naphthoquinodimethane, an example used by Morgan to illustrate the extended connectivity scheme,[7] gives complete agreement in the numbering of atoms (Table III). Could this be a coincidence or is there some underlying substance to the finding? The question is too important to be overlooked and deserves further attention. To find an answer we proceeded to examine some 35 skeletons of conjugated cyclic and acyclic hydrocarbons. In many cases we found the two schemes, one based on extended connectivity and the other on eigenvectors, to give the same numbering (Table IV). The sample includes benzenoid and nonbenzenoid systems, quinoid structures, polyphenyls, polyacenes, polyenes, vinyl compounds, and several nonalternant hydrocarbon skeletons. Some of the molecules considered have more than ten nonequivalent atoms (*e.g.*, 1,2,3,4-dibenzanthracene), and the full agreement in the numbering strongly suggests a causal relationship. Sometimes, however, the two schemes give different numbering (Table V). It is remarkable that only a few vertices are differently labeled. If one bears in mind that the differences between neighboring vertices in the sequencing procedure is frequently based on differences of a few per cent in the magnitudes of the associated extended connectivity values and that the coefficients may also differ by a few per cent or less, then the partial agreement should be suggestive of some intimate relationship between the coefficients of the first eigenvector and extended connectivity values.

The concept of extended connectivity, though simple, occasionally leads to practical difficulties. First there is sometimes nonuniform convergence (for instance, in 1,8-naphthoquinodimethane; Table III), but more troublesome is the occurrence of oscillatory behavior which ascribes to a nonequivalent set of nuclei the same extended connectivity values. Examples are to be found for *m*-xylylene and *p*-divinylbenzene (Table VI). Also, not infrequently the calculated extended connectivity values show instability. That is to say, that although all nonequivalent vertices have different connectivity values, the relative values change with subsequent iteration. Examples are 8,8-divinylstyrene and 1,2-naphthoquinodimethane (Table VI). Finally, in rare situations we have the difficulty that nonequivalent vertices have the same connectivity values in every iterative step. This is the case for styrene and *o*-divinylbenzene (Table VI) which are special cases for the scheme based on the coefficients of the first eigenvector as well. We know that such a situation reflects the ability of such frames to support a family of isospectral graphs.

It seems clear that the schemes based on eigenvectors and on extended connectivity are related although the relationship is not transparent. The deficiencies of the extended connectivity procedure are conspicuously absent in the procedure which uses eigenvectors, which makes the latter method advantageous.

ACKNOWLEDGMENTS

I wish to thank Professor W. Todd Wipke (Princeton University) for correspondence and Dr. J. W. Sutherland (Yale University) for suggesting improvements in the presentation.

LITERATURE CITED

(1) We give a brief selection of representative literature: A. T. Balaban and F. Harary, *Tetrahedron*, **24,** 2505 (1968); L. Spialter, *J. Chem. Doc.*, **4,** 261 (1964); R. H. Penny, *ibid.*, **5,** 113 (1965); H. Hiz, *ibid.*, **4,** 173 (1964). A number of systems attempts to generate standard codes: W. J. Wiswesser, "A Line-Formula Chemical Notation," Crowell, New York, N. Y., 1954; L. Lederberg, G. L. Sutherland, B. G. Buchanan, E. A. Feigenbaum, A. V. Robertson, A. M. Duffield, and C. Djerassi, *J. Amer. Chem. Soc.*, **91,** 2973 (1969); J. E. Dubois, D. Laurent, and H. Veillard, *C. R. Acad. Sci.*, **263,** 764, 1245 (1966).

(2) Ugi, I., P. Gillespie, and C. Gillespie, *Trans. N.Y. Acad. Sci.*, 416 (1974); J. Blair, J. Gasteiger, C. Gillespie, P. D. Gillespie, and I. Ugi, *Tetrahedron*, **30,** 1845 (1974); E. J. Corey, W. T. Wipke, R. D. Cramer, and W. J. Howe, *J. Amer. Chem. Soc.*, **94,** 421, 431 (1972); W. T. Wipke and T. M. Dyott, *ibid.*, **96,** 4825, 4834 (1974); C. K. Johnson and C. J. Collins, *ibid.*, **96,** 2514 (1974); C. J. Collins, C. K. Johnson, and V. F. Raaen, *ibid.*, **96,** 2524 (1974); J. B. Hendrickson, *ibid.*, **93,** 6847, 6854 (1971); H. Gelernter, N. S. Sridharan, A. J. Hart, S. C. Yen, F. W. Fowler, and H. J. Shue, *Fortsch. Chem. Forsch.*, **41,** 113 (1973).

(3) Harary, F., *Colloq. Math. Soc. J. Bolyai, Balatonfured, Hungary, 1969*, p 625, and therein cited work of H. J. W. Duijvesteijn and B. R. Heap; R. C. Read in "Graph Theory and Computing," Academic Press, New York, N. Y., 1972, p 153.

(4) Randić, M., *J. Chem. Phys.*, **60,** 3920 (1974). In Figures 4 and 5 of this paper several incorrect labels are shown. The following exchanges should be made: adamantane 5/6; cubane 3/4; nortricyclene 2/3; snoutene 2/3, 6/7, 9/10; benzphenanthrene 2/3, 8/9, 15/16, 17/18; chrysene 2/3, 8/9, 17/18. I am grateful to Dr. A. Mackay (University of London) for corrections.

(5) Lefkovitz, D., quoted in ref 3.

(6) Stockton, F. G., Report on Standardization of Graphs, National Technical Information Service, Springfield, Va., 22151, No. CFSTI: PB 179837.

(7) Morgan, H. L., *J. Chem. Doc.*, **5,** 107 (1965).

(8) Wipke, W. T., and T. M. Dyott in ref 2.

(9) Mackay, A., *J. Chem. Phys.*, in press; however, see also M. Randić, *ibid.*, in press.

(10) Collatz, L., and U. Sinogowitz, *Hamburg Univ., Abh. Math. Sem.*, **21,** 63 (1957).

(11) Terminology used by mathematicians for the first eigenvalue is *graph index*.

(12) Wild, U., J. Keller, and H. H. Günthard, *Theor. Chim. Acta*, **14,** 383 (1969).

(13) Živković, T., N. Trinajstić, and M. Randić, *Int. J. Quantum Chem.*, submitted for publication.

(14) Živković, T., N. Trinajstić, and M. Randić, *Mol. Phys.*, in press; W. C. Herndon, *Tetrahedron Lett.*, 671 (1974); for additional work on isospectral molecular graphs and their constructions, see W. C. Herndon and M. L. Ellzey, Jr., *Tetrahedron*, in press; W. C. Herndon and M. L. Ellzey, Jr., submitted to *Discrete Math.*

(15) Schwenk, A., in "New Directions in the Theory of Graphs," F. Harary, Ed., Academic Press, New York, N. Y., 1973, p 275.