Molecular Identification Number for Substructure Searches

FRANK R. BURDEN

Chemistry Department, Monash University, Clayton, Victoria 3168, Australia

Received January 13, 1989

A method for producing molecular identification numbers for hydrogen-depleted organic structures from the eigenvalues of a connectivity matrix is presented. Over 20 000 structures have been successfully tested, and the method can also be used to produce a unique numbering for the atoms in a structure and to identify which atoms belong to each of the substructures of a disconnected main structure.

INTRODUCTION

Codes that can classify, identify, or index a large number of molecular structures have been the subject of investigation for decades. There are several types of use to which such codes can be put, each with different criteria as to their properties such as reversibility (molecular formula to code and code to molecular formula), uniqueness, compactness, and ease of production. Two important uses are for indexing large databases, such as the Chemical Abstracts Online service and the lists of reaction transforms in programs that attempt to suggest synthetic pathways. Of course the molecular formula itself is a code that is used in everyday chemical discourse but is itself not readily amenable to the architecture of present-day computers, though perhaps the development of neural networks will alter this. A further use for some indices has been in the area of quantitative structure–activity relationships, though an index for use in substructure searching need not necessarily have a direct functionality in this manner.

Three main approaches have been taken. The molecular formula is made more compact by removing redundant information and compressing the rest. The Wiswesser line notation[1,2] is a well-known and widely used example, though having the disadvantage of being difficult for the inexperienced to encode. A second method is to make use of graph invariant properties such as path counts, which has the main advantage of being nonempirical and easily programmable for computation, though the number of digits in the index for large structures is somewhat wasteful. A third method uses the connectivity matrix as a part of an index-producing algorithm, with or without empirical parameters, of which the characteristic polynomial is an example, albeit in another form. Reviews of these various approaches can be found in the literature.[3,4]

An aim of this work is to produce an index that is highly compact but with minimal redundancy in terms of the number of characters (decimal digits) needed. These two aims are in conflict, so a compromise must be reached, and although it would also be more aesthetic to do without empirical parameters, this is a luxury worth sacrificing for a usable code, providing the parameters are reasonably intuitive. The molecular or empirical chemical formula is itself immediately available for any structure and can be used as a key to reduce the requirements on the computed index since it then need only be nondegenerate for identical chemical formulas. The most likely case for degeneracy to occur is with ring structures, and here a ring count can be used to reduce the chance.

Given the above, a method is outlined below that meets these requirements, though whether the indices will prove useful in areas apart from structure searching remains to be investigated.

OUTLINE OF THE METHOD

Connectivity Matrix. The essence of the method is to solve the eigenvalue equation

$$[\mathbf{B}][\mathbf{V}] = [\mathbf{V}][\mathbf{e}] \quad \text{or} \quad [\mathbf{V}]^t[\mathbf{B}][\mathbf{V}] = [\mathbf{e}]$$

where [**B**] is a real symmetric connectivity matrix to be defined, [**V**] is a matrix of eigenvectors, and [**e**] is a diagonal matrix of eigenvalues. The identification number will consist of the *n* smallest eigenvalues taken to *m* significant figures, where *n* and *m* are chosen according to the size of the database to be indexed.

The rules defining [**B**] used here are as follows:

(a) Hydrogen atoms are not included.

(b) The heavy atom structure is arbitrarily numbered.

(c) The diagonal elements of [**B**], B_{ii}, are the atomic numbers of the atoms.

(d) The element of [**B**] connecting atoms *i* and *j*, B_{ij}, is 0.1 for a single bond, 0.2 for a double bond, 0.3 for a triple bond, and 0.15 for an aromatic delocalized bond.

(e) Elements of [**B**] corresponding to bonds to terminal atoms (i.e., atoms with one connection only) are augmented by 0.01.

(f) All other elements of [**B**] are set at 0.001.

The original rationale for this form of the matrix was derived from consideration of molecular orbital calculations. The lowest eigenvalues will belong to the totally symmetric representation of the molecular point group, and the eigenvectors will have coefficients belonging to every atom, except in highly symmetric cases. This means that the lowest eigenvalues contain contributions from all atoms and so reflect the topology of the whole molecule. Identical eigenvalues for different molecules are therefore expected to occur rarely.

The actual values for the matrix elements of [**B**] were chosen to be powers of 10 apart. Diagonal elements are equal to 1 or greater, bonds 0.1 or greater, terminal bonds 0.01 larger, and finally nonbonded elements of the order 0.001. These ratios enable each type of element to contribute without being swamped by another type, so preserving the detail. Of course, the actual values assigned are arbitrary, though it is most important to set the nonbonded elements to a nonzero value to reduce the chance of degeneracy.

Eigenvectors. The eigenvectors are not stored as they are of no use as identification numbers. They can, however, be used in one circumstance. Given a molecule such as

$$S_1–S_2–S_3$$

where S_1, S_2, and S_3 are any arbitrary substructures, if the connecting bonds are broken so that S_1, S_2, and S_3 are independent, the eigenvectors may be used to distinguish which atoms belong to which substructure as follows.

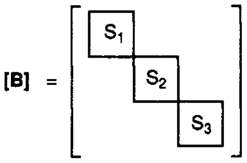
The [**B**] matrix is set up as defined above except that the elements corresponding to the nonbonded connections are set

0095-2338/89/1629-0225$01.50/0

Table I. Number of Indexing Digits Needed for the Alkanes

max no. of carbon atoms	total no. of alkanes	no. of eigenvalues used	total no. of indexing digits
10	149	1	6
12	663	1	7
14	3323	2	12
16	18029	2	16

to zero as are the S_1–S_2 and S_2–S_3 connecting elements. In principle, the [**B**] matrix could now be rearranged so that it would block diagonal in S_1, S_2, and S_3



with zero elements outside the S blocks.

The eigenvectors of this variation of [**B**], corresponding to the smallest eigenvalues, will have both zero and nonzero components that can be used to ascribe each atom to one of the substructure fragments or another. The presence of nonzero coefficients in any particular vector indicates that the corresponding atoms all belong to the same fragment. Examination of several vectors can then be used to sort out one fragment from another.

A further possible bonus from using this method is that structures can be uniquely numbered. Following the diagonalizing transform

$$[\mathrm{V}]^t[\mathbf{B}][\mathrm{V}] = [\mathrm{e}]$$

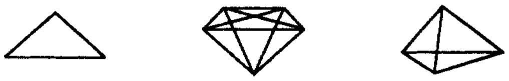
the diagonal elements of [e] are ordered numerically $[\mathrm{e}] \Rightarrow [\mathrm{e}]_0$ and for every transposition of the diagonal elements $\mathrm{e}_i \Rightarrow \mathrm{e}_j$ the ith and jth columns of [V] are transposed, $[\mathrm{V}] \Rightarrow [\mathrm{V}]_0$, as are the ith and jth rows of the transposed matrix $[\mathrm{V}]^t$, $[\mathrm{V}] \Rightarrow [\mathrm{V}]_0^t$, from which by the reverse transformation

$$[\mathbf{B}]_0 = [\mathrm{V}]_0[\mathrm{e}]_0[\mathrm{V}]_0^t$$

$[\mathbf{B}]_0$ now reflects the ordering imposed on [e] from which the atoms can be labeled.

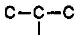
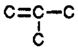
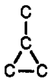
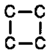
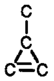
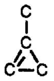
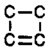
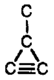
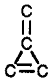
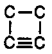
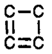
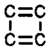
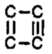
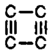
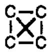
Implementation. The method was tested with Pascal programs and matrix routines from *Numerical Recipes*.[5] It was found that the 18 029[6] possible alkanes with up to 16 carbon atoms had different indices by using the two lowest eigenvalues taken to eight decimal places. Since the eigenvalues of all the alkanes range between 5 and 6, the compounds can be indexed with the 16 digits after the decimal point. Table I shows indexing for various numbers of alkanes.

In the case of the alkenes it was found that accidental degeneracies had an earlier onset, though of the 1560 structures with carbon numbers up to 11 only 2 structures had the same indices when 16 digits were used in the index. Again, for the 1696 alcohols plus ethers with empirical formulas $C_nH_{2n+2}O$ all indices were distinct. These were checked against the numbers quoted by Lederberg et al.,[7] though they give no absolute guarantee that their values are accurate. Because ring structures may cause more degeneracies than allylic structures, the indices for the ring structures shown by Randic were evaluated. Different indices were obtained for all different structures (noting that there are some repeats) by using two eigenvalues taken to six decimal places, except the structures



which is not a serious drawback, especially since they differ in their carbon count. Even here the first two structures are differentiated by the third eigenvalue.

Table II. Indices for Some C_4 Fragments

formula	fragment structure	lowest eigenvalue	second eigenvalue
C_4H_{10}	C—C—C—C	5.829 786 80	5.928 962 97
C_4H_{10}		5.810 471 79	5.999 000 00
C_4H_8	C═C—C—C	5.762 128 24	5.902 907 01
C_4H_8	C—C═C—C	5.751 915 28	5.951 432 06
C_4H_8		5.739 507 99	5.999 000 00
C_4H_8		5.844 782 90	5.900 000 00
C_4H_8		5.801 000 00	5.999 000 00
C_4H_6	C═C—C═C	5.734 718 13	5.833 771 48
C_4H_6	C≡C—C—C	5.672 652 32	5.895 837 23
C_4H_6	C—C≡C—C	5.664 483 11	5.964 297 82
C_4H_6		5.800 000 00	5.857 550 77
C_4H_6		5.771 768 91	5.916 528 91
C_4H_6		5.739 090 13	5.960 909 87
C_4H_4	C═C═C═C	5.668 023 76	5.867 216 37
C_4H_4	C≡C—C═C	5.664 949 43	5.805 700 27
C_4H_4		5.700 000 00	5.865 078 30
C_4H_4		5.766 145 82	5.800 000 00
C_4H_4		5.659 283 97	5.940 716 03
C_4H_4		5.684 671 65	5.999 000 00
C_4H_4		5.701 000 00	5.899 000 00
C_4H_4			
C_4H_4		5.639 090 13	5.860 909 87
C_4H_4		5.601 000 00	5.799 000 00
C_4H_4	 tetrahedron	5.900 000 00	5.900 000 00

Although the actual values produced for the index are arbitrary (to the extent that the rules are themselves arbitrary), it might be interesting to note some actual values for a number of C_4 fragments as presented in Table II. The term fragment is used here to represent a heavy atom skeleton (i.e., hydrogen depleted) so that a fragment such as C═C—C—C can also represent the molecule 1-butene as well as the chain of four carbon atoms, as shown, within a larger fragment or molecule.

CONCLUSION

A method has been presented that allows a molecular identification number for a substructure to be calculated in a straightforward manner by using commonly available matrix manipulation routines. Over 18 000 alkanes were indexable, using 14 digits from two eigenvalues, as were 1700 alkenes and 200 different ring systems. These indices are suitable for indexing a database of structures such as those used for encoding reaction transforms and can be made more or less discriminating by using an appropriate number of eigenvalues. However, the method is not reversible in that the index cannot be used to derive the original structure. On the other hand, it can be used to derive a unique numbering for substructures by ordering the eigenvalues of the defined connectivity matrix.

Another useful benefit is associated with the eigenvectors, which can be used to determine the attribution of each atom to substructures upon disconnection of the main structure into distinct fragments. This could be useful when the possible disconnections of a main structure into reacting substructures are considered.

REFERENCES AND NOTES

(1) Wiswesser, W. J. *J. Chem. Inf. Comput. Sci.* **1982**, *22*, 88.

(2) Wiswesser, W. J. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 258.

(3) Randič, M. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 164.

(4) Ash, J. E.; Chubb, P. A.; Ward, S. E.; Welford, S. M.; Willett, P. *Communication, Storage and Retrieval of Chemical Information*; Ellis Horwood Ltd.: Chichester, U.K., 1985.

(5) Press, W. H.; Flannery, B. P.; Teukolsky, S. A.; Vetterling, W. T. *Numerical Recipes: The Art of Scientific Computing*; Cambridge University Press: Cambridge, U.K., 1986.

(6) Henze, H. R.; Blair, C. M. *J. Am. Chem. Soc.* **1931**, *53*, 3077.

(7) Lederberg, J.; Sutherland, G. L.; Buchanan, B. G.; Feigenbaum, E. A.; Robertson, A. V.; Duffield, A. M.; Djerassi, C. *J. Am. Chem. Soc.* **1969**, *91*, 2973.

Topological Organic Chemistry. 1. Graph Theory and Topological Indices of Alkanes

HARRY P. SCHULTZ

Department of Chemistry, University of Miami, Coral Gables, Florida 33124

Received January 24, 1989

By use of the adjacency, degree, and distance matrices that describe the structure of an alkane, a method is outlined for deriving with simple matrix algebra a molecular topological index number of the alkane. Additionally, the calculations offer a picture of the relative intricacy (branched) value for each carbon atom of the alkane.

INTRODUCTION

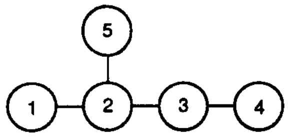
Graph theory offers the means to numerically characterize structures of chemicals. Single-number representations of alkanes have been reported by Hosoya[1] and Randič.[2] Balaban,[3] Randič,[4] and Hansen and Jurs[5] have summarized methods of calculating molecular topological indices, observing that problems, mainly of degeneracy, exist in the methods they reviewed.

This paper outlines a technique to determine molecular topological indices (MTI) by subjecting the adjacency, valence, and distance matrices describing the structures of alkanes to matrix algebraic operations. The procedure results in a single-number solution of high discrimination for each molecular graph. The method is simple to use. It is entirely objective, and the results are essentially monotonic. The representative list in Table I includes many of those pairs of compounds (**17** and **18**, **19** and **20**, **24** and **25**, **27** and **28**, **29** and **31**, **31** and **32**, **21** and **33**, **41** and **42**) for which earlier work presented ambiguous indices.

COMPUTATIONS

The method described is illustrated with a specific example, 2-methylbutane (**7**), hydrogen-suppressed, and with the interatomic carbon–carbon bond distances set at unity. The procedure uses the simplest of matrix algebraic operations.[6]

(**1**) The structural (molecular) graph for 2-methylbutane is drawn.



(**2**) Its distance (**D**) matrix is constructed.

$$\mathbf{D} = \begin{bmatrix} 0 & 1 & 2 & 3 & 2 \\ 1 & 0 & 1 & 2 & 1 \\ 2 & 1 & 0 & 1 & 2 \\ 3 & 2 & 1 & 0 & 3 \\ 2 & 1 & 2 & 3 & 0 \end{bmatrix}$$

(**3**) Its adjacency (**A**) matrix is constructed.

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

(**4**) Its valency (**v**) (degree) matrix is constructed.

$$\mathbf{v} = [1 \quad 3 \quad 2 \quad 1 \quad 1]$$

Müller et al.[7] demonstrated that the **D** matrix can be derived from the **A** matrix. Summing the elements in either the rows or columns of the **A** matrix yields the elements of the **v** vector; inspection of the molecular graph is the fastest way to construct the **v** vector.

(**5**) The **D** and **A** matrices of 2-methylbutane are summed to give the (**D** + **A**) matrix.

$$\overset{\mathbf{D}}{\begin{bmatrix} 0 & 1 & 2 & 3 & 2 \\ 1 & 0 & 1 & 2 & 1 \\ 2 & 1 & 0 & 1 & 2 \\ 3 & 2 & 1 & 0 & 3 \\ 2 & 1 & 2 & 3 & 0 \end{bmatrix}} + \overset{\mathbf{A}}{\begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix}} = \overset{(\mathbf{D} + \mathbf{A})}{\begin{bmatrix} 0 & 2 & 2 & 3 & 2 \\ 2 & 0 & 2 & 2 & 2 \\ 2 & 2 & 0 & 2 & 2 \\ 3 & 2 & 2 & 0 & 3 \\ 2 & 2 & 2 & 3 & 0 \end{bmatrix}}$$

(**6**) Its (**D** + **A**) matrix is multiplied by the row vector **v** used as the premultiplier, thus affording conformability to the expression