Topological Index as a Sorting Device for Coding Chemical Structures*

HARUO HOSOYA
Department of Chemistry, Ochanomizu University,
Bunkyo-Ku, Tokyo 112, Japan

Received June 1, 1972

Although the topological index, an easily calculable quantity, does not uniquely correspond to the individual structure of a graph, it roughly represents the topological nature of the graph—i.e., branching and cyclization. Examples are given for using the topological index as a first sorting device for coding and retrieving structures, especially of fused polycyclic systems.

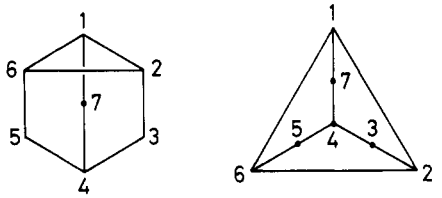
In spite of the continuous effort by many researchers, the proposed methods for coding the skeletons of fused ring systems are rather difficult to master and sometimes betray arbitrariness in the coding system.[2,3,9-14] For example, to get the IUPAC name for a polycyclic compound as in Figure 1, one must write out as many pairs of bridgehead atoms as possible, each of which gives a candidate for the correct name. The next step is to select the name with the highest priority, say (A), from the group of the names (A) to (F). As the number of rings increases, the procedure and the names become lengthy and are liable to be erroneous. Less formal names, for example, perhydro-1,4-ethanoindene for the skeleton (D), in general use sometimes make things more complicated. The problem is that one can never get access to the items in a file under the name (A) so long as he looks for the compounds with names (B), (C), and so on.

This problem might be resolved if one puts the compound into matrix form such as an adjacency matrix A, which takes account of the neighborship and, if modified, can discriminate the atom species and the bond multiplicities.[14] However, this is a rather time-consuming method even with a computer, since a matrix with N degree has $N!$ different ways of representations. Information on the topological nature of the skeleton of a molecule seems to be condensed in the characteristic polynomial $P(X) = \det|A + XE|$ or the set of its coefficients.[4] (Another definition can be used as $P(X) = (-1)^N \det|A - XE|$.) Very recently, however, it was shown that the characteristic polynomial does not uniquely determine the topology of a molecule.[1,7,8] Namely, some distinct graphs have the same polynomial.

I have defined the topological index Z to characterize the topological nature of a molecule.[4] The topological index for graph G is defined as the sum of the nonadjacent number, $p(G,k)$, which is the number of ways of choosing k disconnected bonds from G. Although the quantity Z, as well as $P(X)$, does not uniquely characterize the topology, it does roughly represent the topological nature of a graph. Further, it can be obtained very quickly even by hand by the aid of two composition principles, the prescription being illustrated in the Appendix.[4,5] Tables of Z values for smaller graphs have been published.[7,8] Application of Z to other chemical problems has been proposed.[5,6]

*Supported in part by a fund from the Ministry of Education of Japan, Tokutei-Kenkyu I, Showa 47, No. 92427.

We now propose to use the topological index as a first sorting device for coding or retrieving the structure of the compounds either with or without rings. Let us see how it works. The carbon atom skeleton of tricyclo[2.2.1.$0^{2,6}$]-heptane has C_{2v} symmetry in the projection 1.



If three methyl branches are attached to skeleton 1, we get thirty-one isomers as in Figure 2. They are classified into thirteen by the values of Z. On the other hand, structure 1 can be projected into 2 so as to have C_{3v} symmetry, which makes some group of isomers derived from projection 1 identical—e.g., 1,2,4- and 2,4,6-trimethyl derivatives are the same. Thus it turns out to be that there are only 14 isomers which exactly correspond to the classification by the topological index with only one exception: two different isomers, cyclofenchene (1,3,3-derivative) and isocyclene (3,3,4-derivative), have $Z = 115$.

Figure 1 shows the next example, where the topological indices of all six pseudo-isomers are found to be 290. On the other hand, as in Figure 3, all of the structures (G) to (L) except (J), with indene skeletons, are found to have Z values different from the group (A) to (F). Thus all of

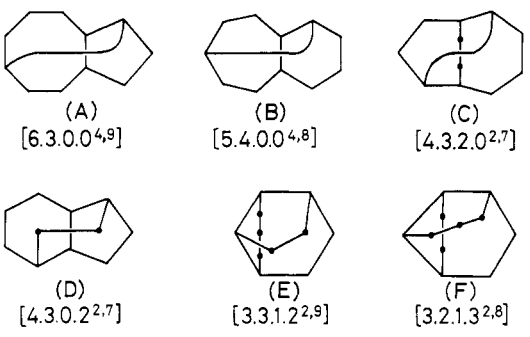


Figure 1. Various projections and names for tricyclo[6.3.0.$0^{4,9}$]undecane

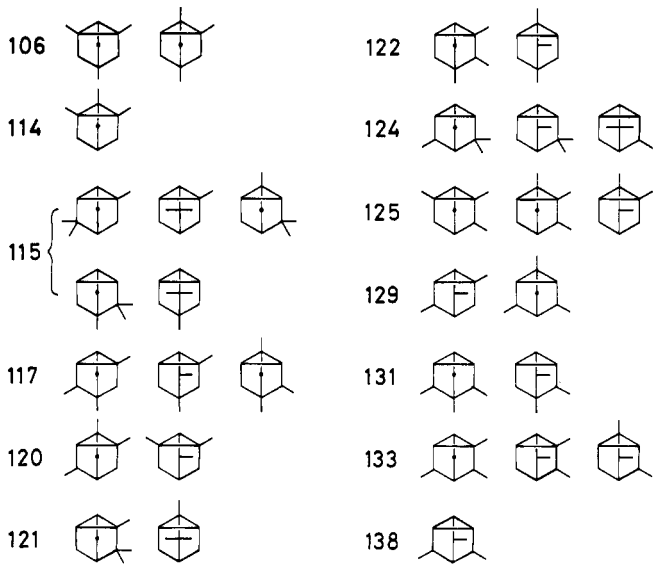


Figure 2. Pseudo-isomers of trimethyltricyclo[2.2.1.$0^{2,6}$]heptanes sorted according to Z values

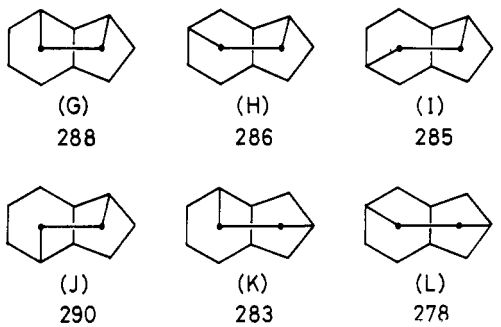


Figure 3. Perhydroethanoindene isomers and their Z values

the structures (G) to (L) are shown to be different from each other.

That the topological indices of two given graphs are equal is not sufficient for identity of structures. But if the Z values of two given graphs are different, one need not try further comparison, while if they are equal one must resort to finer comparison or identification. Study is in progress on this problem.

APPENDIX

Calculation of the Topological Index[4,5]

For graph G, Z is defined as

$$Z = \sum_{k=0}^{m} p(G,k). \qquad (1)$$

Tables I and II give examples of $p(G,k)$ and Z for the two simplest series of graphs, linear chains and cycles. Note that the Z values for both the series form the well-known Fibonacci series. By counting the $p(G,k)$ numbers for graph 1 or 2, we get $Z = 1 + 9 + 21 + 10 = 41$.

This value can be obtained more efficiently by use of the composition principle I (CPI). Another composition principle (CPII) is explained elsewhere.[5] Delete bond l from G and we get subgraphs L and M (Figure 4). (If l is a member of a cycle deletion of l gives only one subgraph.

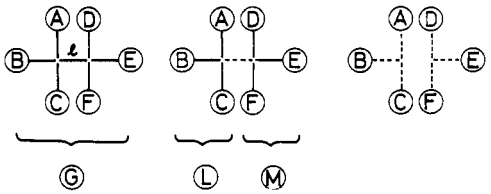


Figure 4. Illustration for the compositpon principle I (CPI)

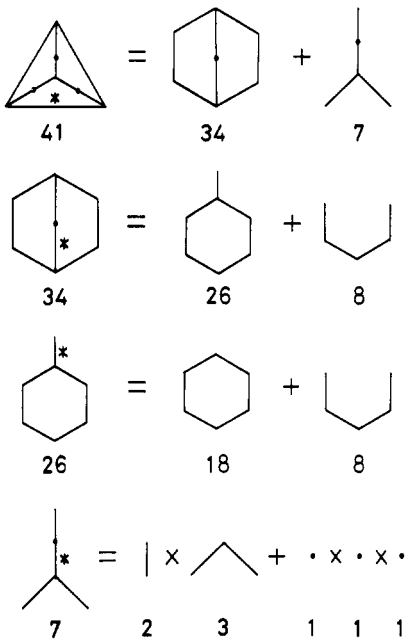


Figure 5. Schematic diagram for enumerating the Z value of graph 1 or 2, the asterisks showing the bonds to be deleted for applying CPI

Table I

G	$p(G,k)$ $k=0$	1	2	3	Z
•	1				1
—	1	1			2
— —	1	2			3
— — —	1	3	1		5
— — — —	1	4	3		8
— — — — —	1	5	6	1	13
— — — — — —	1	6	10	4	21

Table II

G	$p(G,k)$ $k=0$	1	2	3	Z
•	1				1
⬭	1	2			3
△	1	3			4
□	1	4	2		7
⬠	1	5	5		11
⬡	1	6	9	2	18

See the example.) Next delete all the bonds that were incident to l in the original graph G, and we get subgraphs A, B, ··· F. The value of Z for G is given by

$$Z_G = Z_l Z_M + Z_A Z_B Z_C Z_D Z_E Z_F \quad (2)$$

and is shown to be independent of the choice of bond l. For graph 1 or 2 let us choose bond 2—6 (asterisked) for deletion as in Figure 5, where the procedure for obtaining Z is illustrated. The Z values for smaller graphs are taken from Tables I and II.

Relations of $p(G,k)$ and Z with the characteristic polynomial $P(X)$ are discussed elsewhere.[4,5]

ACKNOWLEDGMENT

The author is indebted to Kenzo Hirayama of Fuji Photo Film Co. for arousing his interest in this problem and esepcially in the preparation of Figure 1. Shizuo Fujiwara and Takeo Yamamoto of the University of Tokyo are gratefully acknowledged for their encouragement and discussion.

LITERATURE CITED

(1) Balaban, A.T., and Harary, F., "The Characteristic Polynomial Does Not Uniquely Determine the Topology of a Molecule," *J. Chem. Doc.* **11,** 258-9 (1971).
(2) Frome, J., and O'Day, P.T., "A General Chemical Compound Code Sheet Format," *Ibid.*, **4,** 33-42 (1964).
(3) Gluck, D.J., "A Chemical Structure Storage and Search System Developed at Du Pont," *Ibid.*, **5,** 43-51 (1965).
(4) Hosoya, H., "Topological Index. A Newly Proposed Quantity Characterizing the Topological Nature of Structural Isomers of Saturated Hydrocarbons," *Bull. Chem. Soc. Japan* **44,** 2332-9 (1971).
(5) Hosoya, H., "Graphical Enumeration of the Coefficients of the Hückel Molecular Orbitals," *Theor. Chim. Acta* **25,** 215-22 (1972).
(6) Hosoya, H., Kawasaki, K., and Mizutani, K., "Topological Index and Thermodynamic Properties. I. Empirical Rules on the Boiling Point of Saturated Hydrocarbons," submitted to *Bull. Chem. Soc. Japan.*
(7) Kawasaki, K., Mizutani, K., and Hosoya, H., "Tables of Non-Adjacent Numbers, Characteristic Polynomials and Topological Indices. II. Mono- and Bicyclic Graphs," *Natural Science Report, Ochanomizu Univ.* **22,** 181-214 (1971).
(8) Mizutani, K., Kawasaki, K., and Hosoya, H., "Tables of Non-Adjacent Numbers, Characteristic Polynomials and Topological Indices. I. Tree Graphs," *Ibid.* **22,** 39-58 (1971).
(9) Morgan, H.L., "The Generation of a Unique Machine Description for Chemical Structures—A Technique Developed at Chemical Abstracts Service," *J. Chem. Doc.* **5,** 107-13 (1965).
(10) Patterson, A.M.P., Capell, L.T., and Walker, D.F., "Ring Index," 2nd ed., American Chemical Society, 1960.
(11) Plotkin, M., "Mathematical Basis of Ring-Finding Algorithms in CIDS," *J. Chem. Doc.* **11,** 60-3 (1971).
(12) Smith, E.G. ed., "The Wiswesser Line Formula Chemical Notation," McGraw-Hill Book Co., New York, 1968.
(13) Spann, M.L., and Willis, D.D., "A Comparative Study of a Fragmentation vs. a Topological Coding System in Chemical Substructure Searching," *J. Chem. Doc.* **11,** 43-7 (1971).
(14) Spialter, L., "The Atom Connectivity Matrix (ACM) and its Characteristic Polynomial (ACMP): A New Computer-Oriented Chemical Nomenclature," *J. Amer. Chem. Soc.* **85,** 2012-13 (1963); *J. Chem. Doc.* **4,** 261-74 (1964).

Search of CA Registry (1.25 Million Compounds) with the Topological Screens System

MARGARET MILNE,* DAVID LEFKOVITZ, HELEN HILL, and RUTH POWERS
Office of Engineering Research, University of Pennsylvania,
Philadelphia, Pa. 19104

Received June 2, 1972

The TSS (Topological Screens System) for substructure search was applied to the CAS Registry file of 1.25 million compounds, making it searchable on-line. The TSS screens and the use of the screen indexes are described. Statistics on screen assignment are provided, and the strengths and weaknesses of the system in general and in particular for a large file are discussed.

The Topological Screens System (TSS) is a screening system for substructure search which was developed at the University of Pennsylvania under NSF support. The system has been applied to the Chemical Abstracts Service 1.25 million compound Registry File making it searchable on-line. This application was carried out under the U.S. Army Chemical Information and Data Systems (CIDS) contract with the University and used a version of the CIDS on-line search system which allowed single terminal access of the file. This paper describes the TSS as applied to the CAS file, and relates preliminary experiences in its search.

The TSS was applied to the CAS Registry file using an IBM 7040. Because of limited disk storage capacity (one 1301 disk module), the CAS registry file was divided into two parts. The first contained the compounds with CAS

*To whom correspondence should be addressed.