for helpful discussions, and to the Department of Education and Science for the award of an Information Science Research Studentship.

REFERENCES AND NOTES

(1) R. E. Tarjan, "Graphic Algorithms in Chemical Computation", *Am. Chem. Soc. Symp. Ser.*, No. 46, 1–19 (1977).
(2) M. F. Lynch, "Screening Large Chemical Files", in "Chemical Information Systems", J. E. Ash and E. Hyde, Eds., Ellis Horwood, Chichester, 1975.
(3) A. Feldman and L. Hodes, "An Efficient Design for Chemical Structure Searching. I. The Screens", *J. Chem. Inf. Comput. Sci.*, **15**, 147–152 (1975).
(4) G. W. Adamson, J. Cowell, M. F. Lynch, A. H. W. McLure, W. G. Town, and A. M. Yapp, "Strategic Considerations in the Design of a Screening System for Substructure Searches of Chemical Structure Files", *J. Chem. Doc.*, **13**, 153–157 (1973).
(5) M. F. Lynch, "The Microstructure of Chemical Data-Bases and the Choice of Representation for Retrieval", in "Computer Representation and Manipulation of Chemical Information", W. T. Wipke, S. R. Heller, R. J. Feldman, and E. Hyde, Eds., Wiley, New York, 1973.
(6) C. E. Shannon, "A Mathematical Theory of Communication", *Bell Syst. Tech. J.*, **27**, 379–423, 623–656 (1948).
(7) G. W. Adamson, D. R. Lambourne, and M. F. Lynch, "Analysis of Structural Characteristics of Chemical Compounds in a Large Computer-Based File. Part III. Statistical Association of Fragment Incidence", *J. Chem. Soc. C*, 2428–2433 (1972).
(8) L. Hodes, "Selection of Descriptors According to Discrimination and Redundancy. Application to Chemical Structure Searching", *J. Chem. Inf. Comput. Sci.*, **16**, 88–93 (1976).
(9) M. Bersohn, "Rapid Generation of Reactants in Organic Synthesis Programs", *Am. Chem. Soc. Symp. Ser.*, No. 61, 128–147 (1977).
(10) H. L. Morgan, "The Generation of a Unique Machine-Description for Chemical Structures—a Technique Developed at Chemical Abstracts Service", *J. Chem. Doc.*, **5**, 107–113 (1965).
(11) L. A. Evans, M. F. Lynch, and P. Willett, "Structural Search Codes for On-Line Compound Registration", *J. Chem. Inf. Comput. Sci.*, **18**, 146–149 (1978).
(12) M. F. Lynch and P. Willett, "The Automatic Detection of Chemical Reaction Sites", *J. Chem. Inf. Comput. Sci.*, **18**, 154–159 (1978).
(13) S. H. Unger, "GIT—a Heuristic Program for Testing Pairs of Directed Line Graphs for Ismorphism", *Commun. ACM.*, **7**, 26–34 (1964).
(14) D. Cooper and M. F. Lynch, "The Compression of Wiswesser Line Notations Using Variety Generation", paper published in this issue.
(15) P. W. Williams, "Criteria for Choosing Subsets to Obtain Maximum Relative Entropy", *Comput. J.*, **21**, 57–62 (1978).
(16) E. J. Schuegraf and H. S. Heaps, "Selection of Equifrequent Word Fragments for Information Retrieval", *Inf. Storage Retr.*, **9**, 697–711 (1973).
(17) M. F. Lynch, "Variety Generation—a Reinterpretation of Shannon's Mathematical Theory of Communication and Its Implications for Information Science", *J. Am. Soc. Inf. Sci.*, **28**, 19–25 (1977).
(18) M. F. Lynch, Principal Investigator, "Comparison of the Efficiency of Bibliographic Search Codes", British Library, Research and Development Department Report No. 5422, 1978.
(19) P. Willett, "Computer Analysis of Chemical Reaction Information for Storage and Retrieval", unpublished Ph.D. thesis, University of Sheffield, 1978.
(20) J. E. Ash, "Connection Tables and Their Role in a System", in J. E. Ash and E. Hyde, Eds., ref 2.
(21) G. W. Adamson, M. F. Lynch, and W. G. Town, "Analysis of Structural Characteristics of Chemical Compounds in a Computer-Based File. Part II. Atom-Centred Fragments", *J. Chem. Soc. C*, 3702–3706 (1971).
(22) Alternative descriptions of circular substructures have been described by several authors, but the methods of feature description and manipulation are very different; see, e.g., J. E. Dubois, "Ordered Chromatic Graphs and Limited Environment Concepts" in "Chemical Applications of Graph Theory", A. T. Balaban, Ed., Academic Press, London, 1976; W. Schubert and I. Ugi, "Constitutional Symmetry and Unique Descriptors of Molecules", *J. Am. Chem. Soc.*, **100**, 37–41 (1978); M. Randic, "Fragment Search in Acyclic Structures", *J. Chem. Inf. Comput. Sci.*, **18**, 101–107 (1978).

Plausible Paths in the Rearrangement Reaction of Polycyclic Hydrocarbons Searched by the Graph-Theoretical Method and Computer Techniques†

NOBUHIDE TANAKA and TADAYOSHI KAN

Department of Physics, Faculty of Science, Gakushuin University, Mejiro, Tokyo, 171, Japan

TAKESHI IIZUKA*

Department of Chemistry, Faculty of Education, Gunma University, Maebashi, Gunma, 371, Japan

Received August 11, 1978

We have formulated a graph-theoretical concept "transmutation" which corresponds to the change of the skeleton of a molecule. In order to see the relations between two given graphs, we have devised two algorithms, mono-source and di-source propagation algorithms. We applied these algorithms to a transmutation process corresponding to adamantane and diamantane rearrangements and obtained relationships between the transmuted graphs. In the di-source propagation algorithm, we show the "shortest paths" which correspond to the plausible paths of the rearrangement reaction.

We have studied graph-theoretical and computational isomer enumeration,[1] representation of molecular structures,[2] and analysis of rearrangement paths of polycyclic hydrocarbons.[3] In this paper we focus on methods of studying the rearrangement paths of polycyclic hydrocarbons by graph theory and computer methods.

Balaban et al.[4] first studied graph theoretically 1,2 shifts of acyclic hydrocarbons and tried to classify the rearrangement reactions.[5] A study of the cyclic hydrocarbon adamantane, based on graph theory and a computer method, was reported by Whitlock and Siefken.[6] They found the rearrangement paths from twistane to adamantane by graph-theoretical predictions. Using calculations by molecular mechanics, Schleyer et al. challenged the diamantane rearrangement.[7] We think, however, that the rearrangement reaction paths become more complicated as the number of carbons and/or rings increases. The complexity is far beyond the limit of manual work. Thus it is essential to devise efficient computer methods for this problem.

We define the graph-theoretical concept, which we call "transmutation", as corresponding to chemical rearrangement, and devise computer methods to obtain plausible[13] information without imposing physicochemical conditions. We have established two algorithms to find relationships between isomers

† A part of this paper was presented by N. T. in the 26th IUPAC Congress.[3d]

0095-2338/79/1619-0162$01.00/0

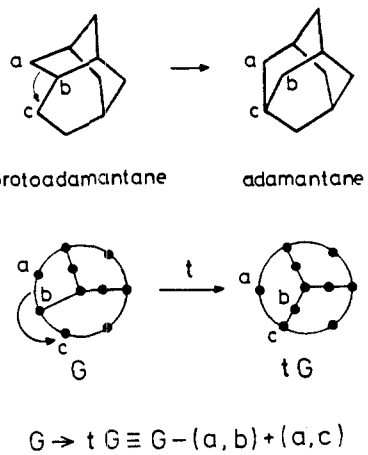

Figure 1. Rearrangement and transmutation.

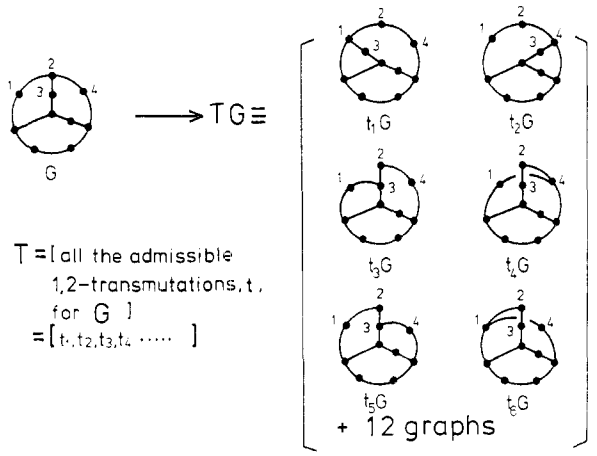

Figure 2. Transmutation on a graph.

and obtained fruitful results as shown below.

FORMULATION

We have graph theoretically studied the change of hydrocarbon skeletons and found algorithms to obtain relationships between the isomers in question.

We follow two assumptions adopted by Whitlock et al. and Schleyer et al. to investigate adamantane and diamantane rearrangements.[6,7] (1) The reaction proceeds via 1,2 shifts. (2) The isomers which take part in the reaction (a) have neither three- nor four-membered rings, and (b) have only one or no quaternary carbon atom.[15] For the description of the reactions we use a graph which shows the intramolecular relationships between skeleton carbon atoms. Roughly speaking we make a bond migration in the chemical reaction correspond to an edge migration of a skeleton graph, which we call a transmutation (Figure 1).

We define a transmutation of a given graph G for more theoretical study as follows.

(a) An element of a transmutation of G is an operation which consists of removing an edge (a, b) incident on vertices a and b in G, and adding an edge (c, d) incident on vertices c and d in G. Then we obtain a graph $G' = G - (a,b) + (c,d)$ by this operation.

(b) An element of a restricted transmutation of G is an element of a transmutation of G under the condition, $a = c$, that is, the removed edge and the added edge have the same vertex.

(c) An element of the 1, n-transmutation of G is an element of a restricted transmutation of G under the condition that the distance between the vertex b and the vertex d is equal to $n + 1$.

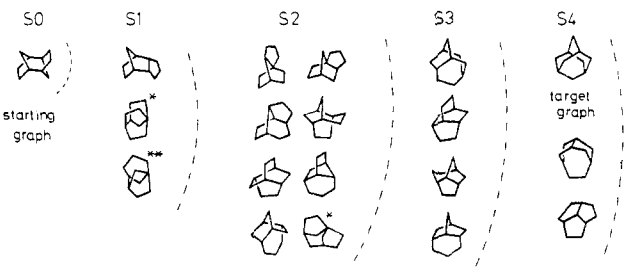

Figure 3. MSPA in adamantane rearrangement. The isomers with * have large strain and cannot be constructed with molecular models except one with **.

Table I

generation:	0th	1st	2nd	3rd	4th	total
starting from	no. 6	5	39	213	939	1197
starting from	diaman.	4	68	445	1571	2089

An element of a transmutation t is determined by choosing vertices a, b, c, d which satisfy the conditions in question. A transmutation T is the set of all the elements mentioned above.

We shall study graph theoretically diamantane rearrangement under 1,2 shifts as the 1,2-transmutation whose element switches an edge off from a vertex of degree 3 or 4 and switches the edge on a neighbor vertex of degree 2 or 3 in a skeleton graph, because the 1,2 shift from the tertiary or the quaternary carbon atom to the secondary or the tertiary carbon atom is a chemically very realizable reaction. It is known that under the 1,2-transmutation, a graph of a diamantane isomer does not become disconnected.[8] We apply this operation to all the vertices of degree 3 or degree 4 in a graph g and obtain the set Tg of graphs.

ALGORITHMS

We have devised two algorithms to find relationships between a starting graph and a target graph. The first is the "mono-source propagation" algorithm (MSPA) and the second is the "di-source propagation" algorithm (DSPA).

Mono-source Propagation Algorithm:

(1) Input a graph g in question and a limit of i.
(2) Initialization: put $S_0 = \{g\}$, $S_{-1} = \phi$, and set $i = 0$.
(3) Put $S_{i+1} = \phi$.
(4) Take a graph g in the set S_i which we call the ith generation.
(5) Do a transmutation T on g under the adopted conditions and get the set Tg.
(6) If $Tg = \phi$ then go to (8).
(7) For each graph f in Tg, do as follows: If f were in S_{i-1}, S_i, or S_{i+1}, output the relationship between g and f, else add f to S_{i+1} and output the relationship.
(8) If all elements of S_i were not used then go to (4).
(9) If the target graph were found in S_{i+1} then go to (11).
(10) Set $i = i + 1$ and if i is less than the limit then go to (3).
(11) End.

Thus the successive generations of graphs are produced like propagation of waves which generate around a stone thrown on the surface of water. Using MSPA the successive generations to adamantane starting from tricyclo[4.2.1.1^{2,5}]decane are shown in Figure 3. Starting from the graph, no. 6[9] (Figure 4), of diamantane isomers, the number of related graphs increases as shown in the first row of Table I. The target graph which corresponds to diamantane is found in the 4th generation. By inversion of the starting graph and the target graph, that is, MSPA from diamantane, the numbers in the second row of Table I are obtained.

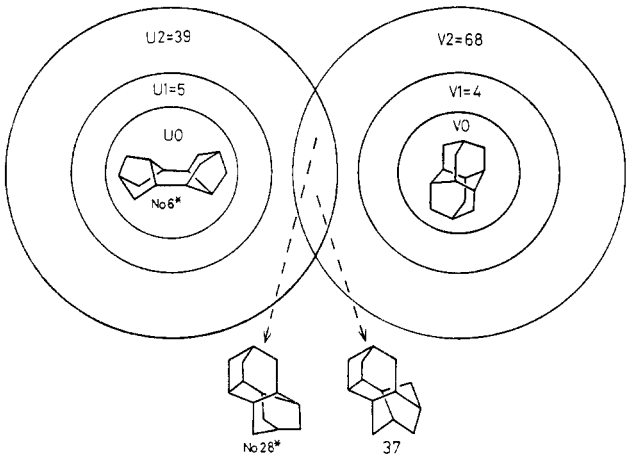

Figure 4. DSPA in diamantane rearrangement and the common set. The isomer 37 which has been newly found by DSPA can be constructed with molecular models without much strain. It would be better to consider it a member of the rearrangement path.

Di-source Propagation Algorithm

(1) Input two sources which are a starting graph and a target graph as an element of the 0th generation U_0 and an element of the 0th generation V_0, respectively.
(2) Put $V_{-1} = U_{-1} = \phi$ and set $i = j = 0$.
(3) If $|U_i| \leq |V_j|$, then construct the next generation U_{i+1} by using MSPA and set $i = i + 1$, else construct V_{j+1} in the same way and set $j = j + 1$.
(4) If $U_i \cap V_j = \phi$, then go to (3).
(5) Set the common sets $X_i \equiv Y_j = U_i \cap V_j$.
(6) Pick up the elements in U_{i-1} which can be traced back from X_i, then construct the predecessor X_{i-1} by them, and set $i = i - 1$.
(7) If $i > 1$, then go to (6).
(8) Pick up the elements in V_{j-1} which can be traced back from Y_j, then construct the predecessor Y_{j-1} by them, and set $j = j - 1$.
(9) If $j > 1$ then go to (8).
(10) End.

The sequence $U_0 = X_0, X_1, \ldots, X_i = Y_j, \ldots, Y_1, Y_0 = V_0$ is the "shortest paths" between two sources under the transmutation.

Taking no. 6 and diamantane as two sources, we can find the two common graphs in the second generation from each source by this method (Figure 4).

RESULTS

Comparing the number of graphs in each step of MSPA with that of DSPA, 1197 graphs in the case of MSPA and 118 graphs in the case of DSPA were generated to find all the transmutational relationships between diamantane and no. 6. The latter is one-tenth of the former! As an extention of DSPA, the poly-source propagation algorithm is easily formulated. As an example of results obtained by the poly-source propagation algorithm, Figure 5 shows an aspect of the transmutational relationships between no. 9 and diamantane through no. 6.

We have programmed MSPA and DSPA using Fortran IV for MELCOM7700. Graphs are stored and retrieved by using the "edge table" as a list of edges.[10] Our modified Morgan method[11] is adopted for uniquely naming and for identification of graphs. The computing time to find diamantane from no. 6 in Figure 4 by MSPA is 22 min. Also it takes 13 min to find the shortest paths between no. 6 and diamantane in Figure 4 by DSPA.

DISCUSSION

From our viewpoint we can say that Whitlock and Siefken adopted the global method[12] and succeeded in giving all the transmutational relationships of the graphs appearing in the adamantane rearrangement.

However, we think, it would be difficult to use the global method to study the paths in the diamantane rearrangement, because more than 5000 plausible[13] graphs[14] appear in the diamantane rearrangement. Therefore, it is quite clear that reasonable participants should be selected. By using DSPA we have succeeded in selecting plausible graphs as the participants needed to find the shortest paths in the diamantane rearrangement.

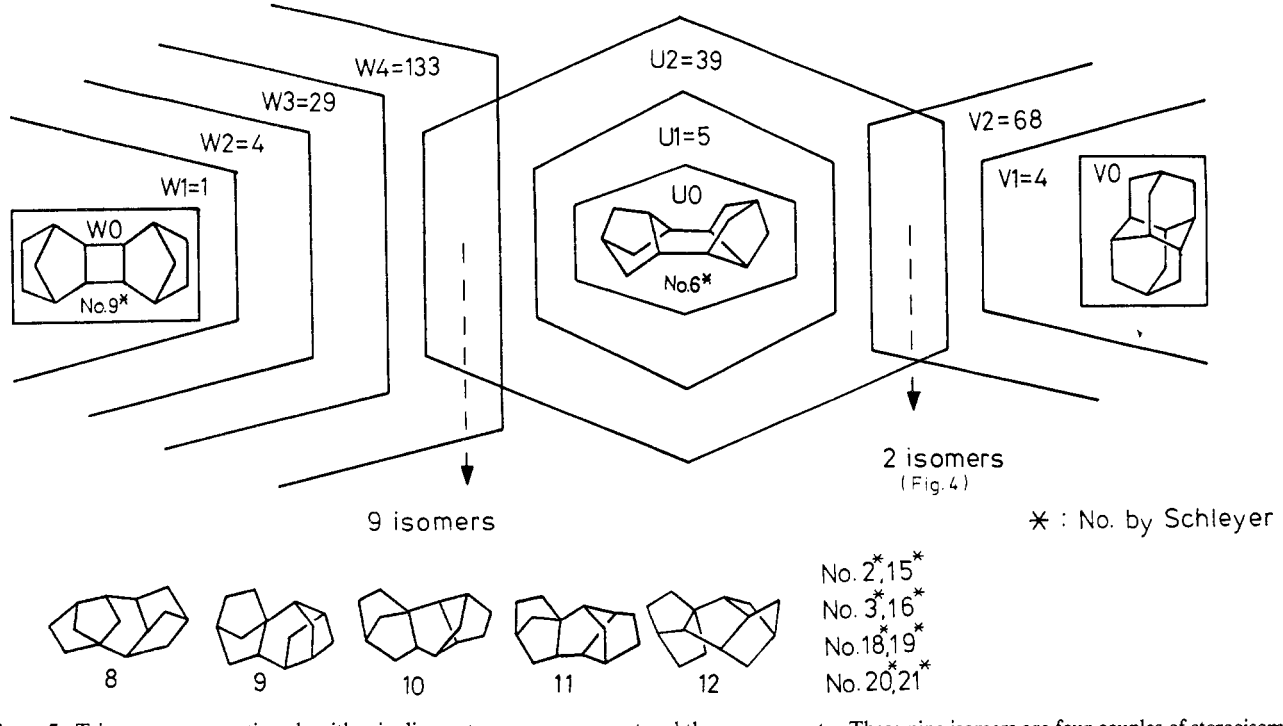

Figure 5. Tri-source-propagation algorithm in diamantane rearrangement and the common sets. These nine isomers are four couples of stereoisomers by Schleyer[7] and five isomers, 8–12, which can be constructed with molecular models even though with more or less strain.

Meanwhile, Schleyer et al. gave some diamantane rearrangement paths by molecular mechanics, but it is questionable whether there are other paths or not. As they rambled walking paths with many branches, they did trial calculations by molecular mechanics to find realizable paths.

If we select graph-theoretical applicants by our methods and carry out the calculations by molecular mechanics, we can predict more realizable paths with less work than using Schleyer's paths. Because even if the graph-theoretical 1,2-transmutation is possible, the chemical 1,2 shift is not always possible; however, when the former is impossible, the latter is also impossible. Therefore the shortest path which is derived by DSPA is not always a realizable path and the most favorable (thermodynamically) path is sometimes longer than the shortest path as in a tetramethylene adamantane rearrangement.[15]

We believe that the graph-theoretical methods with computer techniques developed by us would be an efficient way to study the rearrangement paths in heptacyclic triamantanes.[16]

ACKNOWLEDGMENT

This work was supported in part by a Scientific Research Grant from the Ministry of Education, Japan.

REFERENCES AND NOTES

(1) (a) N. Tanaka, T. Iizuka, and T. Kan, "A Graph-Theoretical Derivation of The Isomers of Tricyclic Hydrocarbons", *Chem. Lett.*, 539–544 (1974); (b) T. Iizuka, N. Tanaka, and T. Kan, "On the Isomers of Tricyclic Hydrocarbons C_nH_{2n-4}", *Sci. Rep. Fac. Educ., Gunma Univ.*, **22**, 47–56 (1973).

(2) T. Iizuka, H. Miura, and T. Kan, "Polycyclic Blocks and Bridged Polycyclic Compounds", *Sci. Rep. Fac. Educ., Gunma Univ.*, **26**, 79–93 (1977).

(3) (a) T. Iizuka, N. Tanaka, and T. Kan, "On the Rearrangement Reactions of Tricyclic Hydrocarbons $C_{10}H_{16}$ and $C_{11}H_{18}$", *Sci. Rep. Fac. Educ., Gunma Univ.*, **23**, 105–117 (1974); (b) "On the 1,2-, 1,3-, 1,*n*-Graph Rearrangements Corresponding to Adamantane Rearrangements", *ibid.*, **25**, 57–70 (1975); (c) "Graph-Transmutation Applied to the Rearrangement Reaction of Pentacyclo Hydrocarbons", *ibid.*, **25**, 81–89 (1976); (d) T. Kan, N. Tanaka, H. Miura, and T. Iizuka, "Chemical Rearrangement and Graph Transmutation", 26th IUPAC Congress in Tokyo, Session IV 7B408, 1977, p 1033.

(4) A. T. Balaban, D. Farcasiu, and R. Banica, "Graphs of Multiple 1,2-shifts in Carbonium Ions and Related Systems", *Rev. Roum. Chim.*, **11**, 1205–1227 (1968).

(5) A. Balabas and A. T. Balaban, "Isomerizations and Isographic Transformations", *Rev. Roum. Chim.*, **19**, 1927–1940 (1974).

(6) (a) H. W. Whitlock, Jr., and M. W. Siefken, "The Tricyclo-[4.4.0.0^{3,8}]decane to Adamantane Rearrangement", *J. Am. Chem. Soc.*, **90**, 4929–4939 (1968). (b) Smith et al. considered isomers which have two quaternary carbon atoms in adamantane rearrangement pathways: T. H. Varkony, R. E. Carhart, and D. H. Smith in "Computer-Assisted Organic Synthesis", W. T. Wipke, Ed., American Chemical Society, Washington, D. C., 1977, p 188.

(7) (a) T. M. Gund, P. v. R. Schleyer, P. H. Gund, and W. T. Wipke, "Computer Assisted Graph Theoretical Analysis of Complex Mechanistic Problems in Polycyclic Hydrocarbons. The Mechanism of Diamantane Formation from Various Pentacyclotetradecanes", *J. Am. Chem. Soc.*, **97**, 743–747 (1975); (b) E. Osawa, K. Aigami, N. Takaishi, Y. Inamoto, Y. Fujikura, Z. Majerski, P. v. R. Schleyer, E. M. Engler, and M. Farcasiu, "The Mechanisms of Carbonium Ion Rearrangements of Tricyclodecanes of Elucidated by Empirical Force Field Calculations", *ibid.*, **99**, 5361–5373 (1977).

(8) No generation of the disconnected graphs is examined in the five cyclic block graph.[2]

(9) See Schleyers's notations in ref. 7.

(10) E. M. Reingold, J. Nievergelt, and N. Deo in "Combinatorial Algorithms", Prentice-Hall, New York, 1977, Chapter 8.

(11) (a) H. L. Morgan, "The Generation of a Unique Machine Description for Chemical Structures—A Technique Developed at Chemical Abstracts Service", *J. Chem. Doc.*, **5**, 107 (1965); (b) W. T. Wipke and T. M. Dyott, "Stereochemically Unique Naming Algorithm", *J. Am. Chem. Soc.*, **96**, 4834–4842 (1974); (c) H. Miura, N. Tanaka, T. Iizuka, and T. Kan, unpublished data. We have used the other weight, prime number, rather than the extended connectivity for classification of points in a graph.

(12) As a classification of the method to find the relationships between graphs, we defined[3d] two methods called "global method" and "local method", which consist of MSPA and DSPA, and TSPA, etc.

(13) We differ between plausible and realizable; the former is mathematically possible, the latter is chemically or physically possible.

(14) H. Miura, T. Iizuka, and T. Kan, 35th Autumn Annual Meeting of the Chemical Society of Japan, 1976, Abstracts, I, 47.

(15) H. Tahara, E. Osawa, T. Iizuka, N. Tanaka, and T. Kan, "Mechanism of Tetracyclic Hydrocarbon $C_{14}H_{22}$," 1979 Winter Meeting in Hokkaido District, Abst. no. 1B05, Feb. 1–2, at Sappro.

(16) We have already applied DSPA to triamantane rearrangement, using molecular force field calculation: N. Tanaka, M. Imai, T. Kan, and T. Iizuka, "Selective Di-source Propagation Algorithm," presented at the ACS/CSJ Chemical Congress, Honolulu, Hawaii, April 1–6, 1979.

Compression of Wiswesser Line Notations Using Variety Generation

DAVID COOPER and MICHAEL F. LYNCH*

Postgraduate School of Librarianship and Information Science, University of Sheffield, Western Bank, Sheffield, S10 2TN, United Kingdom

Received October 17, 1978

The use of variety generation for reversible text compression is described briefly, and it is shown how the technique may be applied to compress Wiswesser Line Notations. The notations may be compressed, using 8-bit codes to represent variable-length character strings, to occupy an average of just under 3.6 bits per original character, an improvement of just over 55% on a fixed-length representation using 8 bits per character. This is similar to the amount of compression given by the same technique on natural language texts.

INTRODUCTION

The Wiswesser Line Notation (WLN) is widely used for storing chemical structure information in a linear form. For many applications it is a more convenient vehicle than connection tables for storing structures in machine-readable files, since it is more compact and can be input and output in a recognizable form without complex manipulations. Details of the WLN are given by Smith[1] and Ash and Hyde.[2]

When large files of WLNs are stored in machine-readable form or transmitted over telecommunication channels, it is clearly desirable that they occupy as small a space as possible. Certain contractions were introduced soon after the invention of the WLN, and can give a reduction of something like 5% in the number of characters needed.[3] However, the contraction and expansion are difficult to perform automatically, and it is often thought that the small saving in space does not compensate for the difficulties involved.

The basic WLN alphabet consists of 40 characters: the 26 letters of the Roman alphabet, the 10 digits, and the characters /, –, &, and space. (Other characters, such as asterisks in a notation for polymers, are now sometimes used, but their presence in a file makes no difference to the techniques

0095-2338/79/1619-0165$01.00/0