(14) Silverstein, R. M.; Bassler, G. C.; Morrill, T. C. *Spectrometric Identification of Organic Compounds*, 4th ed.; John Wiley & Sons: New York, 1981.
(15) AMPAC: QCPE No. 506, Indiana University, Chemistry Department.
(16) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902.
(17) Program runs on the entire range of VAX computers (VMS 5.1 or higher). The minimum hardware required to run MACCS-3D (rev. 1.0), which includes MACCS-II and the Customization Module, is a MicroVax-2000 with 8-Mbyte memory (16 Mbyte recommended). The program code takes less than 25 000 blocks of disk space; depending on the size of the database(s), up to 300 000 blocks of disk space is recommended initially (FCD-3D with 65 000 compounds takes 180 000 blocks). The price of the software depends on the size of the CPU and the number of users.
(18) Customization Module: a module of MACCS-II that enables users to write applications and front-ends to the software, available from Molecular Design Limited.
(19) Pople, J. A.; Segal, J. A. CNDO/2. *J. Chem. Phys.* **1966**, *44*, 3289.
(20) Jain, P. C.; Mukerjee, Y. N.; Anand, N. *J. Am. Chem. Soc.* **1974**, *96*, 2996.
(21) For an excellent review of the frontier molecular orbital concept see: Woodward, R. B.; Hoffmann, R. *Conservation of Orbital Symmetry*; Verlag Chemie GmbH: Weinheim, 1970; and Fleming, I. *Frontier Orbitals and Organic Chemical Reactions*; Wiley, New York: 1976. Results of ab initio calculations in concert with the observation in text is published: Güner, O. F.; Ottenbrite, R. M.; Shillady, D. D.; Alston, P. V. *J. Org. Chem.* **1987**, *52*, 391.
(22) Sustmann, R. *Tetrahedron Lett.* **1971**, 2717 and 2721.
(23) Güner, O. F.; Ottenbrite, R. M.; Shillady, D. D.; Alston, P. V. *J. Org. Chem.* **1988**, *53*, 5348.
(24) Brown, H. D. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 155-158.

# A New Method of Computer Representation of Stereochemistry. Transforming a Stereochemical Structure into a Graph

TATSUYA AKUTSU

Mechanical Engineering Laboratory, 1-2 Namiki, Tsukuba, Ibaraki, Japan 305

Received February 2, 1991

A new method of computer representation of stereochemical structures, which include double bonds and asymmetric carbon atoms, is described. The method is very simple, and a stereochemical structure is transformed into a graph. From the results, graph algorithms, which have been intensively studied in computer science, can be directly applied to chemical structures. Especially, a polynomial time algorithm for stereochemically unique naming is implied, for which SEMA (stereochemically extended Morgan algorithm) does not work in polynomial time.

## INTRODUCTION

Representation and manipulation of chemical structures are very important for database systems and expert systems in chemistry. Especially, unique naming[3,6,7,9,11] and substructure matching[12] are most important. Usually, a chemical structure is represented as a graph, in which an atom corresponds to a vertex and a chemical bond corresponds to an edge. However, a graph is not sufficient for representing a chemical structure. Stereoisomers must be distinguished. Though they have the same graph structures, their geometric structures are different and they show different properties. How to represent stereoisomers in computer systems has been studied well. Especially, the works of Wipke and Dyott are well known. They developed the stereochemically unique naming algorithm[13,14] (stereochemically extended Morgan algorithm, abbreviated as SEMA) based on the ordered list representation of stereochemical structures by Petrarca et al.[10]

In this paper, another approach to distinguish stereoisomers is presented. A basic technique in computer science "transformation" is employed. A chemical structure is transformed into a usual graph (a structure which does not have stereochemical information). Besides, two structures are transformed into isomorphic graphs, if and only if they are stereochemically isomorphic.

## ORDERED LIST METHOD

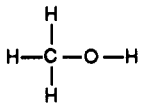
Before describing the transformation method, the ordered list method is reviewed. It was developed by Petrarca et al.[10] and adapted in the SEMA algorithm by Wipke and Dyott.[13,14] However, the original method is not described, but the method proposed in the CHAUS system[2] is described since it is simpler.

In this paper, only stereoisomers caused by the following local structures are considered (see Figure 1).

(1) A pair of carbon atoms connected with a double bond
(2) An asymmetric carbon atom adjacent to four atoms

Other cases (such as conformation) are considered to be handled in a similar way.

Basically, chemical structures are represented as graphs in the ordered list method. In graph representation, an atom corresponds to a vertex and a chemical bond corresponds to an edge. Note that, in this paper, hydrogen atoms are not graphically abbreviated. For example, methanol is represented as



but is not represented as $CH_3$–OH.

The adjacency list, which is a famous data structure in computer science, is employed to represent a graph (see Figure 2). The adjacency list is essentially the same as the connection table, which is popular in chemical information processing. In the adjacency list, there is a list of attached atoms for each atom. The ordering of atoms in the list has no meaning, and an arbitrary ordering is allowed. However, the ordering is used to represent stereochemical information in the ordered list method.

At first, consider the case of a pair of carbon atoms connected with a double bond. For each carbon atom, adjacent atoms are listed in a clockwise order beginning with the other carbon atom (see Figure 3). Although four (= 2 × 2!) different orderings can be considered, two equivalent orderings are allowed (see Figure 3). The other two equivalent orderings are generated for the stereoisomer. Due to the ambiguity of the orderings, stereochemically isomorphic structures are not

0095-2338/91/1631-0414$02.50/0 

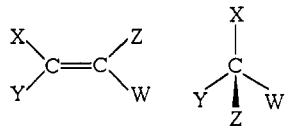


**Figure 1.** Two types of local stereochemical structures.

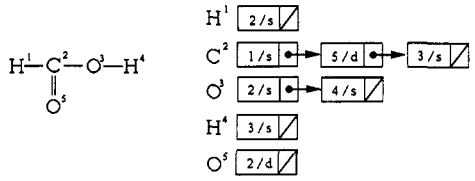


**Figure 2.** Example of the adjacency list. Superscripts on atom names do not indicate isotope. They are just numbers to distinguish between atoms. Symbols "s" and "d" in the adjacency list denote a single bond and a double bond, respectively.

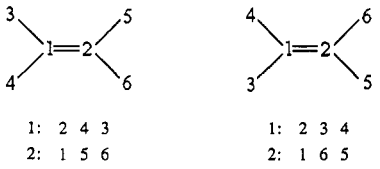


**Figure 3.** Two equivalent orderings for a pair of carbon atoms connected with a double bond.

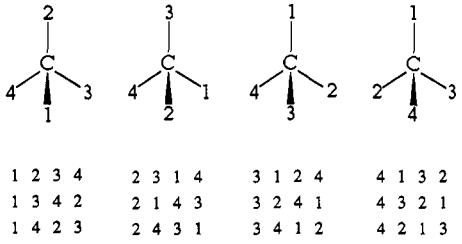


**Figure 4.** Twelve equivalent orderings for an asymmetric carbon atom.

necessarily represented as the same form. However, the equivalent orderings are identified in such algorithms as unique naming algorithms and substructure matching algorithms.

Next, consider the case of an asymmetric carbon atom adjacent to four atoms. If we do not consider stereochemical information, 24 (= 4!) different orderings are generated. However, only 12 orderings are allowed as shown in the Figure 4. These orderings are generated by viewing the bond down from the first adjacent atom to the asymmetric carbon atom; the other adjacent atoms are arranged in a clockwise manner. Of course, there is ambiguity in selecting the first and the second adjacent atoms, so that 12 (= 4 × 3) orderings are generated. The other equivalent 12 orderings are generated for the stereoisomer. Note that equivalent orderings have the same parity, where the parity is odd if the number of pairwise interchanges necessary to sort the ordering in an ascending manner is odd, otherwise the parity is even. As in the previous case, equivalent orderings are identified by the algorithms of chemical structure manipulation.

## TRANSFORMATION METHOD

The ordered list method is simple and practical. However, when we want to apply graph algorithms, they must be modified to identify equivalent orderings. Though modification is easy in most cases, it is very difficult in the case of complicated algorithms such as the polynomial time unique naming algorithm for the graphs of bounded valence.[4] Therefore, a method to use graph algorithms without modification is required.

In the method presented in this paper, a stereochemical structure is transformed into a usual graph. The transformation has the property in which two structures are transformed into isomorphic graphs, if and only if they are stereochemically isomorphic. Moreover, the transformation is simple and can be done efficiently. Only the local transformation of a structure is required.

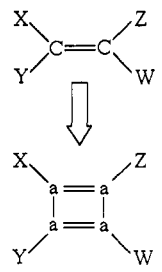


**Figure 5.** Transformation for a pair of carbon atoms connected with a double bond.

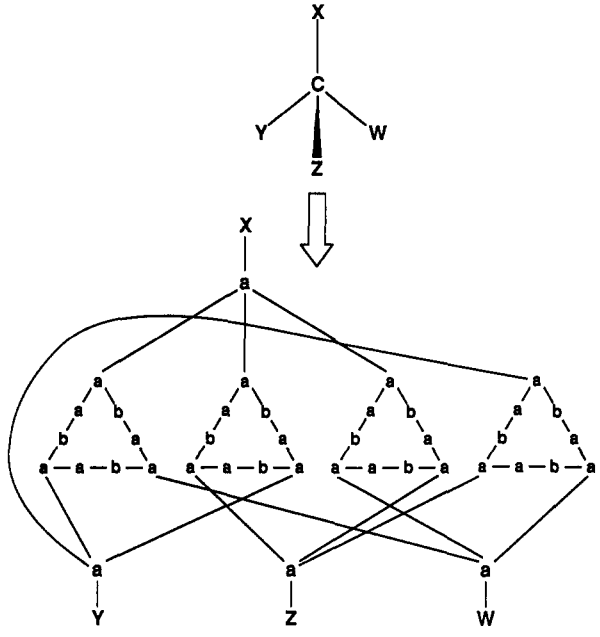


**Figure 6.** Transformation for an asymmetric carbon atom.

In the case of a pair of carbon atoms connected with a double bond, the (local) structure is transformed into one as shown in Figure 5. Note that there is no meaning in the ordering of the adjacency list of the transformed graph.

In the case of an asymmetric carbon adjacent to four atoms, the (local) transformed structure is shown in Figure 6. There is no meaning in the ordering of the adjacency list of the transformed graph, too. (a) and (b) in Figures 5 and 6 are introduced as new types of atoms for convenience. The transformation is developed so that the transformed structures are locally isomorphic, if and only if the orderings of adjacency lists of asymmetric carbon atoms are equivalent. We will explain this using the examples in Figure 7. The orderings of adjacency lists of original chemical structures in (a) and (c) are equivalent, while those in (a) and (b) are not equivalent. Consider the path from X to Y and the path from X to W, both of which pass the same triangle. In the case of (a) and (c), the atom sequence of the path from X to Y is

X a a a b a a Y

and the one from X to W is

X a a b a a a W

However, in the case of (b), the atom sequence of the path from X to Y is

X a a b a a a Y

and the one from X to W is

X a a a b a a W

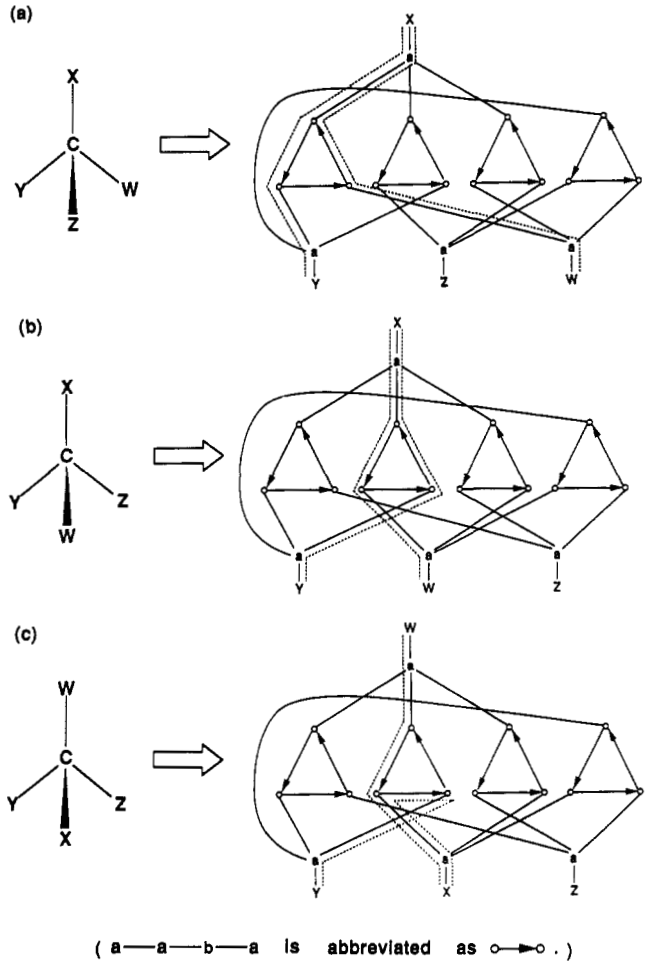



**Figure 7.** Examples for the explanation of the property of transformation.

Although the above discussion is not complete, it can easily be seen that the property mentioned before holds.

If there are $n$ pairs of carbon atoms connected with double bonds and $m$ asymmetric carbon atoms adjacent to four atoms, $n + m$ (local) transformations are required. Examples are shown in Figure 8. The correctness of the transformation is almost obvious from the above discussion. If two asymmetric carbons (each carbon is adjacent to four atoms) have equivalent orderings of the adjacency lists, transformed structures are locally isomorphic. If two asymmetric carbons (each carbon is adjacent to four atoms) do not have equivalent orderings of the adjacency lists, transformed structures are not locally isomorphic. In the case of the pair of carbon atoms connected with a double bond, a similar property holds.

## POLYNOMIAL TIME UNIQUE NAMING ALGORITHM

In computer science, an algorithm is called a polynomial time algorithm[1,5] if it always outputs the correct answer within $p(n)$ time for each input of size $n$, where $p(n)$ is a polynomial of $n$. For example, FFT is a polynomial time algorithm since it computes the Fourier transformation within $k_1 n \log(n) + k_2$ time for each piece of data of size $n$, where $k_1$ and $k_2$ are the fixed constants. Note that, $k_1 n \log(n) + k_2 < n^2$ holds if $n$ is large enough. In theoretical computer science, an algorithm is called good or efficient if it is a polynomial time one. On the other hand, an exponential time algorithm, which requires greater than or equal to $ca^n$ time in the worst case where $c$ and $a$ are the constants, is not called efficient. Note that, $p(n) \leq ca^n$ holds if $n$ is large enough.

Many algorithms have been developed for generating a unique invariant name (a canonical name) for each chemical

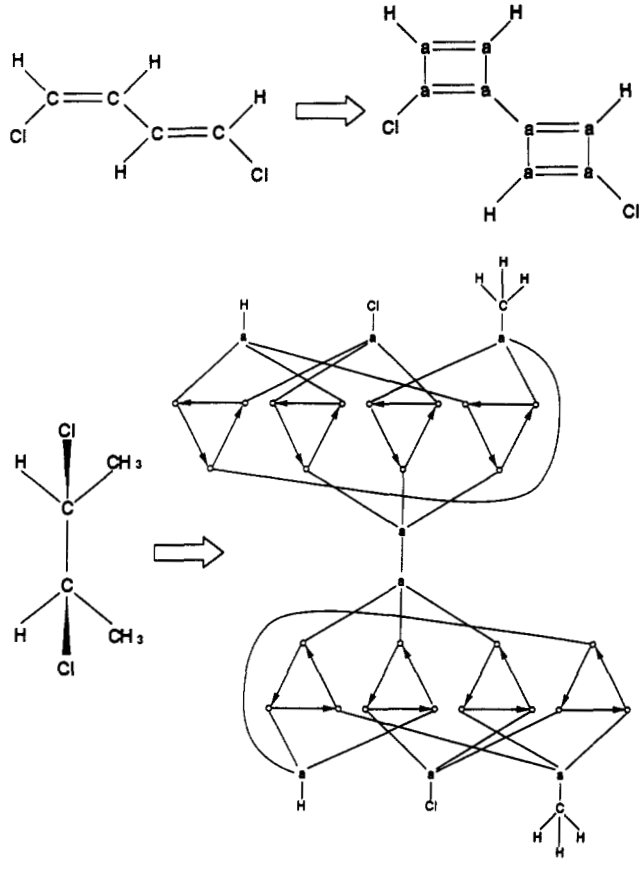


**Figure 8.** Examples of the original chemical structures and their transformed graphs.

structure.[3,6,7,9,11] Unfortunately, they do not seem to be polynomial time algorithms. Nondeterministic choice or enumeration of permutations is included by most of them, though efforts are made to avoid the use of it. It causes exponential processing time. Of course, they seem to work efficiently (i.e., in polynomial time) for most inputs. However, it will take a long time (exponential time) when there is much symmetry in the input chemical structure. Unique naming for each chemical structure is closely related to the graph isomorphism problem since the same name is given to two graphs, if and only if they are isomorphic. It is a famous open problem in computer science whether there is a polynomial time algorithm for testing graph isomorphism or not. However, a polynomial time algorithm of graph isomorphism was found for the graphs of bounded valence.[8] Moreover, a polynomial time unique naming algorithm for the graphs of bounded valence was found.[4] Since every chemical structure is of bounded valence, their results can be applied to one, and a polynomial time algorithm for unique naming can be obtained. Moreover, combining with the results of the previous section, a polynomial time algorithm for stereochemically unique naming can be obtained, since transformed graphs are isomorphic if and only if the original chemical structures are stereochemically isomorphic. The polynomiality of the time complexity (processing time) is stated by the following discussion. Note that the following discussion is done not only for unique naming but also for general cases.

It is easy to see that the size (the number of nodes) of the transformed graph is at most 40 times as large as the original chemical structure (let $c = 40$). Also, the transformation can be carried out within $d_1 n + d_0$ time for a chemical structure which consists of $n$ atoms, where $d_1$ and $d_0$ are the constants. Assume there is an algorithm for usual graphs which works within

$$a_k n^k + a_{k-1} n^{k-1} + \ldots + a_0$$

time for a graph of size $n$. If a chemical structure of size $n$ is transformed into a usual graph and the algorithm is applied, the total processing time becomes

$$c^k a_k n^k + c^{k-1} a_{k-1} n^{k-1} + \ldots + a_0 + d_1 n + d_0 + b$$

where $b$ denotes time required for subroutine call.

However, as the unique naming algorithm for the graphs of bounded valence is too complicated to implement (it is based on the group theory) and the degree of the polynomial is large, it is not considered to be practical. For almost all inputs occurred in practice, such algorithms as the Morgan algorithm seem to work much more efficiently.

What we want to point out here is that there is a possibility that more practical and efficient algorithms for unique naming of graphs will be found. Once such an algorithm is found, it can be applied to stereochemically unique naming by using the results of this paper.

The transformation method can be applied not only to unique naming but also to substructure matching, because transformed graphs are locally ismorphic if and only if original chemical structures are locally and stereochemically isomorphic. Unfortunately, no polynomial time algorithm is known for subgraph matching even if graphs are restricted to the ones of bounded valence.[5] However, once an efficient algorithm for subgraph matching is found, it can be directly applied to stereochemically substructure matching.

## CONCLUSION

A method which transforms stereochemical structures into graphs (structures which do not have stereochemical information) is presented. The transformation is very simple, and two structures are transformed into isomorphic graphs, if and only if they are stereochemically isomorphic. By using the method, graph algorithms for usual graphs such as unique naming algorithms and subgraph matching algorithms can be applied to stereochemical structures. That is, when an efficient algorithm for usual graphs is found, it can be applied to stereochemical structures. A polynomial time algorithm for stereochemically unique naming is implied as an example.

## ACKNOWLEDGMENT

I thank Prof. Ohsuga of the University of Tokyo for giving me a chance to study computer application to chemistry. Also, I would like to express my gratitude to Prof. Sasaki and Dr. Funatsu of Toyohashi University of Technology for their valuable discussions and suggestions on chemical information processing.

## REFERENCES AND NOTES

(1) Aho, A. V.; Hopcroft, J. E.; Ullman, J. D. *The Design and Analysis of Computer Algorithms*; Addison-Wesley: Reading, MA, 1974.
(2) Akutsu, T.; Suzuki, E.; Ohsuga, S. A. Logic Based Approach to Expert Systems in Chemistry. *Knowl. Based Sys.* (in preparation).
(3) Dubois, J. E. French National Policy for Chemical Information and the DRAC System as a Potential Tool of This Policy. *J. Chem. Doc.* **1973**, *13*, 8-13.
(4) Furer, M.; Schnyder, W.; Specker, E. Normal Forms for Trivalent Graphs and Graphs of Bounded Valence. *Proc. ACM Symp. Theor. Comput.* **1983**, *No. 15*, 161-170.
(5) Garey, M. R.; Johnson, D. S. *Computers and Intractability*; Freeman: San Francisco, 1979.
(6) Hendrickson, J. B.; Toczko, A. G. Unique Numbering and Cataloguing of Melecular Structures. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 171-177.
(7) Kudo, Y.; Sasaki, S. Principle of Exhaustive Enumeration of Unique Structures Consistent with Structural Information. *J. Chem. Inf. Comput. Sci.* **1976**, *13*, 43-49.
(8) Luks, E.M. Isomorphism of Graphs of Bounded Valence Can Be Tested in Polynomial Time. *Proc. IEEE Symp. Foundat. Comput. Sci.* **1980**, *No. 21*, 42-49.
(9) Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures—A Techenique Developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5*, 107-113.
(10) Petrarca, A. E.; Lynch, M. F.; Rush, J. E. A Method for Generating Unique Computer Structural Representation of Stereoisomers. *J. Chem. Doc.* **1967**, *7*, 154-165.
(11) Randic, M. On Canonical Numbering of Atoms in a Molecule and Graph Isomorphism. *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 171-180.
(12) Sussenguth, E. H. A Graph-Theoretic Algorithm for Matching Chemical Structures. *J. Chem. Doc.* **1965**, *6*, 36-43.
(13) Wipke, W. T.; Dyott, T. M. Simulation and Evaluation of Chemical Synthesis—Computer Representation and Manipulation of Stereochemistry. *J. Am. Chem. Soc.* **1974**, *96*, 4825-4834.
(14) Wipke, W. T.; Dyott, T. M. Stereochemically Unique Naming Algorithm. *J. Am. Chem. Soc.* **1974**, *96*, 4834-4842.

# Compact Numeric Alkane Codes Derived from IUPAC Nomenclature

SCOTT DAVIDSON*

Computer Data Systems, Inc., One Curie Court, Rockville, Maryland 20850

Received January 18, 1991

A reversible binary coding scheme for storing and ordering all alkane isomers through $C_{21}$ as 32-bit integers is described. The method is derived from a modified set of IUPAC rules formerly utilized to cite side-chain names by increasing complexity (Davidson, S. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 151-5). The ordering enables construction of a bitwise tiebreaker series in which the number of bits assigned at each step is determined by the remaining choices. The compactness of the resulting codes compares favorably with previously reported graph-based codes. Manual encoding/decoding is not difficult because bit fields are small, and the logic is based upon already familiar considerations of chain sizes, lengths, and locants.

## INTRODUCTION

Many numeric codes for alkanes have been reported in the literature. However, only a few have proved to be both unique and reversible. Gordon and Kennedy[1] developed a compact, ordered integer code from combinatorial equations related to enumeration of rooted trees, ordering alkanes by increasing chain length. Decoding is obtained by iterative solution of the same equations. Knop et al.[2] designed an "N-tuple" code for alkanes consisting of a string of $N$ digits for an $N$-carbon alkane. The string represents a maximal sequence of the number of uncounted bonds at each carbon in a traverse of the longest unexplored path from a most substituted carbon, then backtracking to visit all other carbons in that path. Randić[3] has extended this code to polycyclic structures and has reviewed other codes for comparison.

A numeric code derived from standard nomenclature and ordered by size would have the advantage of being a more

* Address correspondence to 240 Manor Circle No. 2, Takoma Park, MD 20912.

0095-2338/91/1631-0417$02.50/0 