Report 78-09
Stanford -- KSL

Scientific DataLink

Exhaustive Generation of Stereoisomers for Structure Elucidation. James G. Nourse, Raymond E. Carhart, Dennis H. Smith, Carl Djerassi. 1979

HDP-78-9

Exhaustive Generation of Stereoisomers for Structure Elucidation [1]

James G. Nourse, Raymond F. Carhart, Dennis H. Smith, and Carl Djerassi

Contribution from the Departments of Chemistry, Computer Science, and Genetics, Stanford University, Stanford, CA 94305.

Abstract. An algorithm and its implementation as a computer program is described which for the first time permits the enumeration and construction of all the distinct stereoisomers possible which are consistent with a given empirical formula. The algorithm finds the stereocenters in a chemical structure, takes full account of any symmetry and produces the stereoisomers with cis/trans and R/S designations along with a canonical (unique) name. Examples of its use and a discussion of potential applications are given.

Determining the structure of an unknown compound from an empirical formula is one of the oldest problems in chemistry. A second very old problem is to determine the number of possible structures for a given empirical formula. A third problem is to generate and display these possible structures. It is these latter two problems to which this work is directed. Particular emphasis is put on stereoisomers since the most significant limitation of our current effort in computer assisted structure elucidation [2a-c] has been its lack of recognition of the stereochemical features of chemical structures. Indeed the wide application of computer methods to structure elucidation depends on the successful solution of the problem of isomer enumeration and generation.

These problems of isomer enumeration (computation of the total number) and generation (construction of all possibilities) have proved to be very difficult [2d] and it was not until 1974 that the problem of generating the possible constitutional isomers from a given empirical formula was finally solved [2a-b]. The only deficiency to this solution was that stereochemistry was not considered so that no stereoisomers were generated. The purpose of this paper is to describe an algorithm and its implementation as a computer program which can generate or enumerate the possible stereoisomers of a structure of given constitution. The algorithm makes use of the novel group theoretical and combinatorial results described in the preceeding paper [3]. The computer program has been combined with the program CONGEN(for CONstrained GENeration) [2c] which generates all constitutional isomers to yield a

program which can now generate all the possible stereoisomers from a given empirical formula.

It is important to be able to generate exhaustively all the possible isomers for a given structural problem to assure that none have been overlooked. However, the complete collection of possible isomers can be extremely large so it is important that the method of generation of these possibilities can be constrained to only a subset of possibilites if partial structures are known. The algorithm for generation of stereoisomers is capable of admitting certain constraints which reduce the number of stereoisomers generated.

1 OVERVIEW AND FLOW-DIAGRAM

When a chemist is faced with the problem of determining the number of stereoisomers of a structure of given constitution, he will probably break the problem into two parts. First, he will try to find the features of the structure which give rise to configurational stereochemistry such as asymmetrically substituted carbon atoms and double bonds. Symmetrically substituted atoms such as methylenes or gem-dimethyls will be rejected as potential sterocenters. Having found n stereocenters he will assume there are 2^n possible stereoisomers unless the structure has some overall symmetry, in which case this total may be reduced. In cases with overall symmetry, the distinct stereoisomers will problably be found by trial and error by varying the

configuration of stereocenters in turn and seeing if new stereoisomers are generated.

The algorithm to solve the problem of stereoisomer generation is summarized in the flow-diagram shown in Figure 1. Just as for the chemist, the two key problems are to determine the potential stereocenters and to correctly guage the effect of any structural symmetry. A brief overview of this algorithm is: (numbers correspond to those on Figure 1)

1. The input structure is processed to find multiply bonded atoms which are potential stereocenters (e. g. olefins, allenes) by the module Process Multiple Bonds. The symmetry group of the input structure is also determined at this stage by the module Find Symmetry Group. Structures A and B in Figure 2 illustrate this process.

2. The stereocenters are found by the module Prefilter. The input is the symmetry group and the input structure with the information about multiple bonds. The output is the set of the stereocenters and the atoms to which they are connected. The symmetry group remains unchanged here. Structure D in Figure 2 shows the stereocenters.

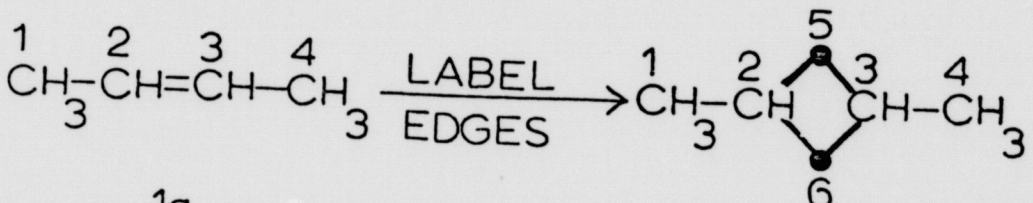
3. The Configuration Symmetry Group (CSG), which is the symmetry group represented on the configurations of the stereocenters [3] is constructed. (Figure 2, structure D)

4. Using both the set of stereocenters and the CSG the possible stereoisomers are generated by the module Generator. These are further processed to give Cis/Trans and R/S designations. The output is a list of stereoisomers.(Figure 2) Using just the CSG a count of the possible stereoisomers can be obtained by using the appropriate combinatorial equation [3]. The output is just the number of stereoisomers. The example shown in Figure 2 is discussed in greater detail later.

2 METHOD

Input Structure. The structure for which the stereoisomers are to be generated must have a definite constitution, i.e., number of atoms and bonds. The algorithm and program considers this structure as a graph [2a] in which the atoms are nodes and the bonds are edges. Each atom is uniquely numbered and while any numbering will do, there is usually one which is preferred for some reason (e.g., one with numbering corresponding to standard nomenclature). The structure is represented as a connection table which has one numbered row corresponding to each atom which consists of all the numbered atoms to which it is connected . Hydrogen atoms are not explicitly considered and are given the number 0. This is a space saving feature of the CONGEN program. Each atom also carries a designation if it is part of an aromatic system and a designation for its atom type (C,N,O, etc).

Process Multiple Bonds. The input structure is searched for atoms involved in multiple bonds which are potential stereocenters. At this stage only atoms involved in aromatic systems, triply bonded atoms, cumulenes with CH_2 ends, and rings of sp-hybridized carbon atoms are rejected as potential stereocenters. The latter hypothetical structure would be generated for an empirical formula with only carbon atoms. Next it is necessary to assign a configuration to these atoms. This is done by labelling the edges of the multiple bond with ficticious bivalent nodes 1a-b. These nodes are given numbers



1a 1b