Computer-Assisted Elucidation of Molecular Structure with Stereochemistry[1]

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

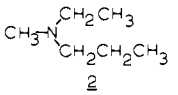
Contribution from the Department of Chemistry, Stanford University, Stanford, California 94305. *Received April 3, 1980*

Abstract: A computer-based method for constrained generation of configurational stereoisomers is described. A stereoisomer generator capable of exhaustive and irredundant generation of configurational stereoisomers is constrained by atom type, substructures bearing configuration designations, and molecular symmetry to generate only those stereoisomers which are allowed by available data. This method, when coupled to a generator of constitutional isomers, provides the complete set of stereoisomers which represent potential solutions to a structure elucidation problem.

In recent years there has been considerable work on applications of computer methods to organic chemical structure elucidation.[2] One approach to computer-assisted structure elucidation has been the CONGEN program (for CONstrained GENeration of isomers).[3] This program was developed with the capability to generate, exhaustively and irredundantly (i.e., without duplicates), all possible constitutional (i.e., bond-connective) isomers for a given molecular formula.[4] Subsequent developments allowed this program to be constrained to give only those constitutional isomers consistent with partial substructural information.[3] These later developments resulted in a program useful in routine chemical structure elucidation.[5] The major deficiency of this program was the lack of any knowledge about the stereochemistry of the structures considered. Recently, this problem was addressed by the development of a program which allows for the first time the exhaustive and irredundant generation of all the configurational stereoisomers (i.e., those differing in orientations around double bonds and chiral centers) consistent with a molecular formula.[6]

It is the purpose of the present paper to describe developments of the stereoisomer generator which permit the *constrained* generation of configurational stereoisomers. These developments provide for the first time the capability, in a computer program, for use of both constitutional and configurational structural constraints to expand efficiently a molecular formula into all possible structures, including stereoisomers, allowed by the constraints. For example, we noted previously that there are seven theoretically possible configurational stereoisomers of twistane **1**.[6] Of these seven, there are only two (the *dl* pair of **1**) which are "reasonable" as they have all-cis bridges. The purpose of the present work is to demonstrate a method, realized in a computer program, which allows an investigator to eliminate "unreasonable" stereoisomers, such as the remaining five for twistane, in an ongoing structure elucidation problem.

1



2

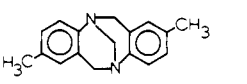
3

This paper is divided into four principal sections. The first section discusses some possible constraints on a structure elucidation problem which depend on the configurations of stereocenters (double bonds and chiral centers). The second section describes the computer program and the types of constraints which can be applied to a generation of stereoisomers with an example of a structure determination involving stereochemistry. The third section presents some results and the fourth discusses some details of the algorithms employed.

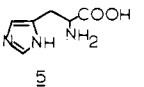
I. Configurational Stereochemical Constraints

A number of constraints on a chemical structure elucidation problem depend on knowledge of the configurations of stereocenters. Three general classes of constraints will be discussed here. These are constraints dependent on: (1) the type of atom, (2) the presence or absence of substructures with definite stereocenter configurations, and (3) symmetry (e.g., equivalent atoms or substructures or overall chirality). The implementation of these constraints in a computer program will be discussed in Section II.

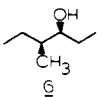
1. Atom Type Constraints. Probably the simplest configurational constraint is based on the fact that only certain tri- and tetravalent atoms are capable of being stereocenters. (Our current program is not capable of considering stereocenters involving atoms of valence greater than four.) In common organic structure elucidation problems, it is always assumed that appropriately substituted carbon atoms can be stereocenters. However, there is often some question whether nitrogen atoms can be stereocenters. Thus, the trisubstituted nitrogen atoms in **2** and **3** would freely invert and would not normally be considered as stereocenters at room temperature. However, the optically resolvable compound **4** owes its chirality to the noninverting, trisubstituted nitrogen atoms (at room temperature). Histidine (**5**) contains one nitrogen in a definite configuration (in the ring double bond), one which can freely invert (the secondary amine) and one which is not a stereocenter (the primary amine).



4



5

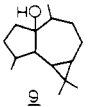


6

2. Substructural Constraints with Configurations. Many constraints on a structure elucidation problem can be expressed by the required presence or absence of substructural fragments (partial structures) which contain stereocenters with definite configurations (i.e., cis or trans double bonds or chiral centers). These constraints are often inferred from various physical or chemical data. Thus it might be known that a structure contains the substructure **6** in which the two stereocenters are known to be in the configurations shown.

7

8



9

Substructure **6** indicates one of the four possible absolute configurations but might also represent one of the two possible relative configurations. The former might have come from a

(1) Applications of Artificial Intelligence for Chemical Inference, Part XXXIII. For part XXXII, see Lavanchy, A.; Varkony, T.; Smith, D. H.; Gray, N. A. B.; White, W. C.; Carhart, R. E.; Buchanan, B. G.; Djerassi, C. *Org. Mass Spectrom.*, in press (1980). We thank the National Institutes of Health for their generous financial support (RR-0612) and for their support of the SUMEX computer resource on which the programs were developed and are made available to outside collaborators (RR-00785).

(2) (a) Smith, D. H., Ed., "Computer-Assisted Structure Elucidation"; American Chemical Society: Washington, D.C., 1977. (b) Djerassi, C.; Smith, D. H.; Varkony, T. H. *Naturwissenschaften* **1979**, *66*, 9.

(3) Carhart, R. E.; Smith, D. H.; Brown, H.; Djerassi, C. *J. Am. Chem. Soc.* **1975**, *97*, 5755.

(4) Masinter, L. M.; Sridharan, N. S.; Lederberg, J.; Smith, D. H. *J. Am. Chem. Soc.* **1974**, *96*, 7702.

(5) Cheer, C.; Smith, D. H.; Djerassi, C.; Tursch, B.; Braekman, J. C.; Daloze, D. *Tetrahedron* **1976**, *32*, 1807.

(6) Nourse, J. G.; Carhart, R. E.; Smith, D. H.; Djerassi, C. *J. Am. Chem. Soc.* **1979**, *101*, 1216.

0002-7863/80/1502-6289$01.00/0

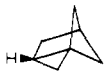
degradation experiment yielding a structure of known configuration whereas relative configuration might be derived from interpretation of vicinal proton NMR coupling constants in a structure possessing **6** in a rigid conformation. For the present purposes, it is important to note that the stereochemical constraint is in the form of a substructure with definite stereocenter configurations. However, in our current substructure representation, there is no information about the conformation (i.e., torsional angles around single bonds). Thus the proton NMR data (the vicinal couplings) which yield information on conformational relationships would have to be interpreted to give a substructure with only configurational information.

If in a structure elucidation problem it were known that substructure **6** was present at least once, then all candidate constitutional structures which lacked substructure **6** (even without any configurations specified) would be eliminated as well as all those stereoisomers of the remaining constitutional structures which did not have at least one occurrence of the substructure **6** with the stereocenter configurations shown. At least some of the theoretically possible stereoisomers of any remaining constitutional isomers would survive this condition.

In addition, *undesired* stereoisomers can be suppressed by such substructural constraints. For example, if it were known that substructure **6** could *not* be present, then *only* those stereoisomers which contained **6** as a substructure with the stereocenter configurations shown would be eliminated. Constitutional isomers which contained no occurrence of **6** (without any configurations specified) would be allowed as well as any stereoisomers which contained a diastereomer (i.e., only differing in one or more stereocenter configurations) of **6**.

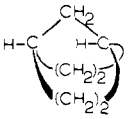
One particularly useful substructural constraint can be derived as a generalization of Bredt's rule. This rule states that structures such as **7** with double bonds at bridgeheads in bicyclic systems are disallowed because of excessive strain. The chains of atoms connecting the bridgeheads can of course vary in length and in larger structures, bridgehead double bonds are known. Smaller structures with bridgehead double bonds are also known and tend to be transient structures or stable only at low temperatures.[7] In common structure elucidation problems, particularly of natural products, such Bredt's rule violators are unlikely to be observed. All structures which violate Bredt's rule contain as a substructure a ring with a trans double bond. This feature has been proposed as the most important in deciding which structures violating Bredt's rule are likely to be observed.[8] The point here is that this rule can be stated as a substructural constraint: no structures are permitted which contain a small ring (usually less than eight members) with a trans double bond. This substructure can be represented by **8**, where X represents any atom type, and the chain connecting the trans positions on the double bond is of length one to five. In a structure elucidation problem, this rule will eliminate some of the stereoisomers (i.e., the trans isomers) of structures with a cyclic double bond and all of the stereoisomers of structures such as **7** since any stereoisomer of **7** will contain a ring with a trans double bond. It is important to note here that a constraint expressed as a substructure with stereocenter configurations on some atoms can eliminate *all* the stereoisomers of some structures. This feature can be particularly useful in structure elucidation problems which involve many structural possibilities (see section III). For example, in the elucidation of the structure of (+)-palustrol **9** the CONGEN program was used as an aid.[5] After the interpretation of all spectral and reaction data, there remained 20 constitutional structural possibilities. Of these, about one third could have been rejected on the basis that their respective reaction products were Bredt's rule violators.

Other substructures can be used in a similar manner to eliminate candidate structures in a structure elucidation problem. Consider structure **10** which differs from **7** by the substitution of a three-membered ring in **10** for the double bond in **7**. All stereoisomers of **10** will contain the trans-fused ring system **11** as a substructure. Trans-fused rings are highly strained when one or both of the rings are small.[9] Another unfavorable substructure is the inverted or "in-out" bicyclic structure **12**. These have been isolated only when one of the chains connecting the two trivalent atoms is quite long (greater than seven atoms).[10] It is interesting to note that all three of these substructures (trans cyclic double bonds (**8**), trans-fused rings (**11**) and inverted bicyclics (**12**)) are related in that they all have two trivalent atoms connected by three "links" of one or more bonds.



10

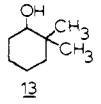
11



12

3. Symmetry Constraints. Some information used in a structure elucidation problem is based on symmetry properties of the unknown structure under investigation. These include observations about numbers of equivalent atoms or equivalent substructures and the chirality of the molecule. Knowledge of the configurations of stereocenters is crucial to the interpretation of these observations.

Consider the problem of equivalent atoms in a structure. A structure's constitution does not by itself determine which atoms are equivalent in any stereoisomer of the structure. As an example consider the dimethyl cyclohexanol **13**. The two methyl groups are constitutionally equivalent yet they are inequivalent in both stereoisomers of **13**. More specifically, the symmetry of both stereoisomers of **13** is lower than the constitutional symmetry of **13**. The symmetry group of both stereoisomers of **13** includes only the identity operation while the constitutional symmetry group (i.e., the symmetry considering only atom–atom connectivity) of **13** includes a permutation which exchanges the two methyl groups. In general, the symmetry group of a stereoisomer will be the same size or smaller than the constitutional symmetry group of the structure. Therefore, in order to utilize observations about equivalent atoms in a structure (for example, from NMR data), it is necessary to know as much as possible about the stereochemistry of the structure. Information about equivalent atoms in stereoisomers can be derived from the symmetry group of the stereoisomer,[11] as is discussed in sections III and IV.



13

Knowing the configurations of stereocenters and double bonds is an improvement over knowing just the constitution but this still does not allow a complete specification of the symmetry of a structure in all cases. For example, the two methyl groups are equivalent in **14a** when the configuration symmetry is considered,[11] yet they are inequivalent in the two conformations **14b** and **14c**. In other words, the symmetry of conformations can be still lower than the configuration symmetry. In many cases, including this example, there will be an equilibration of conformations which results in an observed "average" symmetry. However, there will be many cases in which knowledge of configuration symmetry is inadequate to predict observed equivalent atoms. The point here is that knowledge about stereocenter configurations is crucial to the interpretation of symmetry information in a structure elucidation problem but it does not always provide a complete specification. Knowing stereocenter configurations does, however, represent an improvement over knowing just constitution in problems of this kind.

Making use of observations of equivalent substructures (i.e., those which consist of more than one nonhydrogen atom) represents a slightly different problem. Simply knowing the equivalent

(7) Martella, D. J.; Jones, M., Jr.; Schleyer, P. v. R.; Maier, W. F. *J. Am. Chem. Soc.* **1979**, *101*, 7637.
(8) Wiseman, J. R. *J. Am. Chem. Soc.* **1967**, *89*, 5966.
(9) Liebman, J. F.; Greenberg, A. *Chem. Rev.* **1976**, *76*, 311.
(10) (a) Gassman, P. G.; Thummel, R. P. *J. Am. Chem. Soc.* **1972**, *94*, 7182; (b) Park, C. H.; Simmons, H. E. *J. Am. Chem. Soc.* **1972**, *94*, 7184.
(11) Nourse, J. G. *J. Am. Chem. Soc.* **1979**, *101*, 1210.

Table I. STEREO Program Commands and Their Purposes

command	purpose
COUNTST	count the theoretical number of stereoisomers possible for an individual constitutional isomer or a list of constitutional isomers of identical molecular formula
GENST	generate the stereoisomers, for an individual constitutional isomer or a list of constitutional isomers, consistent with a set of constraints
PRUNST	prune a list of stereoisomers to eliminate those not consistent with a set of constraints
DRAW	draw an individual structure showing its constitution (bond-bond connectivity) only
SDRAW	draw an individual structure or part of a structure showing the configurations of stereocenters
SHOWST	show information such as stereocenter configurations or equivalent atoms for a stereoisomer
DONE	return to CONGEN

atoms in a structure does not allow one to predict with certainty the equivalent substructures. Predictions of equivalent substructures of arbitrary complexity require a more detailed analysis of the symmetry group of the stereoisomer than simple determination of the sets of equivalent atoms. As an example consider the polycyclic structure **15**. All the atoms in the six-membered ring are equivalent but all the bonds (i.e., two atom substructures) are not (there are two equivalent sets).[12] In order to predict the equivalent bonds (and, in general, arbitrary substructures) it is necessary to make use of the symmetry group of the stereoisomer since the equivalence of substructures can be easily ascertained by investigation of the results of the permutations in the group (see section IV).

Another property which depends on a structure's symmetry is chirality. In a structure elucidation problem, it can often be determined if the structure under investigation is chiral (either a single enantiomer or a *dl* pair). The chirality (i.e., either chiral or achiral) of a stereoisomer can be predicted if the configurations of the various stereocenters and the symmetry group are known.[11] Again it is possible for structures to be chiral because of conformational inflexibility (i.e., rotation about single bonds)[13] so that specification of the stereocenter configurations represents a necessary but not sufficient condition for prediction of this property.

II. The STEREO Program and a Sample Application

The CONGEN program without any stereochemical features has been described previously[3] and its use exemplified.[5] The STEREO program will now be described together with an example. Recently, Goldstein et al.[19] used CONGEN as an aid to solving the structure of an unknown compound, as described below. Manual analysis of the structural possibilities was required by them in order to apply stereochemical constraints. In this section we describe the basic functions of the STEREO program and use Goldstein et al.'s structure[19] as an example of how stereochemical constraints can now be applied automatically as an aid to structure elucidation.

(12) The general result follows mathematically from the fact that all transitive permutation groups are not doubly transitive. See Hall, M., Jr., "The Theory of Groups"; Macmillan: New York, 1959.

(13) Mislow, K. "Introduction to Stereochemistry"; W. A. Benjamin: New York, 1966; pp 80-81.

(14) EDITSTRUC is a teletype-oriented "structure editor" based on concepts originally presented by Feldmann (Feldmann, R. J., in "Computer Representation and Manipulation of Chemical Information", Wipke, W. T., Heller, S. R., Feldmann, R. J., Hyde, E., Eds.; Wiley-Interscience, New York, 1974; p 55). EDITSTRUC, recently extended with commands for visualizing and assigning configurations to chiral centers and double bonds, allows creation of substructures of arbitrary complexity within the restrictions of normal chemical valence and tetrahedral stereocenters.

(15) Carhart, R. E. *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 82.

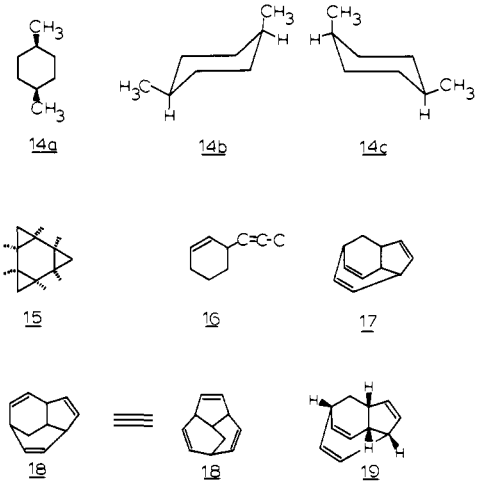
(16) (a) Whitlock, H. W.; Siefken, M. W. *J. Am. Chem. Soc.* **1968**, *90*, 4929; (b) Varkony, T. H.; Carhart, R. E.; Smith, D. H. In "Computer-Assisted Organic Synthesis", Wipke, W. T.; Howe, W, J., Eds.; ACS Symposium Series, No. 61, 1977, p 188.

(17) Paukstelis, J. V.; Kao, J. *J. Am. Chem. Soc.* **1972**, *94*, 4783.

(18) Scott, L. T.; Jones, M., Jr. *Chem. Rev.* **1972**, *72*, 181.

(19) Goldstein, M. J.; Nomura, Y.; Takeuchi, Y.; Tomoda, S. *J. Am. Chem. Soc.* **1978**, *100*, 4899.

Goldstein et al.[19] synthesized a compound of molecular formula $C_{11}H_{12}$ in a study of $C_{11}H_{11}^-$ anion stability. By using proton and carbon NMR they were able to establish that there were three (1,2)-disubstituted double bonds and no quaternary carbons. This information was given to CONGEN in the form of the molecular formula, the three disubstituted double bonds ("Superatoms"), and the restriction ("constraint") of no quaternary carbons in the manner described previously.[3,5] This leads to 91 possible constitutional isomers which include tricyclic structures and monocyclic structures connected to bicyclic structures by chains of bonds ("isthmuses"). Goldstein et al. were able to establish the presence of the substructure, **16**, by extensive proton–proton decoupling. This substructure, **16**, includes four olefinic and five aliphatic carbon atoms. Imposition of this constraint on the list of 91 structures gives just two constitutional possibilities, **17** and **18**.



This example is typical in that use of CONGEN in a structure elucidation problem will result in a list of candidate constitutional structures stored on a disk file. Each structure has a definite constitution and is stored as a compact representation of its connection table. It is at this stage in a structure elucidation problem that the STEREO program (actually implemented as a module of CONGEN) is used. The STEREO program is designed to allow interactive control over generation of stereoisomers using the three types of constraints discussed above. A summary of the command options is given in Table I. These will be discussed in detail and, where appropriate, illustrated with an example.

The STEREO program can be operated in two different modes when called from CONGEN. In one mode, the program will process an entire list of constitutional structures kept on a disk file. This will be termed the "list" mode. In the other mode, the program will process a single constitutional structure and permit a very detailed look at the stereoisomers of that single constitutional structure. This will be termed the "individual" mode.

A. Stereoisomer Counting. It is possible simply to count the number of stereoisomers for either an individual constitutional isomer or for a list (COUNTST command, Table I). This counting is independent of the generation of stereoisomers. The appropriate combinatorial formula has been described previously.[11] This gives the total number of stereoisomers that would be obtained for an unconstrained generation and can be done much faster than the actual generation. This is a useful step to investigate the scope of a problem.

For the example, stereoisomer counting yields 192 possible stereoisomers for the constitutional structures **17** and **18**.

B. Stereoisomer Generation. The STEREO program can generate the possible stereoisomers for a constitutional structure with or without constraints (GENST command, Table I). The capability for unconstrained generation has been described previously and leads to a list, in a compact representation,[6] of stereoisomers for each constitutional isomer. This is an exhaustive and irredundant generation. The current program will, on command, do this in individual or list mode and, in list mode, store the resulting stereoisomers on a disk file. In the list mode, the constitutional

Table II. Available Constraints for Stereoisomer Generation and Pruning in the STEREO Program

constraint	purpose
ATOMTYPEST	designate types of atoms which are (not) capable of being stereocenters
CHIRALITY	generate or keep only those stereoisomers which are (a)chiral
EQUIVPAT	generate or keep only those stereoisomers which have a designated pattern of equivalent atoms or substructures
SUBSTRUCTURE	generate or keep only those stereoisomers which have a designated number of occurrences of some substructure with stereocenter designations
UNSYMMETRICAL	generate or keep only those stereoisomers which have no equivalent atoms

isomers are processed one by one so that lists of hundreds or thousands can be processed with modest memory requirements.

Unconstrained generation for the example yields 192 possible stereoisomers for the constitutional structures **17** and **18**, all saved on a disk file for subsequent application of constraints. Alternatively, available constraints can be applied during the course of stereoisomer generation.

Constrained stereoisomer generation is accomplished by following the request for generation (GENST command, Table I) with requests for the possible types of constraints. These are summarized in Table II and will be discussed in sequence.

1. Atom Type Constraints. The investigator can request (ATOMTYPEST constraint, Table II) that any type of atom be a stereocenter or nonstereocenter. The default setting is that all tri- and tetravalent atoms are potential stereocenters in all environments and the program will process structures with this assumption unless a different request is made. If a type of atom is designated a nonstereocenter then unless atoms of that type appear at bridgeheads or in double bonds in rings, the program treats them as nonstereocenters. This corresponds to the behavior of nitrogen atoms in such environments (e.g., **2–5**) and for common organic structure elucidation problems, nitrogen atoms are the only ones likely to be ambiguous in this respect. The example contains only carbon atoms as potential stereocenters, so this constraint is not modified from the default setting.

2. Substructural Constraints. The investigator can constrain the generation based on the presence or absence of substructures (SUBSTRUCTURE constraint, Table II). The substructure to be used must be defined using the CONGEN structure editor which has been augmented with commands that allow definition of stereocenter configurations.[14] The investigator can then specify the number of occurrences of the substructure desired in all generated stereoisomers. The choice of number of occurrences is one of the following: (a) none, (b) at least *n*, (c) exactly *n*, (d) at most *n*, or (e) a range from *m* to *n*. In addition there is a conditional choice: *if* the substructure occurs (as a constitutional substructure), the configurations of the stereocenters in all generated stereoisomers *must* be the same as in the substructure; if the substructure does *not* occur as a constitutional substructure, then stereoisomer generation proceeds normally for this constitutional isomer. This conditional choice represents a more efficient application of a constraint of this kind because it is implemented before stereoisomer generation (prospectively, see section IV). Thus, if a substructure has many stereocenters and one stereoisomer is desired, it is easier to require the presence of that one stereoisomer than to forbid all its diastereomers. For example, this choice could be used to require any occurrences of a steroid nucleus to possess specified stereocenter configurations.

The final choice is whether the stereocenter configurations in the substructure are to be considered as absolute or relative configurations. If absolute configurations are desired, then the substructure will match only substructures with the same absolute configuration. If relative configurations are desired, then the substructure will also match substructures with all stereocenters in the enantiomeric configurations. The default choice is relative configurations and results in generation of enantiomeric pairs of stereoisomers.

Substructures such as general rings with trans double bonds or the variants of the bicyclic structures **11** and **12** can be conveniently defined using the linknode option of CONGEN's structure editor.[14] A linknode is simply a variable length chain of atoms. Thus, a single structure defined with a linknode actually represents several structures with different length chains. To forbid stereoisomers with rings under some size (e.g., eight) containing trans double bonds, it is only necessary to define a single substructure with a trans double bond in a ring which also contains a linknode of variable size (e.g., length one to five (**8**)). This will forbid all stereoisomers with trans double bonds in rings of less than 8 atoms.

For the example, we performed a constrained stereoisomer generation under the constraints of no trans double bonds in small rings (**8**) and no inverted, "in-out" bicyclics (**12**), of the form [*a.b.c*] where $a < 3$, $b < 4$, $c < 5$ and requested relative configurations. This generation yielded a total of six possible stereoisomers (out of the 192 theoretically possible). Remaining are one *dl* pair for constitutional isomer **17**, **19** and its mirror image, and one *dl* pair, **20** and its mirror image and two meso stereoisomers, **21** and **22**, for the constitutional isomer **18**.

20 **21** **22**

3. Symmetry Constraints. The investigator can require generated stereoisomers to have equivalent atoms or substructures (EQUIVPAT constraint, Table II). The program is told the name of the atom or substructure, the number of equivalent sets, and the number of atoms or substructures in (the size of) each set. For example, in a structure elucidation problem, it may be known (perhaps from NMR data) that there are two sets of equivalent methyl groups, one of size two and one of size three. Providing this information to the program will result only in generation of those stereoisomers which have at least one set of at least two equivalent methyls and another of at least three equivalent methyls. This is interpreted as a "least only" constraint since the apparent symmetry of the configurational stereoisomer can be greater than in the actual conformation of the molecule (see Section I). Thus, if there were five equivalent methyls predicted for some configurational stereoisomer, there might be sets of two and three equivalent methyls observed in some conformation. Hence, the program will keep stereoisomers with more symmetry than that required by the observed sets of equivalent atoms or substructures. This does not generally result in excessive numbers of stereoisomers since most chemical structures have little or no symmetry. In fact, any observation of symmetry is a powerful constraint since this will eliminate most possible structures in a typical problem.

The investigator may also require that there be *no* sets of equivalent atoms of any kind (UNSYMMETRICAL constraint, Table II). This is a much simpler constraint to apply and is handled separately. However, to use this constraint safely, there must be an assumption made that the observed symmetry of any equilibrating conformations (at the conditions of an experiment) must reflect the overall symmetry of the configurational stereoisomer. If individual conformations of an equilibrating set are observed (i.e., the barrier to interconversion is too high), this assumption may result in the discarding of otherwise legal candidate structures.

For the example, Goldstein et al. noted that the structure was unsymmetrical and it is not likely that these structures will be conformationally mobile. Using the constraint of only unsymmetrical structures (with the PRUNST command, below) eliminates the two meso stereoisomers of **18**, **21**, and **22**, and leaves just the two *dl* pairs, **19** and **20** and their mirror images, as possibilities. The resulting stereoisomer of **18**, **20**, would likely be very strained. It contains an inverted ("in-out") bicyclo [5.2.1] substructure which is larger than the substructure used as a constraint. This possibility can probably be safely eliminated leaving only the *dl* pair of stereoisomers of **17**, **19** and its mirror image, which was favored by Goldstein et al.[19]

The investigator can also require that the generated structures be all chiral or all achiral (CHIRALITY constraint, Table II). Again this is based on stereocenter configurations and conformations are not considered; hence, discretion must be used here as well.

C. Stereoisomer Pruning. Lists of stereoisomers generated previously may be tested using any of the above constraints except ATOMTYPEST (PRUNST command, Table I). Stereoisomers which are disallowed are removed, or "pruned" from the list. The new list is again stored on a disk file for subsequent application of additional constraints.

For the example, the constraints used during stereoisomer generation could alternatively have been deferred for application to the entire set of 192 isomers from the unconstrained generation. This is less efficient, however. In general, the STEREO program is used most efficiently by performing stereoisomer generation with available constraints. This is followed by examination of results and perhaps collection of additional data, using the PRUNST command at a later time to reduce further the list of possible stereoisomers. In fact, the application of the constraint specifying unsymmetrical structures (above) was done as a pruning of the set of six stereoisomers from the constrained generation.

D. Character-Based Drawings. When using the STEREO program for an individual constitutional structure, it is possible to get character-based drawings of the structure (DRAW command, Table I) or any of its stereoisomers (SDRAW command, Table I). This is a crucial feature for an interactive, computer-assisted structure elucidation program because there must be some way to visualize the results of the computations. The drawings which show only the constitution (bond–bond connectivity) have been described previously along with the program used to create them.[15] In our program, we can obtain a variety of drawings which show the stereocenter configurations of stereoisomers. These include Fischer projections of individual stereocenters, Fischer projections of chains of atoms, Newman projections along bonds, and ring-based drawings. Some examples are presented in Figure 1.

The ring-based drawings use templates to lay out ring systems and show stereocenter configurations and ring fusions clearly. Bridged structures and complicated polycyclics are less clearly presented. There are several options which allow the user to orient the drawing. These are all character-based drawings which therefore sacrifice the vastly better drawings possible on graphics terminals. However, many stereoisomers can be visualized effectively with these drawings, thereby making it possible for any investigator with the simplest terminal (e.g., a teletype) to use CONGEN and STEREO.

E. Miscellaneous Information. It is also possible to request the STEREO program to show a variety of information about the stereoisomers of an individual constitutional structure (SHOWST command, Table I). Upon request, the program will give the configurations of all stereocenters of a stereoisomer, the symmetry group of that stereoisomer, the symmetry equivalent atoms of that stereoisomer, or a list of all the stereoisomers generated for a particular constitutional structure. These features make use of the representation of stereoisomers discussed previously.[6] For example, a request for the symmetry equivalent atoms of the stereoisomers of **17** and **18** would reveal that **19** and **20** have no equivalent atoms, while **21** and **22** have four pairs of equivalent atoms.

III. Results and Discussion

Often in a structure elucidation problem, a chemist will determine the constitution completely or nearly completely before worrying about the stereochemistry of the structure. An example is the structure elucidation of **19** discussed in section II. For many structures, this approach makes sense since early in a structure elucidation problem there may remain hundreds or thousands of candidate constitutional structures. At this point it is of little interest that there are perhaps 100 stereoisomers for each candidate. However, as has been pointed out earlier, it is possible that stereochemical constraints can eliminate all stereoisomers of a constitutional structure, thereby eliminating the constitutional structure from further consideration. The fact that this possibility

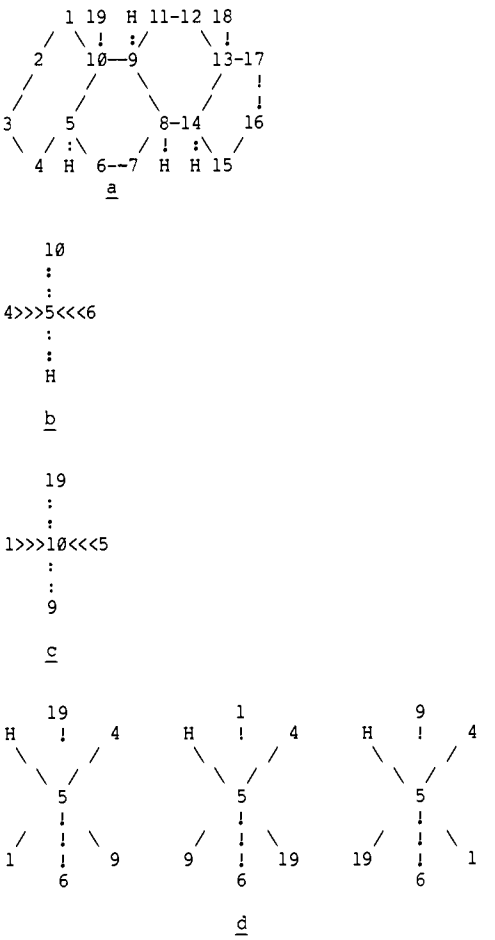


Figure 1. Examples of teletype drawings of the androstane steroid skeleton with the usual stereocenter configurations. (a) Ring-based drawing of the complete skeleton. For substituents, ":" indicates a substituent below the plane of the ring, "!" indicates a substituent above the plane of the ring; (b) Fischer projection of stereocenter 5. These projections follow the usual convention of horizontal substituents above the plane of the paper, vertical substituents below the plane; (c) Fischer projection of stereocenter 10; (d) Newman projections along the C-5,10 bond. Since there is no information about conformation, the program presents three hypothetical rotamers. The middle rotamer corresponds to the correct conformation of the A,B ring juncture in androstane.

can reduce the scope of a structure elucidation problem is perhaps not fully appreciated and will be demonstrated in this section.

The CONGEN and STEREO programs have been used to generate the possible configurational stereoisomers under various constraints for several molecular formulas summarized in Table III. It should be emphasized that only the numbers of isomers are given here, although the program has actually constructed all the isomers and, using commands discussed in section II, could be used to examine the stereoisomers in detail.

The first row in Table III gives the total number of constitutional isomers possible for each of the molecular formulas shown. However, for $C_{10}H_{16}$ the total is constrained to only those isomers with no multiple bonds, methyls, or three- or four-membered rings. These are isomers of adamantane and this system has been the subject of several mechanistic studies.[16] The second row gives the unconstrained total number of configurational stereoisomers for each case. In the third row are the numbers of constitutional isomers which have no triple bonds or cumulated double bonds in rings under eight atoms. This choice of ring sizes corresponds to the observed stabilities of such structures.[9] However, it must be emphasized that any such size can be chosen using the programs; the present choices were made only for illustrative purposes.

The fourth and fifth rows in Table III give the numbers of constitutional and configurational stereoisomers which satisfy four stereochemical constraints. (Note that the results presented in rows four and five are derived in one operation from implemen-

Table III. Results of Constrained and Unconstrained Stereoisomer Generation for Several Molecular Formulas

isomer type and constraint used	C_6H_6	C_6H_8	$(CH)_8$[a]	$(CH)_{10}$[a]	$C_{10}H_{16}$[b]
(1) constitutional total	217	159	20	91	21
(2) stereoisomer total	961	514	476	10 326	162
(3) constitutional, no triple bonds or cumulenes in rings under 8 atoms	158	139	20	91	21
(4) constitutional of (3) with no cyclic trans double bonds, fusions or inverted bicyclics	72	111	17	71	17
(5) stereoisomers of above (4)	79	147	33	253	34
(6) constitutional with at least 2 equivalent CH	33	35	17	61	11
(7) stereoisomers of (6)	34	43	33	187	20
(8) constitutional with at least 3 equivalent CH	5	4	9	26	4
(9) stereoisomers of (8)	5	5	15	47	7
(10) constitutional with at least 4 equivalent CH	4	3	9	24	3
(11) stereoisomers of (10)	4	4	15	44	5

[a] Each carbon bears exactly one hydrogen atom. [b] Constrained generation; see text and ref 16.

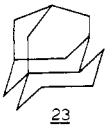
tation of these constraints.) These are:

(1) There are no trans double bonds in rings of less than eight atoms, i.e., **8** where X represents any atom. This corresponds to observations in such systems.[9]

(2) There are no three-membered rings in trans fusions to rings under six atoms (bicyclo [*n*.1.0] systems where $n < 4$). These are structures like **11**. Again this corresponds to observed results as trans fused [4.1.0] structures are known.[17]

(3) There are no trans fused [2.2.0] structures.

(4) There are no inverted ("in-out") bridged bicyclic systems of the form [*a.b.c*] where $a < 3$, $b < 4$, $c < 5$. These are structures like **12**. The only known, open (i.e. no bonds between the chains connecting the two trivalent atoms) systems of this type are much larger,[10] but these have not received as yet a systematic study. An inverted bicyclic substructure can appear in structures with other bonds and an example is the inverted bicyclo [3.3.5] substructure in the diamond lattice structure **23**.

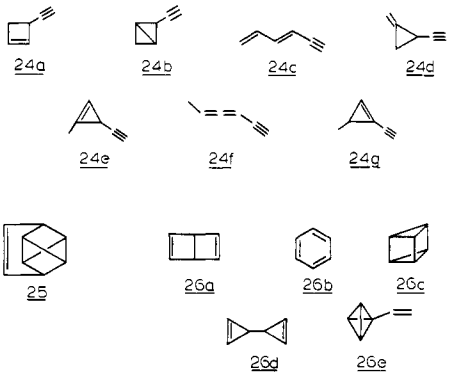


23

Comparison of rows three and four of Table III shows the number of *constitutional* isomers for which there is *no* stereoisomer that survives these four constraints. In the case of C_6H_6 (e.g., benzene) more than half the constitutional isomers fail these constraints along with more than 90% of the stereoisomers (compare rows two and five). The seven structures with more than one stereoisomer (compare rows four and five, column 1) are **24a–24g**. It is exceedingly unlikely that any chemist could have produced these seven structures without the aid of a computer. The key feature of this molecular formula which leads to this reduction is that there are four degrees of unsaturation (rings plus multiple bonds). While structural studies on structures with molecular formula C_6H_6 are relatively rare, studies with molecular formulas containing four or more degrees of unsaturation are common. These stereochemical constraints would be useful in structure elucidation studies in such cases even if only the constitution of the structure is being sought.

The possible C_6H_8 constitutional isomers have been discussed previously.[4] A comparison of rows one, three, and four demonstrates that a significant number of the possible constitutional isomers can be rejected on the basis of these simple stereochemical constraints. The $(CH)_8$ and $(CH)_{10}$ (all carbons have exactly one attached hydrogen) systems have been favorites of organic chemists.[18] In such heavily unsaturated systems, most of the theoretically possible stereoisomers are disallowed (well over 90%, compare rows two and five in Table III). There are also constitutional isomers of $(CH)_8$ and $(CH)_{10}$ which are eliminated by these constraints (compare rows three and four, columns three and four, Table III). These are not Bredt's rule violators since none of these structures can have quaternary carbons. Rather, the eliminated constitutional isomers are structures such as **25** which violate the conditions of no small, trans-fused rings or no in–out bicyclics *several* times and at least once in all stereoisomeric configurations. Formulation of a single rule to eliminate all such structures might be difficult but is easily accomplished by not allowing any occurrences of **11** and **12**.

The final rows (six–eleven) of Table III give the numbers of constitutional and stereoisomers which have at least 2, 3, or 4 equivalent methines (CH groups). The structures with molecular formula C_6H_6 which have three or more equivalent methines are **26a–26e** (column one, row eight, Table III). These results indicate the dramatic reduction in number of possible stereoisomers as a consequence of observation of equivalent atoms or substructures.



24a 24b 24c 24d 24e 24f 24g 25 26a 26b 26c 26d 26e

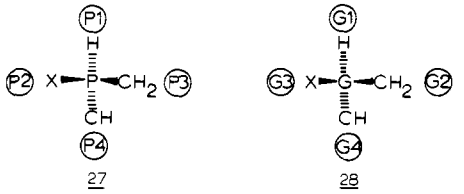
IV. Features of Algorithms

Features of the algorithms used in three parts of the STEREO program will be discussed. These are: (1) the stereochemical graph matching used in finding substructures, (2) the equivalent atom and substructure searching, and (3) the stereochemical constraint application to stereoisomer generation.

1. Stereochemical Graph Matching. Substructural constraints involving stereocenter configurations make use of a graph matching step. The substructure to be matched and the structure in which a match is sought are represented as graphs augmented with designations for the stereocenters which have been assigned configurations.[6] The graph matching step finds all distinct matchings of the substructure within the structure. A distinct graph matching is one in which the edges (bonds), and any isolated nodes (atoms), mapped are unique. (The matching of substructures with linknodes (chains of atoms of variable length) is somewhat different as only one such chain need be found). Matchings are distinct when all edges (bonds) are distinct including the bonds to the linknode (chain of atoms). The atoms in the structure which are part of the variable length chain are not considered in establishing distinct mappings except that no other part of the substructure or any other linknode can match them.

The first part of the procedure is to search for matchings of the constitutional representations of substructure and structure. Once a constitutional matching, or "mapping", of the substructure is found, the stereocenter configurations are compared for agreement. Consider the mapping of the stereocenter, P, in

substructure **27** with neighbors numbered (P1, P2, P3, P4) and configuration (0) to stereocenter, G, in the (partial) structure **28** with neighbors numbered (G1, G2, G3, G4) and configuration (1). The configuration designations 0 and 1 have been discussed earlier;[6] they are unambiguously assigned by the program. While the individual configuration designations depend on the numbering of the attached atoms, the overall comparison is invariant to the numberings because the following parity check is made.



27 28

The parity of the mapping of the neighbors is first determined. This is the parity of the permutation of the neighbors defined by the numerical orderings of the neighbors in the substructure and structure. In the above example P1 goes to G1, P2 goes to G3, P3 goes to G2, and P4 goes to G4. The overall permutation is (1) (23) (4) and is therefore an odd permutation.[6] This means that, in order to match, the stereocenter configurations of the substructure and structure must be opposite. This is the case in the example above so that the substructure matches the structure stereochemically as well as constitutionally. The atoms involved in double bonds can be formally considered in the same way as chiral stereocenters[6,11] so that a similar procedure works for them as well.

2. Equivalent Substructure Search. The algorithms to find equivalent atoms and substructures in stereoisomers make use of the symmetry groups of the structure and its stereoisomers. First, the constitutional symmetry group is found. This is done by first scoring the nodes based on extended connectivity and then doing isomorphism checks. The algorithms used are modified versions of those described previously in greater detail.[20] This symmetry group (of the constitutional isomer) is then used in subsequent procedures (below) to find either equivalent atoms or equivalent substructures.

Since the constitutional symmetry group is represented by permutations of atoms, finding possible sets of equivalent atoms in stereoisomers is not difficult. In the CONGEN program, hydrogen atoms are not explicitly represented so that in this context "atoms" include the number of bonded hydrogens (CH, CH2, CH3, etc.). The STEREO program first finds the number of each type of atom for which a request of equivalent sets has been made. The equivalence sets of each type of atom based on the constitutional symmetry group are computed and checked against the desired equivalence sets. Since the equivalence sets can never be larger in any stereoisomer than that allowed by the constitutional symmetry, all constitutional isomers which have insufficient numbers of equivalent atoms can be eliminated at this point. If the structure passes this test, the sets of equivalent atoms of the generated stereoisomers (obtained from the generally smaller symmetry group of the stereoisomer) are checked against the desired equivalence sets.

Finding equivalent substructures (i.e., those with more than one nonhydrogen atom) requires a different procedure. First, constitutional matchings are sought. Each set of n equivalent substructures is found by graph matching a new substructure, which consists of n copies of the original substructure, into the structure. The time required to do this is reduced by "scoring" the nodes of the structure with the size and identity of its equivalence set based on the constitutional symmetry group. Once the first copy of the substructure is matched, then only those nodes in the structure which are equivalent to part of the first matched substructure are considered further. Once an equivalent set (based on the constitutional symmetry group) is found, the set of permutations in the constitutional symmetry group which take the leading member of the equivalent set to each remaining member of that equivalent set are found and stored. When stereoisomers are generated (with their symmetry groups), it is verified that at least one of each of these sets of permutations are symmetry elements of the stereoisomer. (If all desired equivalent sets of substructures are not found based on the constitutional symmetry, then stereoisomers need not be generated.)

Structures with no symmetry are easily found by first checking that the symmetry group of the constitutional isomer and then, if necessary, the symmetry group of the stereoisomers have no nontrivial elements. This is a simple procedure (compared to finding equivalent atoms or substructures) and is done separately from the finding of equivalent sets.

3. Application of Constraints during Stereoisomer Generation. The method of (unconstrained) stereoisomer generation has been described previously.[6] Constraints are applied to this generation at various times during the process. Each stereoisomer is represented as a bit pattern in which the 0 or 1 of each bit indicates the configuration of one stereocenter. Substructural constraints are processed until they also are represented as bit patterns indicating the desired relative or absolute configurations of the stereocenters in the stereoisomers. These constraints are applied before stereoisomer generation (prospectively) so that only stereoisomers which have the proper relative or absolute stereocenter configurations are ever passed through the generator. The constraints concerning chirality and numbers of equivalent atoms or substructures are applied after generation when the symmetry group of the stereoisomer is known. (A byproduct of the generation of a canonical stereoisomer is the symmetry group of that stereoisomer.) Stereoisomers which fail these constraints are deleted. Atom type constraints are applied during the search for stereocenters.[6] If nitrogen atoms are to be stereocenters only at bridgeheads, then only those nitrogens which when "removed" from the structure do not disconnect it are considered further as stereocenters.

Experimental Section

The CONGEN and STEREO programs are written in the BCPL programming language.[21] The STEREO program is actually implemented as a module of CONGEN so that generation of constitutional isomers can be followed directly by generation of stereoisomers under constraints. The version of STEREO described here is available on an experimental basis, accessible over nationwide computer networks at the SUMEX computer facility at Stanford. The programs are also available for export to Digital Equipment Corporation PDP-10 and -20 systems. Our staff is available to provide guidance on obtaining versions of the programs for other computer systems. Please contact the authors for details.

(20) Brown, H. *SIAM J. Appl. Math.* **1977**, *32*, 534.

(21) Richards, M.; Whitby-Strevens, C. "BCPL—The Language and its Compiler", Cambridge University Press: Cambridge, England, 1979.