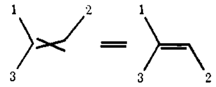
The configuration about double bonds must also be conveyed to the viewer. If the X,Y projection of a double bond, as viewed by the user, is in agreement with the symbolic representation contained in the connection table, then the double bond is displayed normally; otherwise the disagreement is indicated by crossing the double bond, indicating the true configuration is opposite to that shown.



Using these techniques, SECS assures the chemist that the stereochemistry he sees is precisely that which is in the connection table of the precursor. From the distorted diagram **13a**, it is difficult to determine if the implied oxy-Cope is reasonable, but after atoms are moved as in **13b**, it is more apparent that the transform is stereochemically plausible.[16] As the chemist moves atoms manually to obtain an alternate view of the structure, the program dynamically modified the hashing and wedging to maintain the correct stereochemical representation.

Conclusions

We have developed an unambiguous method for describing stereochemical configurations to a computer and for representing them within a computer in a way that facilitates stereochemical analysis. It was shown that stereochemical selection rules for common chemical reactions can be represented and applied by machine to a chemical structure. Finally, a simple algorithm was described that transforms the machine representation of a structure into an unambiguous structural diagram including proper stereochemical designations. This representation of stereochemistry provides the basis for naming stereoisomers uniquely and for recognizing enantiomers.[5] The same configurational information facilitates the generation of a stereochemically correct three-dimensional model,[4] which can be utilized in evaluating steric congestion,[17] and reaction mechanisms.[18] The described algorithms not only increase the selectivity of transforms, but also increase evaluation capabilities, allowing recognition of strained precursors, *e.g.*, those containing a trans double bond in a small ring, or containing a *transoid* bridged ring system. Thus, it is now possible for a computer to assist in synthetic design, not only in the crude connectivity of molecules, but also in the fine details of stereochemistry. Subsequent papers in this sequence will illustrate actual syntheses produced by the SECS program using the principles described here.

Acknowledgment. The authors thank Dr. W. Clark Still and Peter Friedland for their contributions in the development of ALCHEM and Dr. Peter Gund for his work on the model builder which led to stimulating discussions concerning this work. The work was supported by the National Institutes of Health Research Resource Grant RR00578 and in part by the Merck Company Foundation and Eli Lilly and Company.

[16] W. L. Scott and D. A. Evans, *J. Amer. Chem. Soc.*, **94**, 4779 (1972).

[17] W. T. Wipke and P. Gund, *J. Amer. Chem. Soc.*, **96**, 299 (1974).

[18] T. M. Gund, P. v. R. Schleyer, P. H. Gund, and W. T. Wipke, to be submitted for publication in *J. Amer. Chem. Soc.*

Stereochemically Unique Naming Algorithm

W. Todd Wipke* and Thomas M. Dyott

Contribution of the Department of Chemistry, Princeton University, Princeton, New Jersey 08540. *Received August 9, 1973*

Abstract: An algorithm has been developed and implemented to generate for each chemical structure a unique and invariant linear name which includes double bond and asymmetric carbon isomerism. A logical proof is given for the one-to-one correspondence between name and structure. By inspection of the linear names of two structures, one can determine if the two structures are identical, nonisomeric, constitutionally isomeric, diastereomeric, or enantiomeric. The algorithm determines the true stereocenters and calculates a reduced set of chiral centers, S_{RC}. It is proven that if there are any centers in S_{RC} that the compound must be chiral; an achiral compound must have S_{RC} = null. Extensions of the algorithm are outlined to allow uniquely naming conformational isomers.

Nonunique representations of chemical structures are useful for many things: general chemical nomenclature and discourse, chemical synthesis by computer, substructure searches, etc. Registry and storage-retrieval systems, however, require exact structural matches. Searching for such matches is greatly simplified if a canonical name can be assigned, since only one search of the structure file is then required. A canonical name means for each structure there is one name and for each name there is only one structure. This is of great importance in systems, such as the Chemical Abstracts Service registry system, where information pertaining to a compound is stored with an identifying name. In our own work with the Simulation and Evaluation of Chemical Synthesis (SECS) program,[1,2]

[1] W. T. Wipke, P. Gund, J. G. Verbalis, and T. M. Dyott, Abstracts, 162nd National Meeting of the American Chemical Society, Washington, D. C., Sept 1971, No. ORGN-17; W. T. Wipke, T. M. Dyott, P. Gund, and C. Still, Abstracts, 164th National Meeting of the American Chemical Society, New York, N. Y., Aug 1972, No. CHED-39; W. T. Wipke in "Computer Representation and Manipulation of Chemical Information," W. T. Wipke, S. R. Heller, R. J. Feldmann, and E. Hyde, Ed., Wiley (1974). For related work see E. J. Corey and W. T. Wipke, *Science*, **166**, 178 (1969); E. J. Corey, W. T. Wipke, R. D. Cramer III, and W. J. Howe, *J. Amer. Chem. Soc.*, **94**, 421, 431 (1972).

[2] Other stereochemical aspects of the SECS program, including the perception of stereochemistry from two-dimensional structural diagrams, are described in the preceding paper, W. T. Wipke and T. M. Dyott, *J. Amer. Chem. Soc.*, **96**, 4825 (1974).

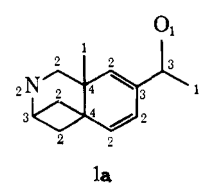
canonical names are used to check synthetic intermediates for uniqueness, and to determine if a synthetic precursor is on a list of readily available compounds.

A number of systems attempt to generate unique canonical names. Among them are the following: the Morgan algorithm used by the Chemical Abstracts Service,[3] Wiswesser Line Notation (WLN),[4] and DENDRAL.[5] The CAS Morgan name is excellent for computer applications, since it is readily converted into a more traditional connection table representation of the structure, but it does not handle stereoisomers. This paper describes an extended Morgan naming algorithm (SEMA) which provides for the proper consideration of stereochemistry. First we present the Morgan algorithm with our symmetry shortcuts, then our stereochemical extensions and the resulting consequences thereof. A complete description of SEMA appears in the Experimental Section. We feel SEMA is well tested for it has been used heavily in SECS for several years without problems.[2] While our primary interest is in the specification of configuration, we also describe provisions for including conformational information, although we have not yet implemented conformational naming.

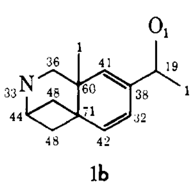
Morgan Algorithm

The generation of an unique linear name for a graph such as a chemical structure requires a set of rules that control the order in which the structure is described. Once this order has been determined, the description itself is a relatively simple process.

A key feature used in the Morgan algorithm in determining this order is called *extended connectivity*. Connectivity could be used to divide the atoms of a structure into four classes, depending on the number of nonhydrogen attachments to each atom, 1, 2, 3, or 4. (The structure is pruned of any explicit hydrogens since they are implicitly defined by the nonhydrogen connectivity, atom types, and charges.) The classification of atoms by their connectivity does not allow for much differentiation between atoms since there are only four classes. Extended connectivity attempts to be more discriminating by considering the connectivity of adjacent atoms as well. In effect it measures how centrally involved an atom is in a structure and is calculated in the following manner: (1) set the extended connectivity (EC) of each atom to its non-hydrogen connectivity; (2) count the number of different EC values (NECV); (3) set the TRIAL EXTENDED CONNECTIVITY (TEC) of each atom to the sum of the EC values of the adjacent atoms, unless the atom is primary, in which case its TEC is 1, (4) count the number of different TEC values (NTECV); (5) if NTECV is not greater than NECV, go to step 7; (6) set the EC value of each atom to its TEC value, set NECV to NTECV, and go to step 3; (7) done, the EC values are the final ones. While this method does not always allow the maximum possible differentiation on the basis of connectivity, it generally allows the atoms to be divided into many more than four classes, as shown in the following example. Diagram **1a** bears the initial connectivity values for the structure, while diagram **1b** bears the extended connectivity values. In this example, the extended connectivity values divide the 15 atoms into 12 classes. If two atoms are equivalent, considering only the connectivity of the structure, as are two of the atoms comprising the four-membered ring in the above structure, their extended connectivity values will be the same.

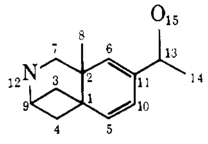


1a



1b

The order in which the structure will be described is then determined by assigning the atoms *sequence numbers* using the following algorithm: (1) choose as the current atom the atom with the highest EC value and give it the squence number 1; (2) if there are any attachments to the current atom which have not been assigned sequence numbers, then assign the unnumbered attachment with the highest EC value the next sequence number and repeat this step, else go to step 3; (3) if the structure is completely numbered, then go to step 4, else the atom with sequence number equal to the current atom plus one becomes the current atom, and go to step 2; (4) done, the sequence numbers have been assigned. (The procedure for resolving choices between attachments with the same EC values is discussed below.) The structure whose extended connectivity was calculated above would be numbered as



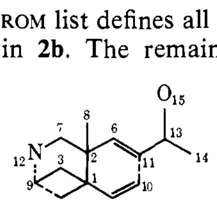
2a

This numbering is the order in which the structure will be described in the name.

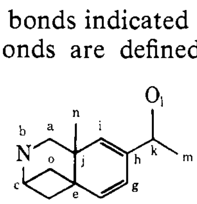
The connectivity of the structure is described by two lists, the FROM list and the RING CLOSURE list. The FROM list contains, for every atom, the number of the central atom from which it was numbered. For example, for atom 3 in the above structure the FROM list would contain a 1, while for atom 14 the FROM list would contain a 13. The complete FROM list would be

Atom Number:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
FROM List:	–	1	1	1	1	2	2	2	3	5	6	7	11	13	13

The FROM list defines all of the bonds indicated by solid lines in **2b**. The remaining bonds are defined in the RING CLOSURE list by the numbers of the atoms the bonds connect. In this case the RING CLOSURE list would be (4,9),(9,12),(10,11), listing them so they are in ascending order (*i.e.*, 0409,0912,1011). The atom and bond types are specified next in the ATOM TYPE and



2b



2c

[3] H. L. Morgan, *J. Chem. Doc.*, **5**, 107 (1965).

[4] W. J. Wiswesser, *Comput. Automat.*, **19**, 2 (1970); E. G. Smith, "The Wiswesser Line-Formula Chemical Notation," McGraw-Hill, New York, N. Y., 1968.

[5] J. Lederberg, G. L. Sutherland, B. G. Buchanan, E. A. Feigenbaum, A. V. Robertson, A. M. Duffield, and C. Djerassi, *J. Amer. Chem. Soc.*, **91**, 2973 (1969).

BOND TYPE lists. The atom types are listed in the order of the atom sequence numbers, while the bond types are listed in the order in which the bonds were defined in the FROM and RING CLOSURE lists. Charges, isotopic masses, and unusual valences, if any, are specified in the next list, which is the MODIFICATION list.

The complete name for the above structure would be

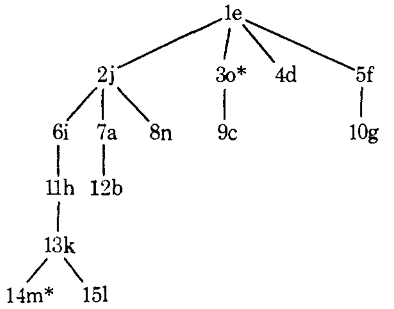
FROM list:	–, 1, 1, 1, 1, 2, 2, 2, 3, 5, 6, 7, 11,13,13
RING CLOSURE list:	4,9;9,12;10,11
ATOM TYPE list:	C,C,C,C,C,C,C,C,C,C,C,N,C, C,O
BOND TYPE list:	–, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 1, 1, 1,1, 1, 1, 1

There is no MODIFICATION list since there are no charges, etc, in the structure.

If at some point during the numbering process, or the selection of the starting atom, a choice must be made between two atoms whose extended connectivity values are the same, both names are generated, and the numbering is used which leads to the "better" name. Better is defined to be lower when the entire name is viewed as one long number. For example, if C, N, and O are represented by 1, 2, and 3, respectively, the Morgan name given above can be viewed as the number

010101010202020305060711131304090912101111111
1111111211311111111221111111

If we assign arbitrary labels to each atom as in **2c** then the sequence numbering process may be represented graphically. The asterisks beside 3 and 14 indicate choice points. An initial choice is made at each point, then alternatives are tried from the "bottom up." Thus all potential assignments are examined. Assuming, as above, that oxygen has a higher value than carbon the better name is that in which the oxygen atom, *l*, has sequence number 15 since the oxygen then appears later in the ATOM TYPE list. As the different assignments are tried, only the best one found so far need be saved and compared with future assignments. Thus assignments 2 and 4 are discarded. When two identical names are generated due to choice points (assignments 1 and 3), then the atoms interchanged by the choices (*o* and *d*) are equivalent with respect to the criteria on which the name is based. Thus, ignoring stereochemistry, atoms *o* and *d* are equivalent.



Assignment	3	4	14	15
1	*o*	*d*	*m*	*l*
2	*o*	*d*	*l*	*m*
3	*d*	*o*	*m*	*l*
4	*d*	*o*	*l*	*m*

In practice we utilize the symmetry as it is discovered in order to reduce the number of assignments that need be tried. Also, generation of a name is terminated as soon as a difference with the current best name is found. We present here a simplified description of the sequence numbering including choice resolution. A detailed description adequate for implementation is offered in the Experimental Section.

1. Let S_1 be the set of atoms having the highest EC value.

2. Choose as the current atom an untried member of S_1 which is not identical by previously discovered symmetry to an already tried member of S_1, and assign it sequence number 1. If no more untried members of S_1, go to step 12.

3. If all attachments to the current atom have been assigned sequence numbers go to step 8, else assign the unnumbered attachment having the highest EC value the next sequence number, but if tied, go to step 4; go to step 3.

4. If there is more than one ring, can not use symmetry shortcut, go to step 6.

5. If the tied attachments are interconverted by a previously discovered symmetry operation then they are equivalent, so arbitrarily choose one to be numbered next; go to step 3.

6. If the tied attachments are terminal, then examine atom type, charge, and bond type until a difference is found—choose the lower valued atom. If no difference, then the attachments are equivalent, so arbitrarily choose one to number next. Go to step 3.

7. If the attachments are not terminal, then choose one arbitrarily and mark this a choice point to try other assignment later. Go to step 3.

8. If the structure is completely numbered go to step 9, else the atom with sequence number equal to the current atom plus one becomes the current atom. Go to step 3.

9. Generate name for this assignment; if this is first assignment, accept it as best, else compare to previous best and keep the lexicographically lower one (viewed as number). If names are equal, record symmetry.

10. Search for most recent choice point (bottom up), reset current atom to the state when this choice occurred. Take next alternative assignment; continue numbering from there. Go to step 3. If no more alternatives unmark this as choice point; go to step 10.

11. If no more choice points go to step 2.

12. Best name so far is the unique name. DONE.

The use of symmetry to reduce choice making is restricted in step 4 because of the effect ring-closure bonds have in polycyclic systems. It is necessary to follow through choices in such systems so the ring closure list can be examined for each choice. The reader can see this by working through decalin as an example. Detection of symmetry is described in the Experimental Section.

Stereochemical Extension of Morgan Algorithm (SEMA)

The FROM list is a way of representing what is known in graph theory as a *spanning tree*. A spanning tree of a graph is an acyclic subgraph of the original graph which contains all of the nodes (atoms), but not necessarily all of the edges (bonds). The edges not contained in the spanning tree are known as the ring closure edges. A spanning tree for a graph with n nodes and m edges will contain $n - 1$ edges. The remaining $m - n + 1$ edges are ring closure edges. The numbering process used by the Morgan algorithm is known as "growing a spanning tree."

We could distinguish between stereoisomers if certain aspects of the three-dimensionality of the structures were incorporated into the spanning tree, causing it to become a 3-D or "stereo" tree. The name must contain the configuration of *all stereocenters*, where a stereocenter is defined to be any structural feature whose inversion produces a different stereoisomer. The two types of stereocenters commonly found in organic structures are asymmetric carbon atoms[6a] and carbon–carbon double bonds capable of cis–trans isomerization.[6b]

In this work heteroatoms or double bonds involving heteroatoms are nòt considered as possible stereocenters since they are generally subject to facile inversion. This restriction is purely arbitrary and could be removed easily if desired for particular reasons.

An easy way to store the needed configurational information is to store the parity or "handedness" of the spatial arrangement of the attachments to stereocenters.[7] First we shall generate an attachment list for each stereocenter, in which the attachment with the lowest Morgan number is placed first, followed by the other attachements, taken in a clockwise manner when viewed looking down the bond from the first. The parity of the stereocenter is then defined as even if the number of pairwise interchanges necessary to order this list in an ascending manner is zero or even; otherwise the parity of the stereocenter is odd. This parity is quite similar in nature and utility to the *RS* nomenclature system developed by Cahn, Ingold, and Prelog,[8] but there is no simple correspondence between these two systems. One is convenient for chemists while the other is convenient for computer implementation.

Conceptually this process can be viewed as noting the configuration of the stereocenters as one goes out from the root on the spanning tree; that is whether the attachments "below" (graph theoretic trees grow downward) the stereocenter on the spanning tree are in clockwise or counterclockwise order. For the top structure in Figure 1 the parity at atom 14 would indicate which of the two possible configurations, depicted on the right, is correct. In a similar manner the parity at atom 14 in the lower structure, which bears an implicit hydrogen, would indicate which of the two configurations shown on the right is correct.

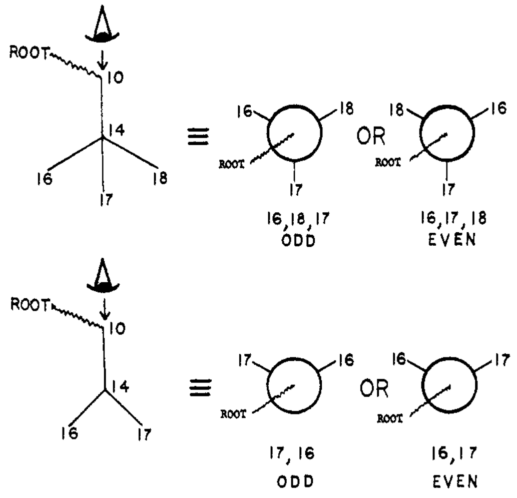
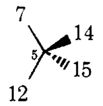
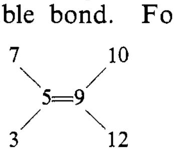


Figure 1. The parity of an asymmetric carbon atom in the stereo spanning tree indicates the ordering of the Morgan numbers as viewed from the top of the tree (top). Implicit hydrogen atoms are not shown (bottom).

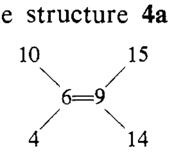
In our work the chemical structure is represented as a connection table,[2] which for each atom contains a list of the attached atoms. The list is ordered, for saturated carbons with three or four non-hydrogen attachments, such that when viewing along the bond from the first attachment the other attachments on the list are seen to be in clockwise order. The configuration is therefore inherently represented by the ordered list of attachments in the connection table.[2] The parity is easily obtained by converting the atoms on the list to their corresponding Morgan numbers and counting the number of pairwise interchanges necessary to get them into ascending order. For example, if the Morgan numbering is as shown in **3**, the connection table would contain for atom 5 the attachment list 14, 12, 7, 15. Since the list requires one interchange (7 and 14) to be put into ascending order (7, 12, 14, 15), the parity of stereocenter 5 is odd. Note that this procedure is independent of the initial ordering of the substituents.



3



4a



4b

Cis and trans double bond stereocenters are treated in an analogous manner. The attachment lists in the connection table are ordered such that the attachments to each end of the double bond are both either clockwise or counterclockwise when viewed from the same side of the plane of the bond. To preserve the configuration of a double bond in the canonical name, it is only necessary to store the *sum* of the parities of the Morgan numberings of the attachments at the two ends of the double bond. For the structure **4a**, the connection table would contain the attachment lists 5: 3, 7, 9 and 9: 5, 10, 12. Since zero interchanges are needed to order the attachments, the total parity for **4a** is even. If the Morgan numbering is as shown in **4b**, the attachment lists become 6: 4, 10, 9 and 9: 6, 15, 14. Each list needs one pairwise interchange to reach ascending order, making the total number of pairwise inter-

[6] (a) K. Mislow, "Introduction to Stereochemistry," W. A. Benjamin, New York, N. Y., 1965, p 25. Note that while *trans*-decalin is achiral, it still has two stereocenters. (b) Allenes and higher cumulenes may be handled by methods analogous to those used for cis and trans double bonds.

[7] A. E. Petrarca, M. S. Lynch, and J. E. Rush, *J. Chem. Doc.*, 7, 154 (1967); J. E. Blackwood, C. L. Gladys, A. E. Petrarca, W. H. Powell, and J. E. Rush, *ibid.*, 8, 30 (1968); A. E. Petrarca and J. E. Rush, *ibid.*, 9, 32 (1969).

[8] R. S. Cahn and C. K. Ingold, *J. Chem. Soc.*, 612 (1955); R. S. Cahn, C. K. Ingold, and V. Prelog, *Experientia*, 12, 81 (1956); R. S. Cahn, C. K. Ingold, and V. Prelog, *Angew. Chem., Int. Ed. Engl.*, 5, 385 (1966).

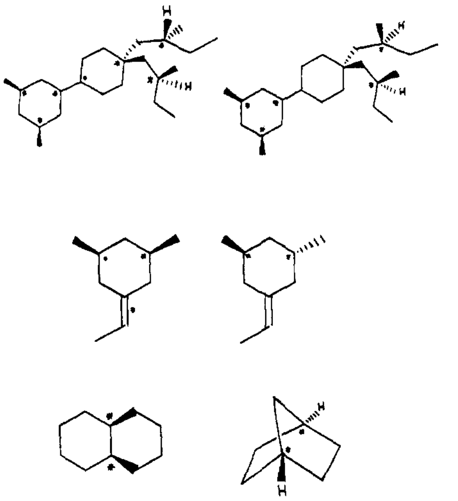


Figure 2. True stereocenters as determined by the extended Morgan algorithm are marked with asterisks.

changes two. The parity of the double bond stereocenter is therefore even.

A list of the parities of all double bonds, in the order in which they are referenced in the BOND TYPE list, is appended to the name. Another list containing the parities of all atoms, in the order of their Morgan numbers (*i.e.*, their order in the FROM list), is also appended to the name. This ordering gives the double bond stereocenters a higher priority in the naming process than the saturated carbon stereocenters. This is similar to the traditional view of double bond isomers as "geometric" isomers as opposed to stereoisomers. The values in both lists are as follows: 0 for nonstereocenter, 1 for stereocenter of odd parity, 2 for stereocenter of even parity, and 3 for stereocenter of unknown parity. The value 3 allows for the naming of structures whose configuration is not completely defined.

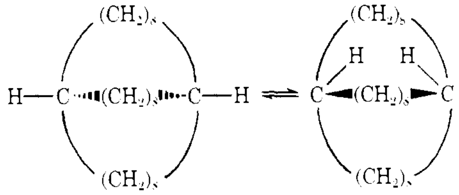
Detection of Stereocenters

In order to specify the stereochemical aspects of a structure, we first must be able to differentiate the true stereocenters from nonstereocenters. A saturated carbon atom is a stereocenter if inversion leads to a different stereoisomer. For a carbon–carbon double bond to be a stereocenter it must have two nonequivalent attachments (one of which may be a hydrogen) to each end of the double bond. The set of chiral centers is a subset of the set of stereocenters.

The determination of stereocenters can be cyclic, since the equivalence or nonequivalence of two large attachments may depend on the configuration of stereocenters within the attachments themselves. For a large multicyclic structure with many possible stereocenters, no simple algorithm immediately presents itself for determining which, if any, are true stereocenters. Most chemists would probably solve the problem, in a complex case, by building two models, one with the two attachments in question interchanged, and comparing the models. Fortunately, the nature of the Morgan algorithm, and our extension of it, allows one to easily determine stereocenters. As the naming algorithm numbers the atoms in a structure, a choice point is reached when numbering the attachments to an atom if two or more attachments have the same EC number. The resolution of a choice point requires generating the name—including the spanning tree, atom types, bond types, charges, and the parity of stereocenters—for each choice. The correct choice is the one leading to the generation of the best name, as described earlier. If the names are identical, then the attachments whose choice led to the identical names are equivalent, and the atom which bears these two attachments cannot be a stereocenter. It is possible to eliminate all but the true stereocenters in this manner, since two identical attachments will always present a choice point to the naming algorithm.

In practice we choose, as potential stereocenters, all tri- and tetrasubstituted saturated carbons and all carbon–carbon double bonds with at least one nonhydrogen substituent at each end. If during the naming, because of a choice point, two names are generated that are identical except for the parity of one possible stereocenter, then two attachments to that center are equivalent, and the center is removed from the set of stereocenters. Figure 2 shows some structures whose valid stereocenters are indicated with asterisks.

It is interesting that the bridgehead atoms in bicyclo[2.2.1]heptane are stereocenters. If one of the bridgehead atoms is inverted, one obtains an isomer with a hydrogen inside the cavity formed by the rings. While this structure would possess considerable strain energy, this "in-out" isomer has been synthesized for the [8.8.8] bicyclic system.[9] The algorithm correctly assigned the "in-in" and "out-out" isomers the same name, while assigning the "in-out" isomer a different name. The "in-in" and "out-out" isomers are merely different conformers, interconvertible by pulling one of the bridges through the ring made by the other two bridges, as depicted below.



Proof of One-to-One Correspondence

For a naming algorithm to be truly useful, there must be a one-to-one correspondence between the different entities to be named and the different names generated. This allows us to determine the equivalence or nonequivalence of two entities by naming each of them and comparing the two names.

The entities with which we must deal are representations of molecular structures—not the structures themselves. Structural diagrams (2- or 3-D), framework models, and connection tables are examples of representations. A representation is well-formed if it conforms to the grammatical rules of chemistry and of that type of representation, *i.e.*, uses the proper symbols, is unambiguous, etc.[2] A representation is a *correct* representation of a given compound if it is well formed and conveys the *constitution* of that compound, *i.e.*, the atom types, bond types, and connectivity of the structure; *and* the *configuration* of each stereocenter, double bonds and asymmetric atoms. For the current dis-

[9] C. H. Park and H. E. Simmons, *J. Amer. Chem. Soc.*, **94**, 7184 (1972).

cussion we will ignore conformational information; thus there is an infinite number of structural diagrams which are correct representations of a given compound. Two representations are *equivalent* if they are *correct representations* of the same compound. We presented in the preceding paper[2] an algorithm for translating a well-formed structural diagram (2- or 3-D) into an *equivalent well-formed* connection table (CT). In this process, for reference purposes, it is necessary to affix to each node in the structural diagram (chemical graph) a unique arbitrary label (in an interactive system this label often reflects the order of input and is truly arbitrary). The information in the CT is ordered according to these arbitrary labels; hence, since there are $n!$ ways to assign the labels, there are $n!$ correct CT representations of a compound. (Actually there are many more than $n!$ since the attachments to stereocenters can also be permuted.[2]) With these definitions and concepts in mind we now show that there is a one-to-one correspondence between the set of correct representations of a given compound and the SEMA name for that compound.

The proof of one-to-one correspondence can be divided into two subproofs: (1) a proof that any correct representations of a given compound will always be given the same name; (2) a proof that correct representations of nonidentical compounds will always be given different names. We now present these two subproofs. TO PROVE: any correct representation of a compound will be given the same name.

METHOD: proof is accomplished by showing the name is invariant with respect to the spatial positioning of the nodes (except that the configuration of stereocenters must be correctly represented) and the order in which the structure is initially described (*i.e.*, the input order).

Given two correct representations, X and Y, of compound Z it is clear that after pruning off all nodes representing hydrogen that the number of nodes in X and Y must be equal. The same is true of the number of edges. For every node in X there must be a corresponding node of the same type with the same connectivity in Y, *i.e.*, X and Y must be isomorphic. In order for the name, that is, the FROM, RING CLOSURE, ATOM TYPE, BOND TYPE, MODIFICATION, DOUBLE BOND CONFIGURATION, and ATOM CONFIGURATION lists, to be invariant, the nodes of the representation must be ordered invariantly (sequence numbers). For the sequence numbers to be invariant, the numbering process must be independent of the initial labels assigned (the input ordering) and position of the nodes. Node position is used only for perception of configuration;[2] thus it only remains to show that sequence numbering is invariant with initial labels.

The primary basis upon which sequence numbers are assigned is extended connectivity (EC). The algorithm to calculate EC as described above utilizes solely the connectivity of the nodes in the representation. The EC values are therefore invariant with respect to initial labels (input ordering).

From the previous four-step description of the sequence numbering algorithm it is seen that if there is only one node of highest EC value, then step 1 of the general algorithm is unambiguous. And if in step 2 we were never faced with a choice in choosing between two or more unnumbered attachments to the current atom which have the same EC value, then all sequence numbers would be assigned in an invariant manner. Consequently, the name generated from this numbering, including atom types, bond types, charges, and configurations, is invariant. (The configurations are derived from the sequence numbers, not the original input numbers.)

However, such choice points may occur and in practice frequently do occur. If all possible combinations of all possible alternatives encountered at the choice points are taken, a set of sequence numberings S is generated. The set S is invariant since during the numbering the choices are a function of the EC and the atom interconnectivity, both of which are invariant. The order in which members of S are discovered is dependent on the input numbering, but it is the population of S, not order of discovery, that is important as we show below.

Each numbering s_i in S leads invariantly to a name n_j, so that the set of names N derived from the set of sequence numberings S is also invariant. (Note that a one-to-one correspondence does not exist between S and N. If the structure is symmetric, there will be numberings in S which are related by the interchange of equivalent atoms. These numberings will lead to the same name.) We can invariantly choose the name n_k in N which has the lowest value when viewed as a number. Thus, given representation X, we will invariantly select the name n_X in N_X and similarly, given Y, we will invariantly select name n_Y in N_Y. Because X and Y are both *correct* representations of the same compound Z, X and Y must represent the same constitution and configuration and may only differ in positioning of nodes and in initial node labels (order of description or input). (Hashed and wedged bonds may vary but stereocenter parity must be the same.) Since S and thus N depend only on constitution and configuration it follows that $N_X = N_Y$ and $n_X = n_Y$. Thus all correct representations of compound Z lead invariantly to a single name.

TO PROVE: correct representations of nonidentical compounds will always be given different names.

METHOD: proof of the contrapositive; if two representations are given the same name then the representations are equivalent and correspond to identical compounds. The names consist of the following: (1) number of atoms, (2) number of bonds, (3) FROM list, (4) RING CLOSURE list, (5) ATOM TYPE list, (6) BOND TYPE list, (7) MODIFICATION list, (8) DOUBLE BOND CONFIGURATION list, (9) ATOM CONFIGURATION list. The name is comprised of these lists appended linearly together. It is essentially a nonredundant connection table. Let us assume the lists within the name are separated by markers. (Actually, the position of the separations need not be marked, but can be calculated from the number of atoms and bonds.) If two names are the same then all corresponding parts of the names are the same. Therefore the representations have the same number of atoms and bonds. Since the FROM and RING CLOSURE lists define all of the bonds between the atoms, representations whose names contain the same FROM and RING CLOSURE lists have the same framework or atom interconnectivity. Identical ATOM TYPE lists indicate the corresponding atoms in the structures are of the same type, while identical BOND TYPE lists indicate the corresponding bonds are of the same type. The

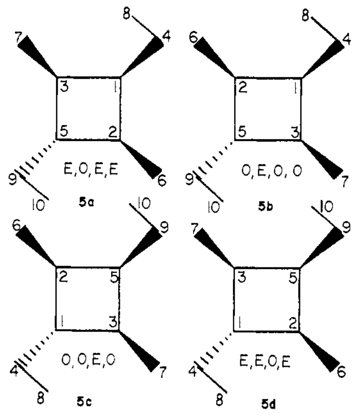


Figure 3. The configurational descriptors must be used to choose between the equivalent sequence numberings for structure **5**.

equivalence of the MODIFICATION lists indicates the charges, special isotopic masses, etc., of the atoms are the same in both representations. The equivalence of the DOUBLE BOND CONFIGURATION lists indicates the configurations of the corresponding double bond stereocenters in each representation are the same. Finally the equivalence of the ATOM CONFIGURATION lists indicates the configurations at the corresponding atoms of each representation are the same. Thus, the representations are *equivalent*, ignoring conformational differences, and correspond to the same compound. It follows that nonequivalent representation must be assigned different names for if they were assigned the same name, the representations must be equivalent—a contradiction.

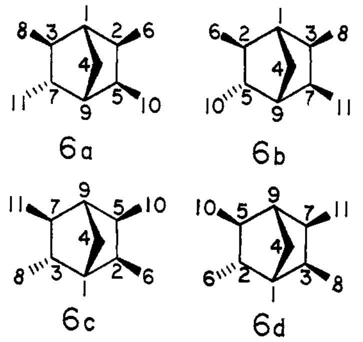
A Necessary Condition

The one-to-one correspondence between name and structure is possible only because the configurations of valid stereocenters are used in determining the unique sequence numbering (in a symmetric structure, one of a unique set of equivalent numberings) on which the name is based. In contrast, if one attempts to derive a unique sequence numbering ignoring configurations, then uses this numbering in assigning configurations which are finally appended to the name, then *the one-to-one correspondence is lost!*

To illustrate this point, let us apply the original Morgan algorithm (without stereochemistry) to structure **5**, producing four equivalent numberings shown in Figure 3. The original Morgan algorithm arbitrarily chooses one of these, the choice being dependent upon the order in which the atoms were initially described, *i.e.*, the input order. But since each numbering leads to the assignment of a different set of parities to be appended to the name (Figure 3), an arbitrary choice of numbering leads to an arbitrary name which varies with input order. *The SEMA algorithm described in this paper, however, arrives at the numbering shown in* **5c**, *irrespective of the input order:* with O = 1 and E = 2, the parity string O,O,E,O is the lowest and hence is preferred.

Enantiomers

The stereochemically extended Morgan algorithm as derived here differentiates between enantiomers, giving each a unique name. In synthetic analysis of a racemic

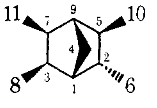


ATOMS:	1	2	3	5	7	9
6a	E	O	E	O	O	O
6b	O	E	O	O	O	E
6c	O	E	E	E	O	E
6d	E	E	E	O	E	O

Figure 4. From the set of possible sequence numberings for this structure, **6b** provides the lowest set of configuration descriptors (O = 1, E = 2). Descriptors for the enantiomer would be opposite those shown.

target structure we would like to treat any enantiomeric precursors produced as duplicates, and represent the *d,l* pair by one structure, just as chemists represent a *d,l* pair by one structural diagram. Thus, we wished to be able to recognize the existence of an enantiomeric relationship by simply comparing the SEMA names of two structures. We implemented a minor addition to the SEMA name to provide this powerful capability.

The names of an enantiomeric pair will be identical up to and including double bond stereodescriptors, but will differ in saturated carbon configuration descriptors. For structure **6** the SEMA generates four sets of sequence numberings and corresponding configuration descriptors (Figure 4) from which it selects **6b** (O,E,O,O,O,E) as the best set of descriptors. The reader should verify that **7**, the enantiomer of **6**, leads to four sets of configuration descriptors different from **6a–d** and in fact the complement of **6a–d**. From these, the best set of descriptors for the enantiomer **7** is O,O,O,E,O,E



7

corresponding to stereocenters 1,2,3,5,7,9, respectively. Comparing the best descriptors for **7** with the best descriptors for the enantiomer **6**, a difference is noted at

Atoms:	1	2	3	5	7	9
6	O	E	O	O	O	E
7	O	O	O	E	O	E

stereocenters 2 and 5, meaning that inverting atoms 2 and 5 in **6b** produces the enantiomer **7**.

For convenience we define for a structure the *reduced set of chiral centers* (S_{RC}) as that set of stereocenters which must be inverted in the SEMA name of the structure to produce the SEMA name of the enantiomeric structure. The set S_{RC} is easily found. (1) Let the set of equivalent sequence numberings (SESN) be the set of best sequence numberings which produce SEMA names differing only in saturated carbon configuration descriptors. (2) For the SESN, compare the lowest and highest configuration descriptor sets. Let $\mathcal{S}_{RC}$ contain those centers which maintain the same parity. If only one sequence numbering was found, SESN contains only one member. In that case, the low and high configuration descriptor sets are identical, hence $\mathcal{S}_{RC}$ = the set of valid stereocenters.

Applying this procedure to **6** we compare O,E,O,O,-O,E with E,E,E,O,E,O, finding 2 and 5 retain the same parity; hence $\mathcal{S}_{RC} = \{2,5\}$. In the SEMA name we mark each center in $\mathcal{S}_{RC}$ by adding 3 to the parity of that center—odd becomes 4, even becomes 5. The final configuration descriptors for **6** and its enantiomer **7** are then

6 1 5 1 0 4 0 1 0 2 0 0
7 1 4 1 0 5 0 1 0 2 0 0

Thus, given the SEMA name of a structure we simply replace 4's by 5's and *vice versa* in the configuration descriptor part of the name to generate the SEMA name of the enantiomeric structure.

To further illustrate the significance of the reduced set of chiral centers, consider structure **5** in Figure 3. Comparing O,O,E,O with E,E,O,E one finds no stereocenters retain parity, hence $\mathcal{S}_{RC}$ = null and the final configuration descriptors are 1,1,2,1. That $\mathcal{S}_{RC}$ = null implies the mirror image of **5** also has configuration descriptors 1,1,2,1; *i.e.*, **5** and its mirror image are identical. By inspection one sees that **5** possesses a plane of symmetry; thus **5** is achiral and has no enantiomer. The utility of $\mathcal{S}_{RC}$ is reinforced by the following theorem.

THEOREM: *if a compound (X) has $\mathcal{S}_{RC} \neq$ null, the compound is chiral.* As mentioned earlier, the mirror image of X, X′ gives configuration descriptor sets (CDS's) which are the complement of the CDS's for X. The highest CDS of X is the complement of the lowest (best) CDS of X′. If $\mathcal{S}_{RC} \neq$ null, then from the algorithm for finding $\mathcal{S}_{RC}$, some centers in the lowest and highest CDS's have the same parity. Consequently, the lowest CDS of X and the lowest CDS of X′ have some centers which differ in parity, and thus X and X′ will have different SEMA names. From the proof of one-to-one correspondence between SEMA name and structure, X and its mirror image, X′, are not identical, hence X is chiral.

COROLLARY 1: *if a compound is achiral, then $\mathcal{S}_{RC}$ = null.* The proof of this is obvious after the example of structure **5**.

COROLLARY 2: *if a compound is chiral due to causes other than conformation or heteroatom stereochemistry, then $\mathcal{S}_{RC} \neq$ null.* The chirality of such a compound must be due to stereocenters; consequently the CDS's of the compound and its mirror image must be different. By the previous discussion, if such a difference exists, $\mathcal{S}_{RC}$ will contain those centers which are different, hence $\mathcal{S}_{RC} \neq$ null.

Figure 5 illustrates these points with more complex examples. In each example the calculated members of $\mathcal{S}_{RC}$ are indicated by large dots. Inversion of these marked centers generates the enantiomer. Note that $\mathcal{S}_{RC}$ is a unique set because of its definition, but there are often other sets of stereocenters, which, when inverted, also produce the enantiomeric structure. The algorithm correctly determines that all structures in Figure 5 are chiral except **9** and **12**. Thus, from the SEMA name one can determine not only enantiomeric relationships, but subject to the exceptions of corollary 2, he can also determine whether the compound is chiral or achiral.

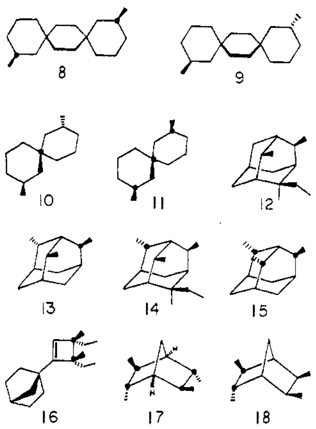
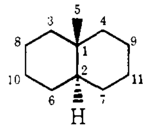


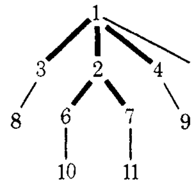
Figure 5. Structures processed by the extended Morgan algorithm. Those centers in $\mathcal{S}_{RC}$ are marked by large dots. Inversion of the marked centers produces the enantiomeric structure.

Conformational Naming

The described extension of the Morgan name allows the complete configurational specification of organic structures. To extend the name to include a complete description of the conformation as well as the configuration, we need only expand the concept of a stereo spanning tree to include rotational angles about bonds in the spanning tree. From the extended Morgan algorithm on *trans*-decalin (**19**) one obtains the spanning tree A. To completely specify the conformation of **19**, we need specify only the configuration of atoms 1 and 2 (O,O, respectively) and the dihedral angles centered on the bold bonds in A. The three bonds spanned by the dihedral angle must be present in the spanning tree; thus, the dihedral angle 1–3–8–10 is not included since the bond 8–10 is absent in the spanning tree. When several dihedral angles exist about the same central bond, the angle specified is that formed by atoms with the lowest SEMA sequence numbers. These rules dictate the minimum member of dihedral angles required for complete specification of conformation. The values representing the dihedral angles are listed in the same order as the central bond of the dihedral angle is referenced in the BOND TYPE list in Table I.



19



A

For many purposes, the dihedral angles need not be precise. For example, ring conformations have been described by a system with only three values (+,0,−).[10] One can represent all staggered and eclipsed conformations with the scheme in Figure 6. One views down the central bond in order of increasing Morgan numbers, *i.e.*, from the top of the spanning tree, and selects one

[10] J. B. Hendrickson, *J. Amer. Chem. Soc.*, 86, 4854 (1964).

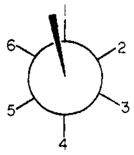


Figure 6. One of six values is chosen to represent the actual dihedral angle as viewed down the spanning tree.

Table I

Atom	From list	Dihedral angle
1		
2	1	3-1-2-6
3	1	4-1-3-8
4	1	2-1-4-9
5	1	
6	2	1-2-6-10
7	2	1-2-7-11
8	3	
9	4	
10	6	
11	7	

of six possible values to represent the dihedral angle as shown. In an 8 valued system (2^3), which is efficient for computer representation, values 0 and 7 remain to represent unspecified dihedral angles or free rotation, respectively.

If the conformational descriptors were added to the linear name after configurational descriptors and were used in choice resolution in the stereochemically extended Morgan algorithm, each conformation, which is unique within the precision of the dihedral angle representation, would receive a unique name. With such complete information concerning the structure of a molecule contained within the name, it would be easy to generate, directly from the name, a three-dimensional model of the molecule.

Conclusions

This research has created a comprehensive machine oriented naming algorithm, capable of generating a unique name for each and every stereoisomer, independent of the conformation of the original numbering of the input structure. The one-to-one correspondence between name and structure was logically proven. This algorithm now permits error-free registration of complete chemical structure without manual intervention. Using this stereochemically unique name, automated searching for structures should be faster and more specific.

The name is so structured that it is possible to determine if two structures are identical, nonisomeric, constitutionally isomeric, diastereomeric, or enantiomeric, simply by comparing the component parts of the two respective names. The name also indicates true stereocenters, and a reduced set of chiral centers, which upon inversion produce the enantiomeric structure. From this latter set one can, subject to the exceptions stated in corollary 2, determine if a compound is chiral or achiral. We expect this algorithm to be helpful in any computer-assisted stereo-mechanistic study as well as in computer-assisted design of organic synthesis.

We demonstrated how this algorithm could be extended further to uniquely name conformers. This extended algorithm would also uniquely name the stereochemical problem cases of enantiomerism arising from restricted rotation and helicity. From such a conformationally unique name the three-dimensional model of the structure could easily be reconstructed.

Although the algorithm and name are machine oriented, they can provide the chemist a nomenclature-free communication path to important files of chemical information through convenient man–machine interfaces[2] which convert the common structural diagram into a stereochemically unique name. All aspects of this work except the conformational extension were implemented in Fortran IV on a Digital Equipment Corporation PDP-10 computer.

Acknowledgments. This research was supported by National Institutes of Health Research Resources Grant RR00578, and in part by the Merck Foundation and Eli Lilly and Company.

Supplementary Material Available. The Experimental Section of this paper describes SEMA in explicit algorithmic form in sufficient detail to enable implementation. We also present execution times and illustrate the relationship between structure and execution time. Finally, symmetry recognition is discussed. This material will appear following these pages in the microfilm edition of this volume of the journal. Photocopies of the supplementary material from this paper only or microfiche (105 × 148 mm, 24× reduction, negatives) containing all of the supplementary material for the papers in this issue may be obtained from the Journals Department, American Chemical Society, 1155 16th St., N.W., Washington, D. C. 20036. Remit check or money order for $3.00 for photocopy or $2.00 for microfiche, referring to code number JACS-74-4834.