match no. 6 pp. 259 - 275 1979

NONNUMERICAL MATHEMATICAL METHODS IN THE PROBLEM OF STEREOISOMER GENERATION

James G. Nourse, Dennis H. Smith

Departments of Chemistry, Computer Science, and Genetics,
Stanford University, Stanford California 94305

Abstract

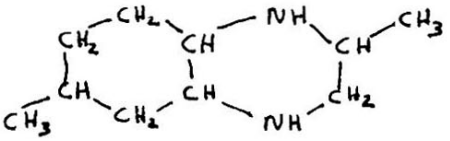
The problem of the generation of all possible isomers consistent with a given empirical formula will be discussed. We factor this problem into two parts. The generation of all constitutional isomers, and the generation of all stereoisomers for each constitutional isomer. The method of generation of all constitutional isomers is briefly reviewed with particular emphasis on the graph theoretic and group theoretic methods used to solve the problem. The method for finding stereocenters and making use of structural symmetry are detailed. Particular emphasis is placed on the chemical graph used and the particular representation of the symmetry group on the configurations of stereocenters.

Introduction

A solution to the problems of exhaustive isomer generation and enumeration is critical to the problem of computer-assisted chemical structure elucidation since this will give the assurance that no possibilities have been overlooked. While the problem is chemical as are the applications of the resulting programs, there is a heavy component of mathematics, particularly theory, combinatorics, and group theory to the solution of the problems of isomer generation and enumeration. Our solution to these problems factors into two parts, the problems of constitutional isomerism and stereoisomerism. The method for the generation of all the constitutional isomers will be mentioned only briefly (1) while the method for stereoisomers will be discussed in greater detail. Emphasis will be placed on the identification and application of the mathematical structures necessary to the solution of the problem.

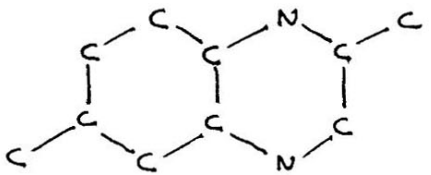
CONSTITUTIONAL ISOMERS

Chemical structures are viewed as graphs in which the atoms are nodes and the bonds are edges. For example, consider the bicyclic structure 1.



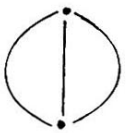
1

The chemical graph is considered to be the structure without hydrogen atoms. It is assumed that all unspecified valences are occupied by hydrogen atoms. This chemical graph is shown in structure 2.



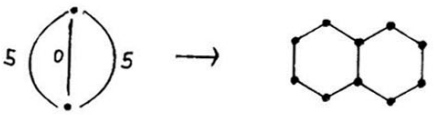
2

This structure has the empirical formula $C_{10}N_2H_{20}$ and is one of a very large number of structures with this empirical formula. In order to construct all structures with any empirical formula we have found it necessary to isolate a key structural feature of the chemical graph called the vertex graph. A vertex graph is definded to be the cyclic skeleton of a structure with all nodes of degree less than 3 removes (1). For structure 1 the vertex graph is just the graph:



with two trivalent nodes and 3 edges. For a given empirical formula only certain vertex graphs will be observed for any possible structure and these can be computed (1,2). Thus the vertex graph is a key mathematical concept to the solution of the isomer generation problem.

A given vertex graph implies the use of a number of atoms corresponding to the number of nodes in the vertex graph. The rest of the atoms in the empirical formula are unspecified by the vertex graph. In the example, only 2 of the 12 nonhydrogen atoms are specified by the vertex graph. This leaves the problem of specifying the other 10 atoms. This problem is solved using two mathematical methods which are used repeatedly throughout the structure generation program. The first method is called partitionlng. The 10 remaining atoms must be partitioned into 3 sets in all unique ways to be placed on the three edges of the vertex graph. At this stage carbon and nitrogen atoms are treated alike. One such partition of 10 is into 5,5,0. This means that 5 atoms are put onto one of the edges, 5 onto another edge, and 0 onto the final edge. The process of placing atoms onto edges uses the second method known as labelling. In the example there is only one unique way to do this because all three edges of the vertex graph are equivalent. This gives the cyclic skeleton of the final structure:



In cases with distinguishable edges on the vertex graph and different numbers in the partition the problem is more complicated and a more elaborate procedure is necessary which takes the symmetry of the vertex graph and the partition into account directly. The solution to problems of this kind requires another common mathematical structure known as a double coset structure. For example consider the vertex graph:

and the partition 2,1,0,0,0. These 5 numbers must be mapped to the 5 edges in all possible ways. The unique mappings correspond to double cosets of the edge symmetry group and the partition symmetry group in the group S5 which correspond to the 5! mappings of the numbers to the edges. These double cosets are:

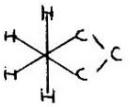
edge symmetry group / S5 / partition symmetry group

S2(S2) / S5 / S3 X S1 X S1

where Sn symbolizes the symmetric group on n objects. The partition symmetry group merely indicates the fact that there are three edges with 0 atoms, one edge with 1 atom and one edge with 2 atoms. The edge symmetry group is a wreath product (3) and indicates there are two sets of two equivalent edges. There are 4 of these double cosets and these correspond to the 4 labelled vertex graphs:

The rest of the structure generation procedure requires a series of partitioning and labelling steps which eventually yields the desired final structures. These have been discussed in detail (1).

The total set of structural isomers for an empirical formula can be extremely large and unmanageable. For most structural determination problems it is possible to reduce this set by requiring all structures to contain a particular substructure which is for some reason known to exist in the structure. For example consider the empirical formula C_9H_{16} with the requirement that all structures contain a 6-membered ring. This splits the set of atoms into two parts, those in the six membererd ring, and those which remain. The six membered ring can be thought of as a single atom called a "superatom" which is hexavalent. One of the possible structures can be symbolized as:

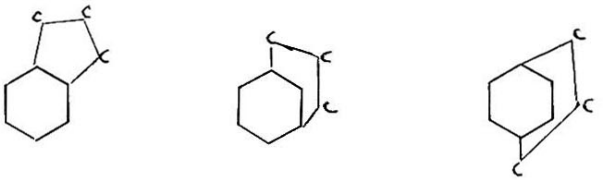


This means the six membered ring is attached to two ends of a carbon chain with three atoms. The next step is to determine the possible ways this attachement can be done. The procedure used is called "embedding" and again requires a double cosed method to properly account for symmetry. The two ends of the 3 carbon chain plus 4 hydrogen atoms must be mapped to the 6 sites on the skeleton of the 6 membered ring in all possible ways. The unique structure correspond to the double coset:

superatom symmetry / S6 / remaining atoms symmetry

D6 / S6 / S4 X S2

where the symmetry of the superatom is D6. There are three of these structures:



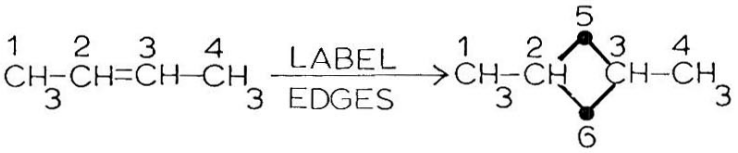
STEREOISOMERS

Once the constitution (atoms and bond connectivity) of a structure has be defined the stereoisomers can be generated. An efficient solution to the problem of stereoisomer generation requires two features:

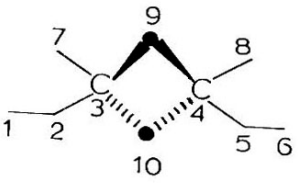
1. A method of establishing which atoms in the structure are capable of existing in more than one configuration.

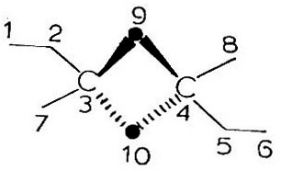
2. A method for making the proper use of any structural symmetry which makes some potential stereoisomers equivalent.

The search for stereocenters requires a special representation of the graph describing the constitution of the chemical structure. Every double bond in the structure is labelled with extra nodes in this manner:

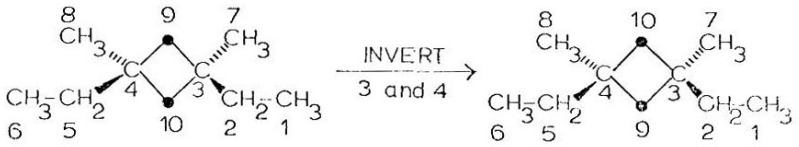


These are ficticious nodes which are formally bivalent. This processing is done so that both the atoms involved in the double bond can be formally assigned a configuration based on the four (numbered) atoms to which it is attached. For the usual olefin structure the cis and trans forms differ in the configuration of one of these atoms.

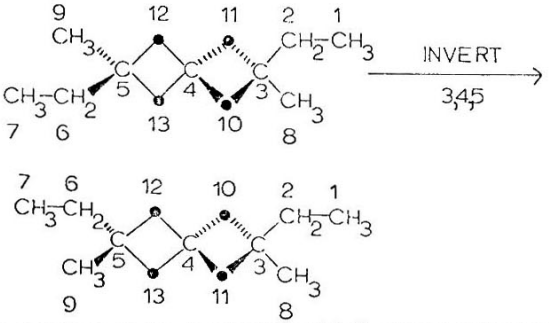




If both of the atoms in the double bond have their configuration inverted, the same structure remains. This can be accomplished by simply switching the two ficticious nodes labelling the two edges.



However if the configurations of the three atoms involved in an allene structure have their configurations inverted, the enantiomeric structure results. This is accomplished as shown. This is consistent with the idea that the enantiomer of a chemical structure can be obtained by reversing the configuration of all chiral centers.



Next it is necessary to establish which atoms are properly substituted to exist in two configurations. This corresponds to the idea that a carbon atom with four different substituents can exist in two configurations while those with some identical substituents cannot. In reality the problem is somewhat more complicated and five cases must be considered. For each potential stereocenter the key mathematical structure is the subgroup of the overall graph symmetry group which fixes that atom. The permutations in this group will then exchange the four substituents on that atom among themselves. The possible symmetries then correspond to the symmetric group on 4 objects or S4. This group has five conjugacy classes and it is these which must be surveyed in order to establish whether a potential stereocenter can exist in two configurations. The possibilities are summarized:

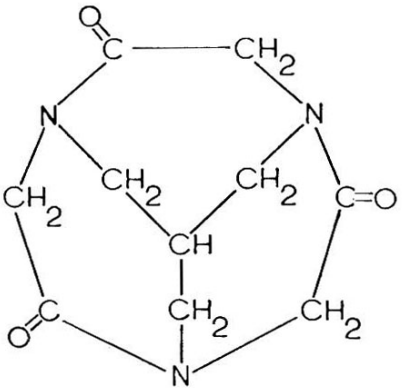
Permutation cycle type	Example	Even/Odd	Eliminate stereocenter?
(1)^4	(1)(2)(3)(4)	E	no
(1)^2(2)	(12)(3)(4)	O	yes
(2)^2	(12)(34)	E	no
(1)(3)	(123)(4)	E	no
(4)	(1234)	O	yes

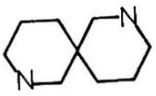
Any symmetry element in the group which fixes a potential stereocenter must be one of these five types with respect to its action on the substituents attached to the potential stereocenter. The question to be asked is whether a permutation of this type is capable of eliminating the possibility of the stereocenter existing in two distinct configurations. Whether it (the permutation) actually does eliminate the stereocenter is established later. These will be surveyed in order.

The permutation which has 4 cycles is the identity of the group. An example of a structure with such symmetry is a carbon with 4 different substituents. Such a structure can exist in two distinct configurations, hence permutations of this kind cannot eliminate a potential stereocenter. This case is trivial.

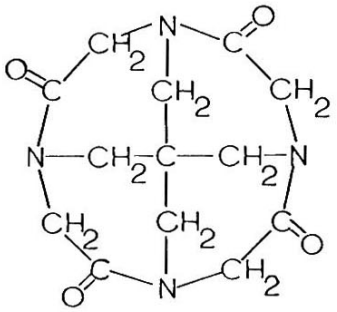
Permutation with three cycles, one of length 2, correspond to planes of symmetry in a geometric representation. These are capable of eliminating potential stereocenters, an example is a gem-dimethyl substituted carbon. However there are cases in which a carbon atom has a plane of symmetry but remains a stereocenter. Examples of such structures are pseudochiral(4) structures in which the two substituents are themselves constitutionally identical chiral ligands which have opposite configurations.

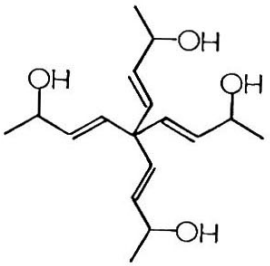
Permutations with two cycles of length 2 cannot by themselves eliminate potential stereocenters. Similarly, permutations with two cycles, one of length 1 and one of length 3 cannot by themselves eliminate potential stereocenters. Structures which have only these kinds of symmetries will exist in two enatiomeric forms.





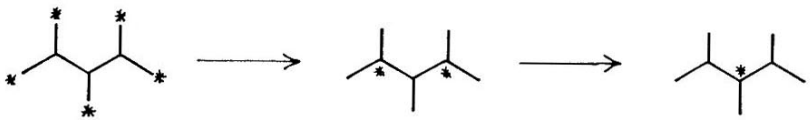
Permutations with one cycle of length 4 may eliminate potential stereocenters. However, it is still possible for a carbon with four constitutionally identical substituents to exist in two configurations. Examples of both these cases are:





In summary, the subgroup of the graph symmetry group which fixes a potential stereocenter can be used to establish whether that potential stereocenter is capable of existing in two configurations.

When a potential stereocenter has constitutionally identical substituents related by the proper type of permutation, there will be two distinct configurations only if the substituents themselves contain stereocenters. Thus substituents must be searched to see if they contain stereocenters. If not, then the potential stereocenter is eliminated. If so, then the procedure is repeated until no new stereocenters are found. To show how this iterative procedure works consider the structures:

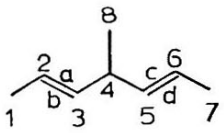


To establish that the central carbon is not capable of existing in two distinct configurations it must first be established that the methyls are not stereocenters and then that the isopropyls are not stereocenters. Since the central carbon has two substituents (isopropyl) which are not stereocenters, the central carbon is not a stereocenter.

Once the stereocenters have been found it is next necessary to establish the effect of any structural symmetry on the number of distinct stereoisomers. If there is no symmetry then there are of course 2^n stereoisomers where n is the number of stereocenters. However, if there is structural symmetry then the total number of distinct isomers can be reduced considerably.

To establish the effect of the symmetry group it is first necessary to compute it. For the particular chemical graphs being considered here, this group will always be factorizable into two parts. The first part is the group of all permutations of the ficticious nodes labelling the double bonds. This group will be a direct product of m groups of order 2 where m is the number of double bonds. This group is easily constructed. The other group is the symmetry group of the graph without the double edges labelled. This group can have an arbitrary structure and is found using a published algorithm with only slight modification (5). The desired group is the semidirect product of these two (6). This can be illustrated with an example:

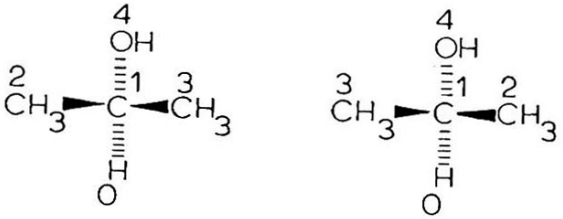
	(2)(3)(4)(5)(6)	(4)(26)(35)
(a)(b)(c)(d)	(2)(3)(4)(5)(6)(a)(b)(c)(d)	(4)(26)(35)(ac)(bd)
(ab)(c)(d)	(2)(3)(4)(5)(6)(ab)(c)(d)	(4)(26)(35)(adbc)
(a)(b)(cd)	(2)(3)(4)(5)(6)(a)(b)(cd)	(4)(26)(35)(acbd)
(ab)(cd)	(2)(3)(4)(5)(6)(ab)(cd)	(4)(26)(35)(ad)(bc)



The edge symmetry group is of order 4 and is indicated with letter permutations. The node symmetry group is of order 2 and is indicated with number permutations. The symmetry operation corresponds to the intuitive twofold symmetry axis. The product of these two groups is computed by noting the effect of the node permutations on the edges. There is no effect of the edge permutations on the nodes. There are 8 permutations of the nodes and the edges in this group. The permutations of the edges could have been just as easily represented on the ficticious nodes labelling the double edges.

It is next necessary to determine the effect of the graph symmetry permutations on the configurations of the stereocenters. Since all four substituents on each stereocenter have different numbers, the two possible configurations are defined to be the two configurations based on these numbers. Each graph symmetry operation will have one of two effects on each stereocenter. First, the graph symmetry operation may fix the sterocenter. If this is the case then the substituents on that stereocenter will be permuted. If this permutation is even, then the graph symmetry operation leaves the configuration of that stereocenter unchanged. If the permutation is odd, then the graph symmetry operation inverts the configuration of that stereocenter. In this latter case, the number of the stereocenter is given a superscript (′). For example, consider isopropanol:

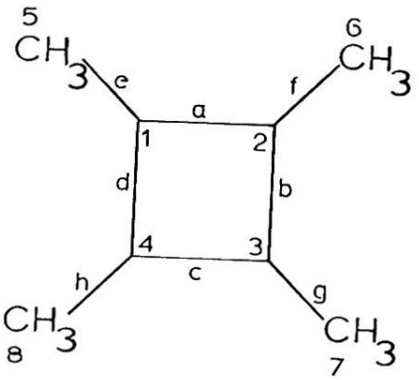
GSG	CSG
(1)(2)(3)(4)	(1)(2)(3)(4)
(1)(23)(4)	(1′)(23)(4)



There are two configurations based on the four different numbers on the substituents attached to stereocenter 1. (The hydrogen has the number 0). The only nontrivial graph symmetry operation exchanges the two methyls. This has the effect of inverting the configuration at stereocenter 1 and this is indicated by a superscript. The resulting group is called the Configuration Symmetry Group (CSG).

The second possibility is that the graph symmetry operation moves the stereocenter to a symmetrically equivalent one. In this case the four substituents on the first stereocenter A go to the four substituents on the second stereocenter B. Thus there is a mapping (a1, a2, a3, a4) $\longrightarrow$ (b1, b2, b3, b4). The substituents are ordered based on the atoms numbers. There is a permutation of these four ordered substituents caused by the graph symmetry operation. If this permutation is even then the configuration of stereocenter A remains unchanged as it goes to stereocenter B. If this permutation is odd, then the configuration of stereocenter A inverts as it goes to stereocenter B. This operation is repeated for all graph symmetry operations and all stereocenters. The resulting Configuration Symmetry Group is given for tetramethylcyclobutane:

GSG	CSG
(1)(2)(3)(4)	(1)(2)(3)(4)
(1)(24)(3)	(1')(24)(3')
(12)(34)	(12)(34)
(1234)	(12'34')
(13)(2)(4)	(13)(2')(4')
(13)(24)	(1'3')(2'4')
(1432)	(1'43'2)
(14)(23)	(1'4')(2'3')



Note that the CSG is not the same as the point group in which reflective operations invert the configurations of all the stereocenters. In the CSG for tetramethylcyclobutane shown above, there are some permutation inversions which invert only two stereocenters. Intuitively the CSG correspond to the invariance group for a stereoisomer in which the relative configurations of constitutionally identical stereocenters are taken into account. See ref. (4) for more pictorial example.

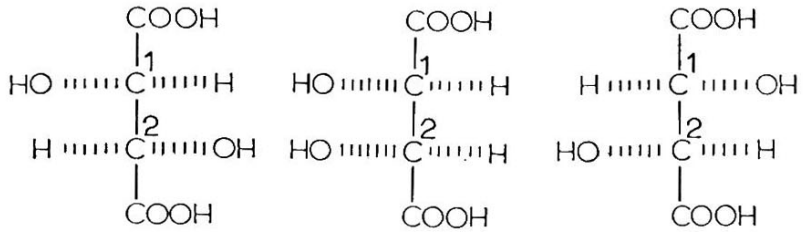
The Configuration Symmetry Group can be used to generate the distinct stereoisomers of a chemical structure. This is done by making use of the idea of an equivalence class of the CSG. If there are n stereocenters, then there are 2^n possible stereoisomers and these can be symbolized by n-tuples of the form (+-++...). This symbolizes the stereoisomer with stereocenter 1 in the (+)configuration based on the node numbers, stereocenter 2 in the (-)configuration etc. The CSG elements (called permutation inversions) act on these n-tuples in the following manner:

$$(++--) \xrightarrow{(12'34')} (--++)$$

This is read: stereocenter 1 goes to stereocenter 2 and inverts configuartion, stereocenter 2 goes to stereocenter 3 unchanged, stereocenter 3 goes to stereocenter 4 and inverts configuration, stereocenter 4 goes to stereocenter 1 unchanged. When all of the permutation inversions in the CSG act on all the n-tuples in this manner, the n-tuples are collected into equivalence classes which correspond to the distincht stereoisomers. For example, consider the stereoisomers of tartaric acid:

	(1)(2)	(12)	stereoisomer
[++]	[++]	[++]	d
[+-]	[+-]	[-+]	meso
[-+]	[-+]	[+-]	meso
[--]	[--]	[--]	l

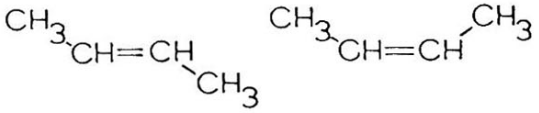
The two permutations in the Configuration Symmetry Group for tartaric acid act on the four possible stereoisomers to give three equivalence classes which correspond to the three distinct stereoisomers.



There are four elements in the CSG for 2-butene since there is the symmetry which exchanges the two edges. This permutation has the effect of inverting the configuration of both stereocenters and is therefore symbolized (1')(2'). These four permutation inversions collect the four possible stereoisomers into the two equivalence classes shown.

	(1)(2)	(1')(2')	(12)	(1'2')	stereoisomer
[++]	[++]	[--]	[++]	[--]	trans
[+-]	[+-]	[-+]	[-+]	[+-]	cis

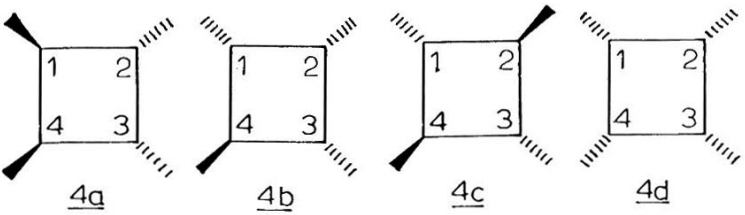
The four permutations in the Configuration Symmetry Group for 2-butene act on the four possible stereoisomers to give two equivalence classes which correspond to the two distinct stereoisomers. Each row is an equivalence class.



Finally the CSG for tetramethylcyclobutane has the 8 permutation inversions shown and collect the 16 possible stereoisomers into 4 equivalence classes which correspond to the 4 distinct stereoisomers. Each column is an equivalence class. The first element in each column is used to generate the equivalence class by the action of the permutation inversion beginning each row. Note the correlation between the number of times an n-tuple is repeated in an equivalence class with the symmetry of the stereoisomer.

	4a	4b	4c	4d
(1)(2)(3)(4)	[++++]	[-+++]	[--++]	[-++-]
(1')(24)(3')	[-+-+]	[++-+]	[++--]	[+--+]
(12)(34)	[++++]	[+-++]	[--++]	[+--+]
(12'34')	[+-+-]	[+++-]	[++--]	[-++-]
(13)(2')(4')	[+-+-]	[+---]	[++--]	[+--+]

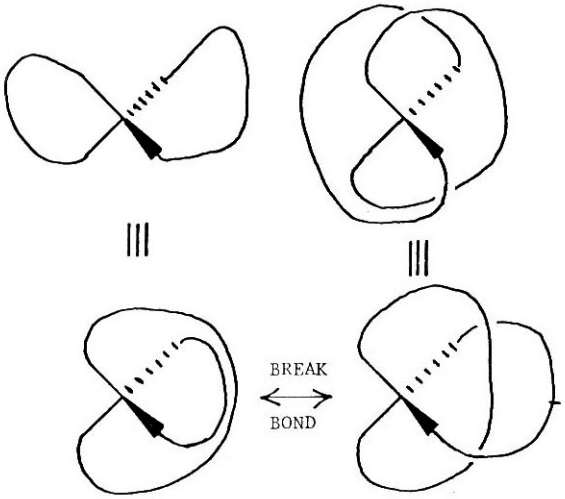
(1'3')(2'4')	[----]	[--+-]	[--++]	[-++-]
(1'43'2)	[-+-+]	[-+--]	[++--]	[-++-]
(1'4')(2'3')	[----]	[---+]	[--++]	[+--+]



4a 4b 4c 4d

This method has been converted into an algorithm which has been programmed for a computer to generate all the possible stereoisomers of a chemical structure of defined constitution. This has been combined with the structure generated discussed above to give a program which generates all the possible isomers consistent with a given empirical formula (7). The program also produce a canonical (unique) representation of the configuration of the configurational stereochemistry of a structure which can be appended to a canonical numbering of the constitutional graph of the structure (7,8). This is an advantageous feature for computer-assisted structure elucidation (9).

It is interesting to note that the limitation of this method is that it considers only the connectivity of a chemical structure and augments these with configuration parity labels. While this implicitly defines a geometrical feature of the structure, it is actually devoid of any geometric information, hence all notions of conformation are missing. Furthermore, no energetic information is used so that the relative stability of stereoisomers is not treated. Since only connectivity is considered, structures which can be interconverted by passing bonds through bonds are considered the same. This includes catenanes and interesting structures such as the inverted spiran shown. This can be interconverted with the usual spiran by passing bonds through bonds as shown.


BREAK
BOND

We gratefully acknowledge financial support from the National Institutes of Health (2R24 RR 00612-08).

REFERENCES

(1) L.M. Masinter, N.S. Sridharan, J. Lederberg, and D.H. Smith, J. Amer. Chem. Soc., 96, 7702 (1974)

(2) R.E. Carhart, D.H. Smith, H. Brown, and N.S. Sridharan, J. Chem. Inf. Comp. Sci., 15, 124 (1975).

(3) F. Harary, "Graph Theory", Addison-Wesley, Reading, Mass., 1969, ch. 14

(4) J.G. Nourse, J. Amer. Chem. Soc., 97, 4594 (1975).

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

(6) (a) M. Hall, Jr., "The Theory of Groups", Mac Millan, New York, N. Y. 1959. (b) M.G. Hutchings. J.G. Nourse, and K. Mislow, Tetrahedron, 30, 1535 (1974)

(7) (a) J.G. Nourse, submitted. (b) J.G. Nourse, R.E. Carhart, D.H. Smith, and C. Djerassi, submitted.

(8) W.T. Wipke and T.M. Dyott. J. Amer. Chem. Soc., 96, 4834 (1974)

(9) R.E. Carhart, D.H. Smith, H. Brown, and C. Djerassi, J. Amer. Chem. Soc., 97, 5755 (1975).