Algorithm for Exhaustive and Nonredundant Organic Stereoisomer Generation†

M. L. Contreras,*,‡ J. Alvarez,§ D. Guajardo,§ and R. Rozas‡

Faculty of Chemistry and Biology, Department of Chemical Sciences, and Faculty of Engineering, Department of Information Technology, University of Santiago de Chile, Casilla 40, Santiago-33, Chile

Received June 30, 2006

Generation of organic stereoisomers with *R*/*S*, *Z*/*E*, and/or *M*/*P* configurations that may contain heteroatoms, multiple bonds, and any kind of cycle (isolated, spiro, condensed, and nested) is described. Inputs for processing are molecular structures in a *N*_tuple format resident on an automatic (canonical) or manual (non canonical) generated file which are processed by doing internal molecular graph construction, a weighted bipartite tree construction for all atoms and bonds to detect stereocenters, and symmetrical atom groups (SAG) with some specific SAG parameters that constitute a novel way for redundancy elimination of *meso* structures. Finally, determination of ligand CIP priorities allows for writing the output *N*_tuples with stereoisomer description. Several examples showing application of this methology to a wide number of structures are also presented.

1. INTRODUCTION

Computer-assisted exhaustive generation of stereoisomers is an important contribution to structure elucidation[1−7] and molecular design.[8−11] At present, areas such as drug or material design are relevant and also require this type of computer system and software. In that context computational generator systems[12−19] are an important complement for systems such as similarity searching,[20−23] database systems,[24−30] descriptor calculators of different types,[31−36] and many others[37−39] that could became part of a virtual high throughput screening system oriented in a predictive way as a useful making decision tool.

Many of these programs generate mainly constitutional or connectivity isomers (also known as topological isomers), under different strategies, i.e., structure assembly,[40−42] structure reduction,[43] or convergent structure generation as is the case of HOUDINI,[19] a program of structure elucidation having two major components: a hyperstructure and a substructure representation. Each of these having its own data structure, the contents of which, at any stage of structure generation, reveal the status of the process. The data structure of the hyperstructure is a two-dimensional, symmetric, square matrix in which the labels of the rows and columns are the individual actual atoms of the molecular formula. The matrix describes all possible bonding relationships. Initial substructural representation is created using the dual output of a program, INTERPRET, which receives as input, the molecular formula of the unknown and the one- and two-dimensional NMR spectral data. The objective of HOUDINI program is to determine all possible ways in which the atoms of the substructural inferences can be mapped to the actual atoms of the molecular formula such that the integrity of all of the information content is preserved. This program is not addressed to configurational generation, symmetry search, or to avoiding generation of *meso* structures representing redundant structures. *Meso* structures can be defined as chiral molecules having more than one stereocenter and a symmetry plane.

A different approach has been developed in our laboratory before, using a 2-D molecular representation named "*N*_tuple" that can be considered externally as a list or string and internally as a tree.[15−17]

For solving some stereoisomer generation problem we have used molecular graphs processing of increasing complexity, first for acyclic structures[44,45] and then for cyclic ones with isolated and spiro cycles.[46] The approach now is to solve stereoisomer generation for molecular structures containing cycles of whatever type (isolated, spiro, condensed, and nested). An important challenge here for avoiding redundant generation is symmetry detection. One of the pioneers in this field was the work of Nourse et al. (1979)[6] where concept of automorphism groups and construction of configurational symmetry sets were proposed. Nourse defines two standard groups that are used to describe the symmetry of a graph: the node symmetry group and the edge symmetry group. The node symmetry group being the group of all one-to-one mapping of the nodes of the graph onto themselves, which preserves connectivity of the graph. The edge symmetry group is the group of all one-to-one mapping of the edges, which preserves connectivity. A graph symmetry group needed for the discussed chemical problems contains elements that are in both the node and edge groups and is a product of two groups: (i) the node symmetry group mentioned above and (ii) the group of all edge permutations that interchanges the two edges corresponding to a double bond. This graph symmetry group is by definition represented by its action on the configurations of all stereocenters. Nourse defines stereocenters to be any nontriply bonded tetravalent or trivalent atom that has at most one hydrogen substituent. The configuration at a stereocenter is determined by the numbering associated with the attached atoms. Graph symmetry group allows for the construction of a configuration

† Dedicated to Professor Johann Gasteiger.

* Corresponding author phone: (562) 776 2870; e-mail: lcontrer@usach.cl.
‡ Department of Chemical Sciences.
§ Department of Information Technology.

10.1021/ci6002762 CCC: $33.50
Published on Web 10/27/2006

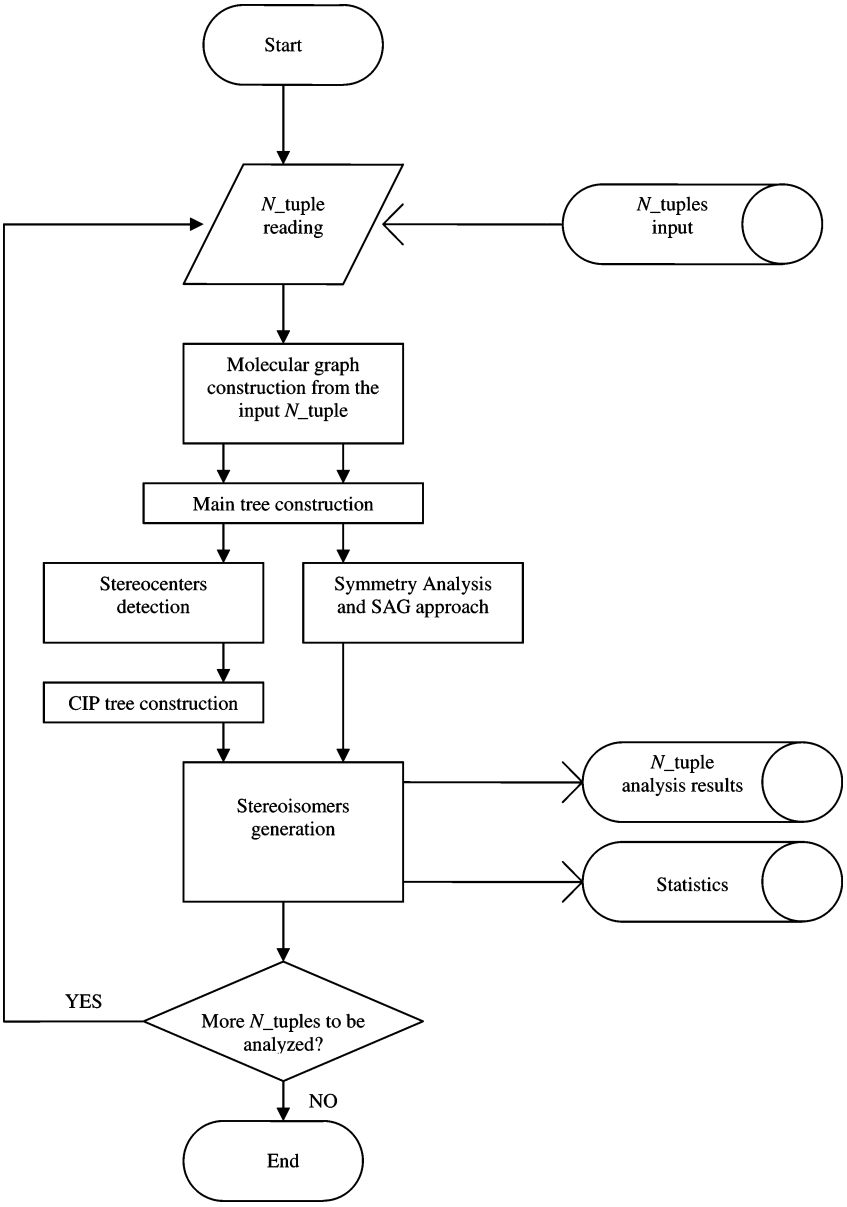


Figure 1. General diagram of the CAMGEC2 system.

symmetry group (CSG) in such a way that an element in the CSG is simply the permutation in the graph symmetry group augmented with superscripts (‘) for the stereocenters that are inverted by the permutation. CSG can be used to generate all distinct stereoisomers of a given chemical structure of specified constitution. Each distinct stereoisomer corresponds to an equivalence class in the set of all 2^n possible stereoisomers, where n is the number of stereocenters.

Nourse ideas were further developed under a mathematical perspective in the construction of MOLGEN program.[7,12,42] Using the chemical formula together with some prescribed and some forbidden substructures as the input, MOLGEN generates the following *output:* the complete list of all the mathematically possible molecular graphs that are compatible with the chemical formula (i.e. the vertices are labeled with the element names, they have the prescribed valences, and the graph is connected). The program considers symmetry and shows global results; however, a particular analysis of the atoms of the graph, for instance the number of atoms that could have identical properties, is not done.

In this work a simplified heuristics (without using permutations) is used for addressing these points in a program implementing these ideas: CAMGEC2 (Computer Assisted Molecular Generation & Enumeration of Compounds).

From a computing point of view we consider molecular graphs as models of organic molecules as an approach based on graph theory[47] where molecular graphs are connected and nondirected graphs (hydrogen atoms suppressed). Basic concepts associated are topology and topography. Topology considers the set of neighbors of any graph node (order does not matter) and topography accounts for sequences of neighbors around topographical centers called stereocenters (order does matter).

Global generation process in general takes into account the following steps: internal graph representation and weighted bipartite tree construction, stereocenter atoms and topological symmetry detection, and a nonredundant generation. The term bipartite refers to the inclusion of both atoms and bonds as nodes of the tree. However both stereocenter detection and topological symmetry detection require establishing of topological and even topographical ligand differ-

Chart 1. Bipartite Tree Making Process and Weight Assignment Algorithm

```
// Atoms and/or bonds will be called elements and/or nodes

      Put element as the tree's root.

      Put all of its neighbours as its sons.

      For each son,

            While it has any neighbour different from its father,

                  Put all of its neighbours (different from its father)

                  as sons.

      End for

      //calculate and assign weights starting from the leaves of the tree

      //(that initially have a weight zero) until arriving to the root.

      //Weight(element) for a node_atom will be the atomic number; for a

      //node_bond the weight(element) will be the type of bond

      //(numerically assigned as 1, 2 or 3 for single, double or triple

      //bonds, respectively, and 1.5 for aromatic system bonds).

      For all of the leave nodes,

            Weight = weight(element) x level

      For all the fathers whereas father is not the root,

Weight = (Addition(sons' weight) + weight(element)) x level.
```

ences. Ligand is referred to the entire group attached to the node in question, not just to the atoms bonded directly to that node. If the root of the tree is a bond, then the number of ligands (or branches) will be two. If the root is an atom, then there will be as many ligands as non-hydrogen atoms or atom groups attached to that atom according to its valence and hybridization.

This new approach uses a weighted bipartite tree making process wherein each atom and bond of the molecule are used as the root, and an algorithm for symmetry detection that allows for grouping all of the atoms of the structure having identical properties in the so-called "symmetry atom groups" (SAG(s)). Definition and analysis of specific parameters assigned to SAG(s) containing stereocenters is a key point for detecting redundant structures prospectively, i.e., before they can be generated avoiding in this way its formation before the generation process takes place thus allowing for a better performance of the system.

Generated structures are written as extended *N*_tuples wherein for components forming part of a stereocenter, priorities of its corresponding ligands are written in a certain particular order that considers both Cahn, Ingold, and Prelog priority rules[48,49] and the kind of stereoisomer being represented. This information is not provided in any other single string molecular representation. The program was developed using Java. A number of examples are presented for a variety of structures to show some of the program capabilities.

2. INTERNAL GRAPH AND TREE CONSTRUCTION

Molecular graph linearly represented as a raw *N*_tuple is read, and an internal molecular graph construction, together with the distance and adjacency matrices and a list of atoms and bonds needed for a bipartite tree construction and weight assignment are done. Figure 1 illustrates the general diagram of CAMGEC2.

Every atom and bond is independently chosen each time as the root of a tree in such a way that as many trees as atoms and bonds are present in the structure will be created. To simplify the algorithm and since is not really needed for detecting stereoisomers or symmetry properties, allene sp carbon atoms (i.e. nodes of the central part of a double bond chain of even length) are excluded as a bipartite tree root. The tree construction and the weight assignment algorithms are done as is shown in Chart 1.

In this way, the tree making process allows for the obtention of the weight of each of the branches (or ligands) coming from the root node (atom or bond) which in turn allows for determination of stereocenters occurrence and the possibility of finding symmetry planes in the molecule. Figure 2 gives an example of a **Z** molecule ($C_3H_5N_1O_1$) and its weighted bipartite tree.

3. STEREOGENIC CENTER DETECTION

Stereocenter and symmetry determination is done using the main bipartite tree representation of a molecular graph for ligands comparison and for establishing any difference among them. For stereocenters determination, the following considerations apply:

- Two kind of centers will be considered: single carbon atoms and chains of cumulated double bonded carbon atoms whose terminal carbon atoms (called chain's head atoms) could have one or two single bonded ligands. A carbon atom

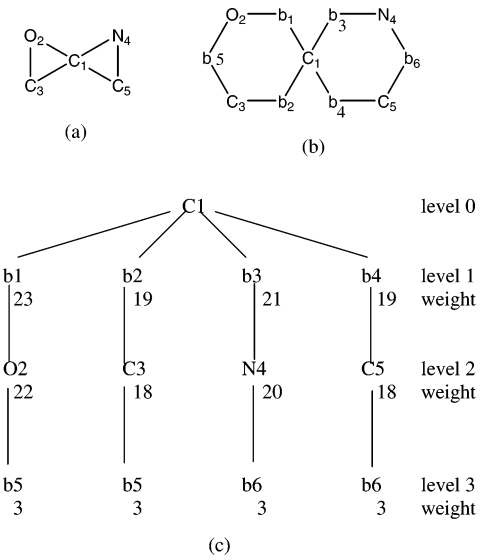


Figure 2. Representation of a **Z** molecule ($C_3H_5N_1O_1$). (a) Molecular graph of a **Z** molecule; (b) modified molecular graph including bonds; and (c) weighted bipartite tree for a **Z** molecule. Ligands b2 and b4 have the same weight.

(chiral center) could generate optical isomers having configuration *R/S* when its four substituents are different (at least topologically). A chain of double bonded carbon atoms could generate geometrical isomers having *Z/E* or *M/P* configuration when substituents at each of the chain's head atoms are different. Configuration *Z/E* could occur if the chain contains an odd number of double bonds (alkene-like compounds) and on the contrary *M/P* configurations could occur (allene-like compounds).

- Cyclic monoalkenes having less than eight members in the cycle cannot generate geometrical isomers, neither can cyclic monoallenes having less than 12 members in the cycle.[50] Cyclic dialkenes having less than 12 members in the cycle can generate *Z,Z* or *E,E*, isomers but not a *Z,E* isomer. Cyclic diallenes with less than 12 members in the cycle can generate *M,M* or *P,P* isomers but not *M,P* isomers, due to steric constrains.

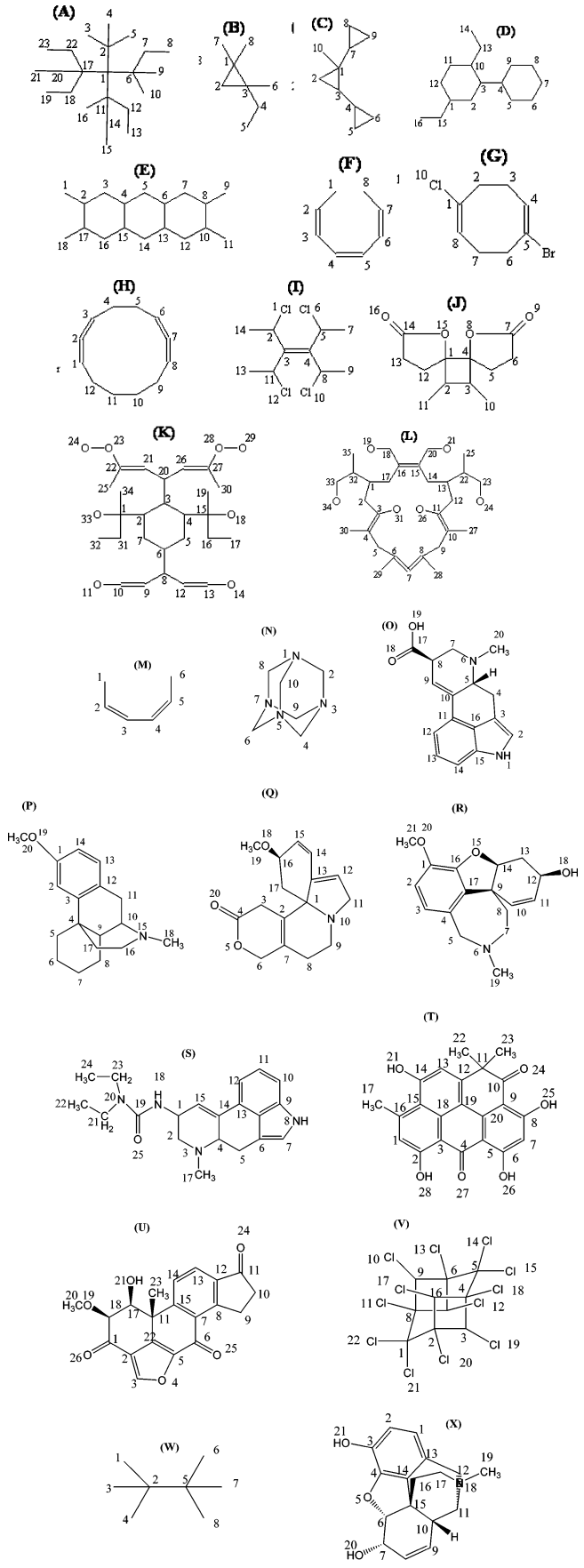
Algorithm for stereocenters detection using these basic rules is outlined schematically in Figure 3. In this diagram, at R1, an atom to be considered as having *R/S* configuration should meet three conditions: (i) being tetrahedral, (ii) having a number of sons bigger than 2, and (iii) not being analyzed before. Similarly, at R3 the atom being a *Z/E* or *M/P* stereocenter type should be a trigonal one having more than 1 son and not being analyzed before. At R5, considerations about cyclic compounds are taken into account as was explained above. At R4, it is determined if groups attached to the sp^2 carbons differ from each other, and at R6 the length of the double bond chain determines the occurrence of *Z/E* or *M/P* stereocenter type.

4. SYMMETRY ANALYSIS

Symmetry analysis first establishes if a symmetry plane passes through a main tree root node,[51] and, if it does, then the node is considered as a symmetry center, and therefore the corresponding symmetry pairs are determined. In that case the following considerations apply:

(i) Symmetry Centers. One bipartite tree root node (atom or bond) can be a symmetry center if (a) is a bond having both substituents topologically equal, i.e., a perpendicular symmetry plane exists (plane is cutting it), (b) is a tetrahedral carbon atom with at least two equal ligands, and (c) is a sp^2

Chart 2. Sample of Structures Processed by CAMGEC2



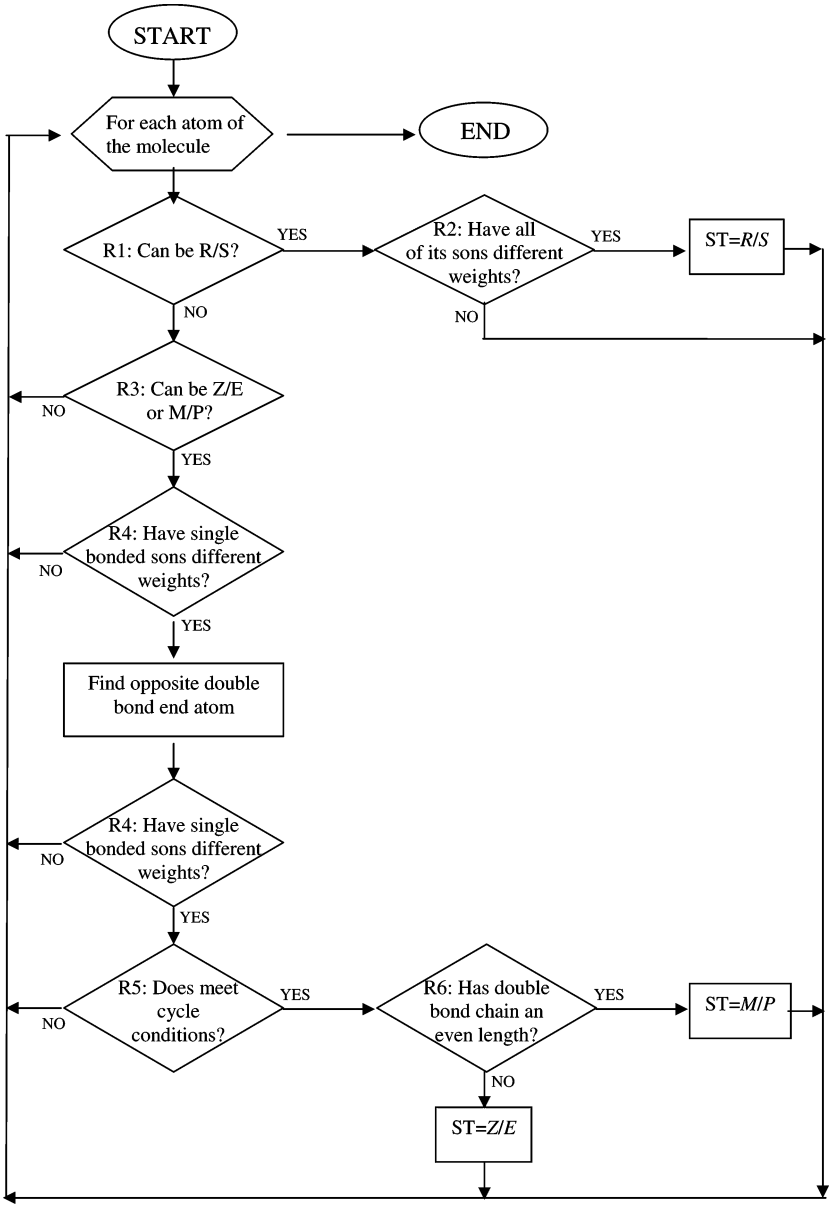


Figure 3. General diagram for stereocenter type (ST) occurrence detection.

carbon atom (alkene compound) with both single bonded ligands being equal as long as at the other end of the odd double bond chain both substituents are also equal (axial symmetry).

(ii) Symmetry Planes. When two branches in the weighted bipartite tree are identical, it is assumed that a symmetry plane occurs which either contains the root if it is an atom or cuts the root if it is a bond.

Also, symmetry plane defines symmetry pairs: two atoms having the same properties (type of atom, number of sons, distance to the root, chirality, and so on). A symmetry pair (not ordered) of atoms A and B is denoted as (A,B), and symmetry planes are represented by a list of symmetry pairs (separated by a semicolon). If a symmetry plane passes just over an atom, its identification number is written twice in brackets with a colon between them. For instance the (**E**) molecule (given in Chart 2) has the following two symmetry planes:

P1:{(1,9);(2,8);(3,7);(4,6);(5,5);(10,17);(11,18);(12,16);(13,15);(14,14)}

P2:{(1,18);(2,17);(3,16);(4,15);(5,14);(6,13);(7,12);(8,10);(9,11)}

(iii) Symmetry Atom Groups. For a set of symmetry planes, we call "symmetry atom group" (SAG) the set of atoms in the transitive closure of the symmetry pairs over the set of planes. That means, if (A,B) and (B,C) are symmetry pairs, then atoms {A,B,C} form a symmetry atom group and have common properties. SAG(s) are used for detecting *meso* isomers.

There are molecular structures that present many symmetry planes which do not add relevant information; for instance, molecule (**W**) in Chart 2 has 15 symmetry planes due to the presence of the single bond between atoms 2 and 5 and the possibility of free rotation. However for getting information related to the atoms having identical properties the system uses only 6 of these planes (see Table 1). Analysis of symmetry planes in molecule (**W**) allows for determination

Table 1. Results for Structures in Chart 2 from CAMGEC2

struct	stereocenters [atoms] (type)	ligands in decreasing CIP priority order	symmetry planes	symmetry atom groups	meso forms	total stereo-isomers
A	[1](R/S)	17,11,6,2	no	no	0	2
B	[3] (R/S)	1,2,4,6	no	no	0	2
C	[1] (R/S) [3] (R/S)	3,7,2,10 1,4,2,0	no	no	0	4
D	[1] (R/S) [3] (R/S) [10] (R/S)	2,12,15,0 4,10,2,0 3,11,13,0	no	no	0	8
E	[2] (R/S) [4] (R/S) [6] (R/S) [8] (R/S) [10] (R/S) [13] (R/S) [15] (R/S) [17] (R/S)	17,3,1,0 15,5,3,0 13,5,7,0 10,7,9,0 8,12,11,0 6,14,12,0 4,14,16,0 2,16,18,0	P1:{(1,9);(2,8); (3,7);(4,6);(5,5);(10,17); (11,18); (12,16);(13,15);(14,14)} P2:{(1,18);(2,17);(3,16);(4,15);(5,14); (6,13);(7,12);(8,10);(9,11)}	G1: {1,9,18,11} G2: {2,8,17,10} G3: {3,7,16,12} G4: {4,6,15,13} G5: {5,14}	182	74
F	[2,3] (Z/E) [4,5] (Z/E) [6,7] (Z/E) 	[2] 1,0 [3] 4,0 [4] 3,0 [5] 6,0 [6] 5,0 [7] 8,0	P1: {(1,8);(2,7);(3,6);(4,5)}	G1: {1,8} G2: {2,7} G3: {3,6} G4: {4,5}	2	6
G	[1,8] (Z/E) [4,5] (Z/E) 	[1] 10,2 [8] 7,0 [4] 3,0 [5] 9,6	no	no	0	2
H	[1,3] (M/P) [6,8] (M/P) 	[1] 12,0 [3] 4,0 [6] 5,0 [8] 9,0	P1: {(1,8); (2,7); (3,6); (4,5); (9,12); (10,11)}	G1: {1,8}, G2: {2,7} G3: {3,6}, G4: {4,5} G5: {9,12}, G6: {10,11}	0	2
I	[2] (R/S) [5] (R/S) [8] (R/S) [11] (R/S)	1,3,14,0 6,4,7,0 10,4,9,0 12,3,13,0	P1: {(1,12); (2,11); (3,3); (4,4); (5,8); (6,10); (7,9);(13,14)} P2: {(1,6);(2,5);(3,4);(7,14);(8,11); (9,13); (10,12)}	G1: 1,6,10,12 G2: 2,5,8,11 G3: 3,4 G4: 7,9,13,14	10	6
J	[1] (R/S) [2] (R/S) [3] (R/S) [4] (R/S)	15,4,2,12 1,3,11,0 4,2,10,0 8,1,3,5	P1:{(1,4);(2,3);(5,12);(6,13);(7,14);(8,15); (9,16);(10,11)}	G1: {1,4}, G2: {2,3} G3: {5,12}, G4: {6,13} G5: {7,14}, G6: {8,15} G7: {9,16}, G8: {10,11}	6	10
K	[1] (R/S) [2] (R/S) [4] (R/S) [9,10] (Z/E) [12,13] (Z/E) [15] (R/S) [21,22] (Z/E) [26,27] (Z/E) 	33,2,31,34 1,3,7,0 15,3,5,0 [9] 8,0 [10] 11,0 [12] 8,0 [13] 14,0 18,4,16,19 [21] 20,0 [22] 23,25 [26] 20,0 [27] 28,30	P1: {(1,15);(2,4);(3,3);(5,7);(6,6);(8,8); (9,12);(10,13);(11,14);(16,31); (17,32); (18,33);(19,34);(20,20); (21,26); (22,27); (23,28);(24,29);(25,30)}	G1: {1,15}, G2: {2,4} G3: {5,7}, G4: {9,12} G5: {10,13}, G6: {11,14} G7: {16,31}, G8: {17,32} G9: {18,33}, G10: {19,34} G11: {21,26}, G12: {22,27} G13: {23,28}, G14: {24,29} G15: {25,30}	166	90
L	[1] (R/S) [3,4] (Z/E) [6,8] (M/P) [10,11] (Z/E) [13] (R/S) [15,16] (Z/E) [22] (R/S) [32] (R/S)	32,2,17,0 [3] 31,2 [4] 5,30 [6] 5,29 [8] 9,28 [10] 9,27 [11] 26,12 22,12,14,0 [15] 20,14 [16] 18,17 23,13,25,0 33,1,35,0	no	no	0	256
M	[2,3] (Z/E) [4,5] (Z/E)	[2] 1,0 [3] 4,0 [4] 3,0 [5] 6,0	P1:{(1,6);(2,5);(3,4)}	G1:{1,6} G2:{2,5} G3:{3,4}	1	3
N	no		P1:{(1,1);(2,8);(3,7);(4,6);(5,5);(9,9);(10,10)} P2:{(1,1);(2,10);(3,5);(4,4);(6,9);(7,7);(8,8)} P3:{(1,1);(2,2);(3,3);(4,9);(5,7);(6,6);(8,10)} P4:{(1,5);(2,4);(3,3);(6,8);(7,7);(9,9);(10,10)} P5:{(1,7);(2,9);(3,3);(4,4);(5,5);(6,10);(8,8)} P6:(1,3);(2,2);(4,10);(5,5);(6,6);(7,7);(8,9)}	G1:{1,5,7,3} G2:{2,8,10,4,9,6}	0	1

Table 1 (Continued)

struct	stereocenters [atoms] (type)	ligands in decreasing CIP priority order	symmetry planes	symmetry atom groups	meso forms	total stereo-isomers
O	[5] (R/S)	6,10,4,0	no	no	0	4
	[8] (R/S)	17,7,9,0				
P	[4] (R/S)	9,3,17,5	no	no	0	8
	[9] (R/S)	10,4,8,0				
	[10] (R/S)	15,9,11,0	1			
Q	[1] (R/S)	10,13,2,17	no	no	0	4
	[16] (R/S)	18,15,17,0				
R	[9] (R/S)	14,17,10,8	no	no	0	8
	[12] (R/S)	18,11,13,0	1			
	[14] (R/S)	15,9,13,0				
S	[1] (R/S)	18,2,15,0	no	no	0	4
	[4] (R/S)	3,14,5,0				
T	no		P1:{(1,1); (2,2); (3,3); (4,4); (5,5); (6,6); (7,7); (8,8);(9,9);(10,10);(11,11);(12,12); (13,13); (14,14); (15,15);(16,16);(17,17); (18,18); (19,19);(20,20);(21,21); (22,23); (24,24); (25,25);(26,26);(27,27);(28,28)}	G1:{ 22,23}	0	1
U	[16] (R/S)	17,22,15,23	no	no	0	8
	[17] (R/S)					
	[18] (R/S)	19,1,17,0	19,1,17,0	19,1,17,0	19,1,17,0	
V	[2] (R/S)	20,1,3,16	P1:{(1,1);(2,8);(3,7);(4,6);(5,5);(9,16); (10,17);(11,20);(12,19);(13,18);(14,15); (21,21); (22,22)}	G1:{2,8}, G2:{3,7} G3:{4,6}, G4:{9,16} G5:{10,17}, G6:{11,20} G7:{12,19}, G8:{13,18} G9:{14,15}	6	10
	[4] (R/S)	18,5,3,16				
	[6] (R/S)	13,5,7,9				
	[8] (R/S)	11,1,7,9				
W	no	no	P1:{(1,1);(2,2);(3,4);(5,5);(6,7);(8,8)} P2:{(1,3);(2,2);(4,4);(5,5);(6,7);(8,8)} P3:{(1,4);(2,2);(3,3);(5,5);(6,7);(8,8)} P4:{(1,1);(2,2);(3,4);(5,5);(6,8);(7,7)} P5:{(1,1);(2,2);(3,4);(5,5);(6,6);(7,8)} P6:{(1,8);(2,5);(3,6);(4,7)}	G1: {1,3,4,6,7,8} G2:{2,5}	0	1
X	[6] (R/S)	5,7,15,0	no	no	0	32
	[7] (R/S)	20,6,8,0				
	[10] (R/S)	11,15,9,0				
	[11] (R/S)	18,10,12,0				
	[15] (R/S)	6,14,10,16				

of two SAG(s), G1:{1,3,4,6,7,8} and G2:{2,5}. That means atoms 1, 3, 4, 6, 7, and 8 have the same properties among them and different from atoms 2 and 5 that constitute another group. In Table 1 symmetry planes are specified for the generated isomers (topological, not topographical symmetry). For instance symmetry plane P1 of (**F**) molecule is valid for the 2*Z*,4*Z*,6*Z* isomer but also for the 2*Z*,4*E*,6*Z* isomer.

5. STEREOISOMER GENERATION

Once stereogenic centers or stereocenters are detected, a basic problem in generation is redundancy avoidance. When symmetry does not exist, for n independent stereocenters, 2^n stereoisomers can be generated just by independent combinations of configurations. In the presence of symmetry this amount is reduced due to *meso* compounds. For addressing this point SAG(s) containing any stereocenter should be analyzed.

Let us consider a vector, a set, or a group of n stereocenters of the graph (stereocenter is a single atom in the case of *R/S* configurations and a pair of terminal atoms belonging to a chain of cumulated double bonds for *Z/E* and *M/P* configurations). Let us call *R*, *Z*, and *M* "inactive" configurations (denoted by 0) and *S*, *E*, and *P* "active" configurations (denoted by 1). The generation algorithm should generate all 2^n binary combinations (words), and for redundancy detection, SAG(s) are taken into account. Note that when a SAG contains an atom being a stereocenter, all remaining atoms in the SAG will be stereocenters too and conversely no one will be.

Redundancy detection is done in the following way: given a binary combination corresponding to the stereocenter set of the structure, an "identification value" for each atom in each SAG containing an stereocenter is calculated, and for each SAG a set of four "identification parameters" are also calculated. Only SAG(s) containing stereocenters are considered in this procedure.

Computation of the "identification value" V_i for an atom i of any SAG is done by adding up (for the m SAG elements) the multiplication of the binary value of that stereocenter atom with the distance to each of the other SAG atoms in the molecular graph as represented by eq 1

$$\mathbf{V}_i = \sum_{j=1}^{m} b \times D_{ij} \tag{1}$$

wherein b is the binary value of configuration (0: inactive, or 1: active, which obviously has to be the same value for both terminal atoms in a cumulated double bond chain), and D_{ij} is the distance in the molecular graph (kept in a distance matrix) of atom i to each atom j of the SAG.

SAG "identification parameters" are specified as follows: active major value (AMV), total active value (TAV), inactive

	1	2	3	4	5	6
1	0	1	2	3	4	5
2	1	0	1	2	3	4
3	2	1	0	1	2	3
4	3	2	1	0	1	2
5	4	3	2	1	0	1
6	5	4	3	2	1	0

Figure 4. Distance matrix for molecule **M**.

Table 2. Identification Values for SAG Atoms Corresponding to Each Binary Combination According to the Stereocenter Set

binary combinations [(2,3),(4,5)]	G2 identification values		G3 identification values	
	atom 2	atom 5	atom 3	atom 4
0,0	0 + 0 = 0	0 + 0 = 0	0 + 0 = 0	0 + 0 = 0
0,1	0 + 3 = 3	0 + 0 = **0**	0 + 1 = 1	0 + 0 = **0**
1,0	0 + 0 = **0**	3 + 0 = 3	0 + 0 = **0**	1 + 0 = 0
1,1	0 + 3 = **3**	3 + 0 = **3**	0 + 1 = **1**	1 + 0 = **1**

major value (IMV), and total inactive value (TIV) which are calculated as

$$\mathrm{AMV} = \max\{V_i\} \quad \text{for } i \text{ such that } b = 1 \tag{2}$$

$$\mathrm{TAV} = \sum_{i=1}^{m} V_i \quad \text{for } i \text{ such that } b = 1 \tag{3}$$

$$\mathrm{IMV} = \max\{V_i\} \quad \text{for } i \text{ such that } b = 0 \tag{4}$$

$$\mathrm{TIV} = \sum_{i=1}^{m} V_i \quad \text{for } i \text{ such that } b = 0 \tag{5}$$

Redundancy detection is based on the following criteria: if all identificacion parameter values for all SAG(s) of a particular binary configuration are identical with values for other already generated configurations, then the actual configuration is a *meso* form of the one previously generated.

As an example, molecule (**M**) in Chart 2 has three symmetry groups G1:{1,6}; G2:{2,5}; and G3:{3,4} (Table 1). Tree processing of this molecule allows for the recognition of two stereocenters: double bonds at atom pairs (2,3) and (4,5), respectively, and accordingly there will be four *Z/E* stereoisomers with binary configurations: [0,0]; [0,1]; [1,0]; and [1,1]. Note that stereocenter atom number 2 has the same configuration as stereocenter atom number 3; the same applies for atoms 4 and 5. The necessary distances are obtained from a distance matrix given in Figure 4.

Table 2 shows "identification values" for each element of both SAG(s) (active values are written in bold). Only G2 and G3 are considered since G1 contains atoms that are not stereocenters.

The four "identification parameters" related to each SAG within each binary combination are presented in Table 3.

It can be seen that both SAG(s) "identification parameters" for G2 and G3 corresponding to configurations [0,1] and [1,0] are identical to each other, having values (0,0,3,3) and (0,0,1,1), respectively, so *E,Z* and *Z,E* configurations are redundant. Then, isomers to be generated are [*Z,Z*], [*Z,E*], and [*E,E*]. The [*E,Z*] isomer is a *meso* compound, and it will not be generated.

6. REPRESENTATION OF THE FINAL EXTENDED *N*_TUPLES

Molecular graphs containing stereocenter configurations are represented in the *N*_tuple linear format. These are written starting from the initial topological *N*_tuple[17] and adding for each atom characterized as stereocenter an appropriate extension that contains the following: configuration symbol plus identification numbers of substituents ordered upon Cahn, Ingold, and Prelog (CIP) priority rules.[48,49] Priority computation is done through a second tree construction process in which root atoms are stereocenters only, nodes are only atoms, and all multiple bonds are transformed correspondingly to single bonds (i.e., a carbonyl group is counted as two single C−O bonds). Then comparison of the atomic number is done level by level, atom by atom until a difference is found. A bigger atomic number means a higher priority.

Procedure for writing down the extension at the *N*_tuple is done by the algorithm given in Chart 3.

For instance for molecule (**M**) the input *N*_tuple is 1c1r 2c1s 3c1d 4c1s 5c1d 6c0s. The corresponding three output extended *N*_tuples are

1c1r 2c1s(Z{3}[1,0]) 3c1d(Z{2}[4,0]) 4c1s(Z{5}[3,0]) 5c1d(Z{4}[6,0]) 6c0s

1c1r 2c1s(Z{3}[1,0]) 3c1d(Z{2}[4,0]) 4c1s(E{5}[0,3]) 5c1d(E{4}[0,6]) 6c0s

1c1r 2c1s(E{3}[0,1]) 3c1d(E{2}[0,4]) 4c1s(E{5}[0,3]) 5c1d(E{4}[0,6]) 6c0s

7. RESULTS AND DISCUSSION

The CAMGEC2 program was developed in Java language. The main new features were focused to (i) determine SAG(s) of the molecule containing stereocenters; (ii) characterize them through the use of specific parameters that allow for a prospective detection of redundant structures; and (iii) write a new molecular representation in the form of an extended *N*_tuple that considers both atom configurational properties and ordering of its neighbors upon CIP priority rules. As was already described SAG(s) are groups of atoms that form part of the symmetry pairs associated with molecule symmetry planes. For instance, in Table 1, **A**, **B**, **C**, and **D** structures do not have any SAG. On the contrary, **E** structure has five SAG(s) due to the presence of two symmetry planes. SAG(s) G1, G3, and G5 do not contain any stereocenter; however, SAG(s) G2 and G4 contains four stereocenters with each one shown in the second column of Table 1. The SAG(s) concept is very different than configurational symmetry set (explained in the Introduction above) and equivalence class[6] concepts, the equivalence class being the different generated final structures as a whole and not portions of them as SAG(s) are.

Complete processing using a file having 53 *N*_tuples as input wherein each structure has between 5 and 35 nonhydrogen atoms (with an average of 12 atoms) takes less than 4 s on a Pentium 4 with 1 GB RAM under Windows

Table 3. SAG(s) Identification Parameters for the Stereocenter Binary Combinations

binary combination [(2,3),(4,5)]	stereocenter configuration [(2,3),(4,5)]	G2 identification parameters				G3 identification parameters			
		AMV	TAV	IMV	TIV	AMV	TAV	IMV	TIV
0,0	Z,Z	0	0	0	0	0	0	0	0
0,1	Z,E	0	0	3	3	0	0	1	1
1,0	E,Z	0	0	3	3	0	0	1	1
1,1	E,E	3	6	0	0	1	2	0	0

Chart 3. Algorithm for Writing Extended N_Tuple

```
//Hydrogen identifier is "0".
For each stereocenter atom,
```

- If the stereocenter atom is *R/S* type, then write accordingly at the end of the N_tuple component corresponding to stereogenic node:
  - character "*R*" and a list of integer identifiers for the 4 atoms bound to the stereocenter in decreasing order upon priorities
  - character "S" and a list of integer identifiers for the 4 atoms bound to the stereocenter being first two identifiers exchanged
- If the sterocenter atom is *Z/E* or *M/P* type, the following should be added:
  - Spatial configuration as a character '*Z*' or '*M*'
  - Integer identifier of the pair atom located at the other end of the double bond chain.
  - A list of two substituent atom identifiers ordered upon decreasing CIP priorities in the case of '*Z*' or '*M*' isomers. For the other configuration write character '*E*' or '*P*' and identifiers in that interchanged atom.

XP operating system. CAMGEC2 provides outputs to text format file containing the information associated with both the molecule analysis (i.e., the identification number and the type of configuration of the atoms being stereocenters, its ligands written in a decreasing order according to CIP priority rules, symmetry planes, SAG(s), number of *meso* forms, and total number of generated stereoisomers) and the system performance (time process for each analyzed *N*_tuple and for the total number of *N*_tuples in the file are written down). No *para*-stereocenters (stereocenters with dependent or interdependent configurations) nor neither any heteroatom as a chiral center are taken into account. Labeled compounds are treated without any problem provided that respective isotopes are defined at the atom.cfg file.

Chart 2 shows a representative sample of 24 molecules of different structures analyzed by CAMGEC2 whose results are presented in Table 1. Total processing time was 4.58 s. Molecules in Chart 2 have between 6 and 35 atoms (with an average of 28.6 atoms). Some of the structures have hydrocarbon skeletons (i.e., **A**−**F**, **H**, **M**, and **W** molecules); others contain heteroatoms (i.e., **G**, **I**−**L**, and **N**−**V** molecules) or contain different kinds of cycles (**B**−**E**, **J**, and **N**−**V** molecules), and others contain cyclic dienes and diallenes (i.e., **G**, **H**, and **L** molecules). These samples also include structures having interesting properties such as the molecules (**N**) (methenamine, a urinary antiseptic); (**O**) (lysergic acid, a psychomimetic); (**P**) (racemethorphan, an antitussive); (**Q**) (β-erythroidine, a skeletal muscle relaxant); (**R**) (galanthamine, a cholinesterase inhibitor); (**S**) (lysuride, a serotonin inhibitor); (**T**) (resistomycin, an antibacterial); (**U**) (viridin, an antifungal); (**V**) (mirex, an insecticide and fire retardant); and (**X**) (morphine, a narcotic analgesic) which present many structural features that can be easily analyzed by CAMGEC2. Actually, in Table 1 results about stereocenter identification, CIP priorities assigned to substituents of the recognized stereocenters, symmetry planes, symmetry atom groups, number of *meso* compounds, and stereoisomer generation (without redundancy) were correctly assessed. For a better understanding some examples will be explained in detail according to Chart 2 and Table 1.

Molecule (**A**) is an acyclic molecule having one chiral stereocenter (atom 1 bonded to atoms (17, 11, 6, 2) written in a decreasing order according to CIP rules) and no symmetry planes. Application of CIP priority rules allows for finding a difference in stereocenter ligands at the third level of comparison.

Molecule (**L**) is a cyclic molecule comprising 8 stereocenters and no symmetry plane due to the allene into the cyclic structure. Stereocenters are four chiral centers (atoms 1, 13, 22, and 32), three *Z/E* stereocenters (atoms 3−4, 10−11, and 15−16), and one *M/P* stereocenter (atoms 6−8). The third column in Table 1 shows corresponding substituents ordered upon decreasing CIP priorities. For stereocenter Nr 1 (chiral) the ordering of ligands priorities found by CAMGEC2 is 32 > 2 > 17 > 0. Zero value is assigned to hydrogen. For *Z/E* stereocenter (atoms 3 and 4) ligands priority order for atom 3 are 31 > 2, and for atom 4 substituens (atoms 5 and 30) have priority order 5 > 30.

Molecule (**N**) has a heterocyclic structure with six planes of symmetry, two SAG(s) (one comprised by nitrogen atoms, and the other group comprised by the carbon atoms all of which having the same properties), and no stereocenter.

Molecule (**X**) has a nested polycyclic structure with five chiral stereocenters. Ligands (or substituents) at stereocenter number 15 are 6, 14, 10, and 16 which are written in a decreasing CIP order. This structure has no symmetry plane and therefore has 32 stereoisomers (2^5).

9. CONCLUSIONS

A system, CAMGEC2, has been created, which allows for the exhaustive generation in a nonredundant way of topographical organic stereoisomers that may contain different types of atoms, bonds, and cycles in their structures. The approach used in the actual implementation of CAMGEC2 is more general than a previous approach since it is independent of the number and type of cycles if they are present. A new strategy for detecting topological symmetry planes and stereocenters (*R/S*, *Z/E*, and *M/P*) based on a bipartite tree making process and assignment of weights to all of the tree nodes proves to be an appropriate and efficient algorithm that does not need to perform a previous classification of the actually occurring cycles to define a searching algorithm. It uses a new strategy for detecting *meso* compounds based on the determination of symmetry atom groups which are groups containing atoms having all the same properties. Computation of SAG(s) identification parameters for each binary combination of stereocenter configurations has proved to be efficient for detecting redundant structures (*meso* compounds) at an earlier step before generation takes place so allowing for a better performance of the system. The system determines priorities of the substituents associated with any stereocenter according to CIP priority rules, which is relevant in 3-D stereoisomers representation and analysis of organic structures both in research and organic chemistry teaching.

A new extended *N*_tuple format containing information about the topology and topography of the molecule is provided which allows for a three-dimensional representation of the structure specifying into a single string existence and characteristics of any stereocenter (identification of its position and type) and also of its substituents for which a decreasing order is given according to CIP priority rules.

ACKNOWLEDGMENT

The authors are pleased to acknowledge support of DICYT (Direction of Scientific and Technical Research), University of Santiago de Chile, Project No. 060441CF.

REFERENCES AND NOTES

(1) Faulon, J. L.; Visco, D.; Roe, D. Enumerating Molecules. In *Reviews in Computational Chemistry*; Lipkowitz, K., Ed.; Wiley-VCH: 2005; Vol. 21, pp 209−286.

(2) Carhart, R. E.; Smith, D. H.; Gray, N. A.; Nourse, J. G.; Djerassi, C. GENOA: A Computer Program for Structure Elucidation Utilizing Overlapping and Alternative Substructures. *J. Org. Chem.* **1981**, *46*, 1708−1718.

(3) Gray, N. A. B. *Computer-Assisted Structure Elucidation*; Wiley-Interscience: New York, 1986; pp 312−324.

(4) Funatsu, K.; Miyabayaski, N.; Sasaki, S. Further Development of Structure Generation in the Automated Structure Elucidation System CHEMICS. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 18−28.

(5) Christie, B. D.; Munk, M. E. Structure Generation by Reduction: A New Strategy for Computer-Assisted Structure Elucidation. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 87−93.

(6) Nourse, J. G. The Configuration Symmetry Group and its Application to Stereoisomer Generation, Specification, and Enumeration. *J. Am. Chem. Soc.* **1979**, *101*, 1210−1216.

(7) Nourse, J. G.; Smith, D. H.; Carhart, R. E.; Djerassi, C. Exhaustive Generation of Stereoisomers for Structure Elucidation. *J. Am. Chem. Soc.* **1979**, *101*, 1216−1223.

(8) Benecke, C.; Grund, R.; Hohberger, R.; Kerber, A.; Laue, R.; Wieland, Th. MOLGEN, A Generator of Connectivity Isomers and Stereoisomers for Molecular Structure Elucidation. *Anal. Chim. Acta* **1995**, *314*, 141−147.

(9) Nilakantan, R.; Bauman, N.; Venkataraghavan, R. A Method for Automatic Generation of Novel Chemical Structures and its Potential Applications to Drug Discovery. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 527−530.

(10) Pivina, T. S.; Molchanova, M. S.; Shcherbukhin, V. V.; Zefirov, N. S. Computer Generation of Caged Frameworks Which Can be Used as Synthons for Creating High-Energetic Materials. *Propell., Expl., Pyrotech.* **1994**, *19*, 286−289.

(11) Pivina, T. S.; Shcherbukhin, V. V.; Molchanova, M. S.; Zefirov, N. S. Computer-Assisted Prediction of Novel Target High-Energy Compounds. *Propell., Expl., Pyrotech.* **1995**, *20*, 144−146.

(12) Clark, D. E.; Frenkel, D.; Levy, S. A.; Li, J.; Murray, C. W.; Robson, B.; Waszkowycs, B.; Westhead, D. R. PRO_LIGAND: An Approach to *de novo* Molecular Design. 1. Application to the Design of Organic Molecules. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 13−32.

(13) Wieland, T.; Kerber, A.; Laue, R. Principles of the Generation of Constitutional and Configurational Isomers. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 413−419.

(14) Molchanova, M.; Zefirov, N. S. Irredundant Generation of Isomeric Molecular Structures with Some Known Fragments. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 8−22.

(15) Agarwal, K. K. An Algorithm for Computing the Automorphism Group of Organic Structures with Stereochemistry and a Measure of its Efficiency. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 402−404.

(16) Contreras, M. L.; Valdivia, R.; Rozas, R. Exhaustive Generation of Organic Isomers. 1. Acyclic Structures. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 323−330.

(17) Contreras, M. L.; Valdivia, R.; Rozas, R. Exhaustive Generation of Organic Isomers. 2. Cyclic Structures: New Compact Molecular Code. *J. Chem. Inf. Comput. Sci*. **1992**, *32*, 483−491.

(18) Contreras, M. L.; Rozas, R.; Valdivia, R. Exhaustive Generation of Organic Isomers. 3. Acyclic, Cyclic, and Mixed Compounds. *J. Chem. Inf. Comput. Sci*. **1994**, *34*, 610−616.

(19) Luinge, H. J.; Van der Maas, J. H. AEGIS, An Algorithm for the Exhaustive Generation of Irredundant Structures. *Chemom. Intell. Lab. Syst.* **1990**, *8*, 157−165.

(20) Korytko, A.; Schulz, K.-P.; Madison, M. S.; Munk, M. E. HOUDINI: A New Approach to Computer-Based Structure Generation. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (5), 1434−1446.

(21) Cramer, R. D.; Jilek, R. J.; Andrews, K. M. Dbtop: Topomer Similarity Searching of Conventional Databases. *J. Mol. Graphics Modell.* **2002**, *20*, 447−462.

(22) Abrahamian, E.; Fox, P. C.; Narum, L.; Christensen, I. T.; Thogersen, H.; Clark, R. D. Efficient Generation, Storage, and Manipulation of Fully Flexible Pharmacophore Multiplets and Their Use in 3-D Similarity Searching. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (2), 458−468.

(23) Raymond, J. W.; Willett, P. Similarity Searching in Databases of Flexible 3D Structures Using Smoothed Bounded Distance Matrices. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (3), 908−916.

(24) Jilek, R. J.; Cramer, R. D. Topomers: A Validated Protocol for Their Self-Consistent Generation. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (4), 1221−1227.

(25) Milman, B. L. Literature-Based Generation of Hypotheses on Chemical Composition Using Database Co-occurrence of Chemical Compounds. *J. Chem. Inf. Model*. **2005**, *45* (5), 1153−1158.

(26) Irwin, J. J.; Shoichet, B. K. ZINC − A Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model*. **2005**, *45* (1), 177−182.

(27) Hewitt, R.; Gobbi, A.; Lee, M.-L. A Searching and Reporting System for Relational Databases Using a Graph-Based Metadata Representation. *J. Chem. Inf. Model.* **2005**, *45* (4), 863−869.

(28) Prakash, O.; Ghosh, I. Developing an Antituberculosis Compounds Database and Data Mining in the Search of a Motif Responsible for the Activity of a Diverse Class of Antituberculosis Agents. *J. Chem. Inf. Model.* **2006**, *46* (1), 17−23.

(29) Pintore, M.; Wechman, C.; Sicard, G.; Chastrette, M.; Amaury, N.; Chretien, J. R. Comparing the Information Content of Two Large Olfactory Databases. *J. Chem. Inf. Model*. **2006**, *46* (1), 32−38.

(30) Luque Ruiz, I.; Cerruela García, G.; Gómez-Nieto, M. A. Representation of the Molecular Topology of Cyclical Structures by Means of Cycle Graphs. 2. Application to Clustering of Chemical Databases. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (4), 1383−1393.

(31) Luque Ruiz, I.; Cerruela García, G.; Gómez-Nieto, M. A. Representation of the Molecular Topology of Cyclical Structures by Means of Cycle Graphs. 3. Hierarchical Model of Screening of Chemical Databases. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (6), 1903−1911.

(32) Merkwirth, C.; Lengauer, T. Automatic Generation of Complementary Descriptors with Molecular Graph Networks. *J. Chem. Inf. Model.* **2005**, *45* (5), 1159−1168.

(33) Katritzky, A. R.; Perumal, S.; Petrukhin, R.; Kleinpeter, E. CODESSA-Based Theoretical QSPR Model for Hydantoin HPLC-RT Lipophilicities *J. Chem. Inf. Comput. Sci.* **2001**, *41* (3), 569−574.

(34) Randic, M.; Balaban, A. T.; Basak, S. C. On Structural Interpretation of Several Distance Related Topological Indices. *J. Chem. Inf. Comput. Sci.* **2001**, *41* (3), 593−601.

(35) Liu, K.; Feng, J.; Young, S. S. PowerMV: A Software Environment for Molecular Viewing, Descriptor Generation, Data Analysis and Hit Evaluation. *J. Chem. Inf. Model.* **2005**, *45* (2), 515−522.

(36) Cramer, R. D.; Patterson, D. E.; Clark, R. D.; Ferguson, A. M. Method for selecting an optimally diverse library of small molecules based on validated molecular structural descriptors. U.S. Patent 6,185,506, 2001.

(37) Cramer, R. D.; Patterson, D. E. Further method of creating and rapidly searching a virtual library of potential molecules using validated molecular structural descriptors. U.S. Patent 6,240,374, 2001.

(38) Firth-Clark, S.; Willems, H. M. G.; Williams, A.; Harris, W. Generation and Selection of Novel Estrogen Receptor Ligands Using the *De Novo* Structure-Based Design Tool, SkelGen. *J. Chem. Inf. Model.* **2006**, *46* (2), 642−647.

(39) Nidhi; Glick, M.; Davies, J. W.; Jenkins, J. L. Prediction of Biological Targets for Compounds Using Multiple-Category Bayesian Models Trained on Chemogenomics Databases. *J. Chem. Inf. Model.* **2006**, *46*, 1124−1133.

(40) Baurin, N.; Mozziconacci, J.-C.; Arnoult, E.; Chavatte, P.; Marot, C.; Morin-Allory, L. 2D QSAR Consensus Prediction for High-Throughput Virtual Screening. An Application to COX-2 Inhibition Modeling and Screening of the NCI Database. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (1), 276−285.

(41) Shelley, C. A.; Hays, T. R.; Munk, M. E.; Roman, R. V. An Approach to Automated Partial Structure Expansion. *Anal. Chim. Acta* **1978**, *103*, 121−132.

(42) Masinter, L. M.; Shridharan, N. S.; Ledergberg, J.; Smith, D. H. Applications of Artificial Intelligence for Chemical Inference. XII. Exhaustive Generation of Cyclic and Acyclic Isomers. *J. Am. Chem. Soc.* **1974**, *96*, 7702−7714.

(43) Kerber, A.; Laue, R.; Grüner, T.; Meringer, M. MOLGEN 4.0. *Match* **1998**, *37*, 205−208.

(44) Christie, B. D.; Munk, M. E. Structure Generation by Reduction: A New Strategy for Computer-Assisted Structure Elucidation. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 87−93.

(45) Contreras, M. L.; Rozas, R.; Valdivia, R.; Agüero R. Exhaustive Generation of Organic Isomers. 4. Acyclic Stereoisomers with One or More Chiral Carbon Atoms. *J. Chem. Inf. Comput. Sci.* **1995**, *35* (4), 752−758.

(46) Contreras, M. L.; Trevisiol, G. M.; Alvarez, J.; Arias, M. G.; Rozas, R. Exhaustive Generation of Organic Isomers. 5. Unsaturated Optical and Geometrical Stereoisomers, and a New CIP Subrule. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 475−482.

(47) Contreras, M. L.; Alvarez, J.; Riveros, M.; Arias, G.; Rozas, R. Exhaustive Generation of Organic Isomers. 6. Stereoisomers Having Isolated and Spiro Cycles and New Extended *N*_Tuples. *J. Chem. Inf. Comput. Sci.* **2001**, *41* (4), 964−977.

(48) Read, R. C. In *Chemical Applications of Graph Theory*; Balaban, A. T., Ed.; Academic Press: New York, 1976; pp 11−60.

(49) Prelog, V.; Helmchen, G. Basic Principles of the CIP−System and Proposals for a Revision. *Angew. Chem., Int. Ed. Engl.* **1982**, *21*, 567−583.

(50) Cusler, R. H. Mathematical Statements about the Revised CIP-System. *Match* **1986**, *21*, 3−31.

(51) Boss, C.; Keese, R. Diastereoselective Synthesis of Cyclododeca-1,6-diallenes (cyclododeca-l,2,6,7-tetraenes). *Helv. Chim. Acta* **1997**, *80* (2), 414−420.

(52) Contreras, M. L.; et al. Algorithm for Determining Occurrence of Symmetry Planes in Organic Structures. Manuscript in preparation.

CI6002762