ESSESA: An Expert System for Structure Elucidation from Spectra. 4. Canonical Representation of Structures

Hong Huixiao* and Xin Xinquan

Department of Chemistry, Nanjing University, Nanjing 210008, People's Republic of China

Received July 8, 1993[⊗]

LNSCS (*l*inear *n*otation *s*ystem of *c*hemical *s*tructures), a chemical structure representation system, is designed to support the arbitrary notation for structures in ESSESA (*e*xpert *s*ystem for *s*tructure *e*lucidation from *s*pectr*a*). This paper describes a method for the canonical representation of chemical structures, the canonical renumbering of structures, the unique notation and the canonicalization algorithm, and the conversion from LNSCS to the canonical connection table. The first stage of the algorithm is the partitioning of the atoms in a structure, the molecule being treated as a graph with nodes (atoms) and edges (bonds). The partition is done by a product of the extended order value method; the initial order value is generated from the atomic character of the atoms in the structure. The program of converting LNSCS code to the canonical connection table consists of three parts, the compiling procedure of LNSCS and partitioning and generation of the canonical connection table. The algorithm by product of extended order value method is found to be more effective than the Morgan algorithm and other published algorithms.

INTRODUCTION

The chemical structure representation system LNSCS (*l*inear *n*otation *s*ystem of *c*hemical *s*tructures), introduced in a previous paper of this series,[1] is an arbitrary notation system which represents a new approach to computerized chemical nomenclature. For a given chemical structure the LNSCS notation can take many equally valid forms. In ESSESA,[1–3] it is very important to determine whether some library of standard compounds or previously generated candidate structures contains a particular structure as derived by the program. It is therefore necessary to convert the notational forms of such chemical structures to unique representations.

In ESSESA, LNSCS was designed only as an interface between the users and the system. The processing of chemical structures uses the canonical connection table. Atoms of structures in LNSCS are numbered arbitrarily, and canonicalization of the chemical structure representation must be carried out so that the atoms in the structure will be numbered uniquely. Graph theory has become important in chemical information because it provides the basis for codification of nomenclature in chemical computer programs.[4] The classification and ordering of nodes in a "colored" graph is applied here to canonicalization. In chemistry, the colored graph may reflect the structure of a compound, in which the nodes are related to atoms and edges are related to chemical bonds. A structure containing n atoms allows $n!$ possible numberings of the n atoms, and it is thus impractical to rely on a scheme whereby one generates each possible numbering, testing it by comparing its connection table with the best connection table found so far. The problem of devising a reliable canonical method for the numbering of atoms in a structure has attracted the interest of many workers in the field. Many canonicalization algorithms have been published,[5–12] but most of these are elaborations of, and variations on, the algorithm originally outlined by Morgan.[5]

In this paper we describe a program which converts the LNSCS to a canonical connection table. We use the atomic character, a property which is described below and which is a combination of several graph invariants, to obtain the initial order. The initial order and the subsequently generated orders are then used to partition the atoms by the extended order product scores method. This method appears to be more effective than other published methods.

CONNECTION TABLE

The connection table in ESSESA has the form shown in Table 1, which reflects the structural information about the structure, including its topological and stereochemical properties such as the charge, conformation, and atomic connectivity. In the connection table, the neighbors and connections of each atom are described in the brackets by two values. The first of these is the number of the atom to which the atom is connected and the second is the code representing the chemical bond type (1, coordination bond; 2, single bond; 3, double bond; 4, triple bond). Thus, an entry such as [[1,2][3,2]] means connected to atoms 1 and 3 by single bonds. Elsewhere in the table, a zero means the property is zero (e.g. charge) or is not applicable (e.g. configuration, conformation).

ATOMIC CHARACTER

Graph theoretical invariants are properties of graphs that are independent of the way a graph is ordered. For chemical structures atomic invariants are the graph theoretical invariants, and these include connectivities, atomic number, atomic mass, free valencies, charge, configuration, and conformation. The priority of these atomic invariants used in ESSESA is connectivity > free valency > atomic number > atomic mass > charge > configuration > conformation. The priority functions have the form

$$P = P_c P_f P_p P_n P_{ch} P_{cf} P_{cm}$$

The function is an ordered function, generally

$$P_a P_b >< P_b P_a$$

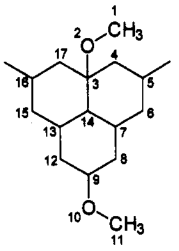
If A_i is an atom in a structure, the atomic character AC_i is generated as follows:

$$\mathrm{AC}_i = P(A_i)$$

[⊗] Abstract published in *Advance ACS Abstracts,* May 15, 1994.

0095-2338/94/1634-0730$04.50/0 © 1994 American Chemical Society

Table 1. Connection Table for Structure 1



Structure 1

LNSCS: COC1,2CC.CC3CC(OC)CC(C2,3)CC.C1

order	atom	free radical	charge	configuration	conformation	connections
1	C	0	0	0	0	[[2,2]]
2	O	0	0	0	0	[[1,2][3,2]]
3	C	0	0	0	0	[[2,2][17,2][14,2][4,2]]
4	C	0	0	0	0	[[3,2][5,2]]
5	C	.	0	0	0	[[4,2][6,2]]
6	C	0	0	0	0	[[5,2][7,2]]
7	C	0	0	0	0	[[6,2][14,2][8,2]]
8	C	0	0	0	0	[[7,2][9,2]]
9	C	0	0	0	0	[[8,2][10,2][12,2]]
10	O	0	0	0	0	[[9,2][11,2]]
11	C	0	0	0	0	[[10,2]]
12	C	0	0	0	0	[[9,2][13,2]]
13	C	0	0	0	0	[[12,2][14,2][15,2]]
14	C	0	0	0	0	[[3,2][7,2][13,2]]
15	C	0	0	0	0	[[13,2][16,2]]
16	C	.	0	0	0	[[15,2][17,2]]
17	C	0	0	0	0	[[3,2][16,2]]

The priority functions are defined here and used to generate the atomic character. In ESSESA, the atomic character is represented as a nine-digit integer. The digits are each assigned values, as below.

9	8	7	6	5	4	3	2	1

digit 1 value of $P_{cm}(A_i)$:
1 = an axial bond; 2 = an equatorial bond

digit 2 value of $P_{cf}(A_i)$: 1 = *R* or *E*; 2 = *S* or *Z*

digit 3 value of $P_{ch}(A_i)$: charge +5

digits 4–7 value of $P_n[Pp(A_i)]$:
(square of atomic no.) + (no. of neutrons)

digit 9 value of $P_c(A_i) = (\mathrm{P1} + \mathrm{P2} + \mathrm{P4})(A_i) =$
$\mathrm{P1}(A_i) + \mathrm{P2}(A_i) + \mathrm{P3}(A_i) + \mathrm{P4}(A_i)$

P1, P2, P3, and P4 are the priority functions of a coordination bond, a single bond, a double bond and a triple bond, respectively. If A_i has *n* coordinate bonds, then $\mathrm{P1}(A_i) = n$; if it has *n* single bonds, then $\mathrm{P2}(A_i) = 2n$; if it has *n* double bonds, then $\mathrm{P3}(A_i) = 3n$; if it has *n* triple bonds, then $\mathrm{P4}(A_i) = 4n$. The values 1, 2, 3, and 4 are the values of the priority functions P1, P2, P3, and P4 for a coordination bond, a single bond, a double bond, and a triple bond, respectively. The stronger the bond, the larger the value.

As an example, the atomic characters of all the atoms in structure **1** are shown in Table 2. The atom originally numbered as 3 has the highest atomic character (800042500) and is thus renumbered as 1. Four atoms (7, 9, 13, and 14) have the next highest atomic character (600042500) and are all renumbered as 2, and this process is continued until all atoms have been renumbered.

Table 2. Atomic Characters and Initial Orders in Structure **1**

atom no.	atom	atomic character	initial order
1	C	200042500	6
2	O	400072500	4
3	C	800042500	1
4	C	400042500	5
5	C	410042500	3
6	C	400042500	5
7	C	600042500	2
8	C	400042500	5
9	C	600042500	2
10	O	400072500	4
11	C	200042500	6
12	C	400042500	5
13	C	600042500	2
14	C	600042500	2
15	C	400042500	5
16	C	410042500	3
17	C	400042500	5

INITIAL ORDER

In ESSESA, atoms in a structure are ordered by the atomic characters, and the order produced is called the initial order. As the atomic character becomes larger, so the inital order decreases. Because the atomic character is based on the structural environment description of the atom in the structure taken only to a connectivity distance of 1, it must be extended to longer connectivity distances; sometimes the full structure if it is to be used to partition the structure completely. For example, the atomic characters and initial orders of structure 1 are generated from these atomic characters and listed in Table 2.

PARTITIONING

The basic Morgan algorithm and the improvements that have been made to it do not use all structural data that could be used to partition the atoms during the analysis. The connectivity is an important property of the atom in chemical structures, but it is not the only property available, and the equivalence of final extended connectivity values does not necessarily imply equivalence of atoms. Any graph invariant can be, and sometimes must be, used in the partitioning process. The graph invariants used in ESSESA are the properties of structure taken to a connectivity distance of 1. They should not be used to partition the chemical structure completely. The graph invariants including the properties of the whole structure should be used, and this is done by extending the order value.

The first step in our algorithm calls for definition of the initial order according to the atomic characters and is shown in Table 2. This initial order is used to generate the first extended order values and order, the first order are used to generate the second order values and order, and this process is repeated until the order obtained no longer changes with further extension. This means that the structure has been completely partitioned. The product method is used for this extension, and the order is obtained according to the preceding order and the extended order value scores. For example, in structure **1**, the initial order of atom 1 is 6. When the extension is done by the product method, the extended order value is

atom 1 = 6(atom 1) × 4(atom 2 connected to atom 1) = 24 and all the non-hydrogens in the structure are dealt with in the same way:

atom 2 = 4(atom 2) × 6(atom 1) × 1(atom 3) = 24

atom 3 = 1(atom 3) × 4(atom 2) × 5(atom 17) × 2(atom 14) × 5(atom 4) = 200

atom 4 = 5(atom 4) × 1(atom 3) × 3(atom 5) = 15

atom 5 = 3(atom 5) × 5(atom 4) × 5(atom 6) = 75

atom 6 = 5(atom 6) × 3(atom 5) × 2(atom 7) = 30

atom 7 = 2(atom 7) × 5(atom 6) × 2(atom 14) × 5(atom 8) = 100

atom 8 = 5(atom 8) × 2(atom 7) × 2(atom 9) = 20

atom 9 = 2(atom 9) × 5(atom 8) × 4(atom 10) × 5(atom 12) = 200

atom 10 = 4(atom 10) × 2(atom 9) × 6(atom 11) = 48

atom 11 = 6(atom 11) × 4(atom 10) = 24

atom 12 = 5(atom 12) × 2(atom 9) × 2(atom 13) = 20

atom 13 = 2(atom 13) × 5(atom 12) × 2(atom 14) × 5(atom 15) = 100

atom 14 = 2(atom 14) × 1(atom 3) × 2(atom 7) × 2(atom 13) = 8

atom 15 = 5(atom 15) × 2(atom 13) × 3(atom 16) = 30

atom 16 = 3(atom 16) × 5(atom 15) × 5(atom 17) = 75

atom 17 = 5(atom 17) × 1(atom 3) × 3(atom 16) = 15

The atoms may now be renumbered according to the preceding order and the extended order value scores. Thus atom 7 is renumbered as atom 3; that is defined as the extended order in the algorithm. The details of this algorithm are shown in Figure 1, in which AC_i are atomic characters, O_j are orders, OVE_{jn} are extended order values, and EO_j are orders. The partitioning process and results from structure 1 are shown in Table 3.

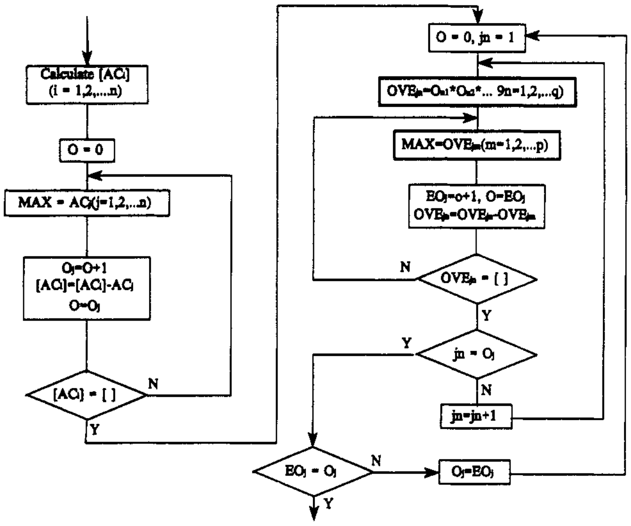


Figure 1. Overview of algorithm.

Table 3. Partitioning Process and Results

			partition			
			first extension		second extension	
atom no.	atomic character	initial order	order value	order	order value	order
1	200042500	6	24	11	77	11
2	400072500	4	24	7	77	7
3	800042500	1	200	1	2800	1
4	400042500	5	15	10	50	10
5	410042500	3	75	5	400	5
6	400042500	5	30	8	120	8
7	600042500	2	100	3	864	3
8	400042500	5	20	9	54	9
9	600042500	2	200	2	972	2
10	400072500	4	48	6	132	6
11	200042500	6	24	11	66	12
12	400042500	5	20	9	54	9
13	600042500	2	100	3	864	3
14	600042500	2	8	4	36	4
15	400042500	5	30	8	120	8
16	410042500	3	75	5	400	5
17	400042500	5	15	10	50	10

Table 4. Canonical Numbering of Structure 1

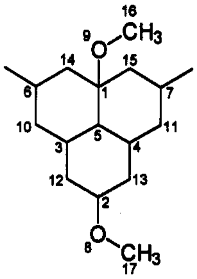
canonical numbering	numbering in LNSCS	canonical numbering	numbering in LNSCS
1	3	10	15
2	9	11	6
3	13	12	12
4	7	13	8
5	14	14	17
6	16	15	4
7	5	16	1
8	10	17	11
9	2		

CANONICAL NUMBERING

If a molecule has no symmetry, the canonical numbering of atoms on the molecule can be generated from partitioning of the molecule. If it has symmetry, the canonical numbering of atoms cannot be generated from only its partitioning. To obtain canonical numbering, that is to say, to describe a molecule canonically, a canonical numbering rule must be given.

The numbering rule in ESSESA is as follows. If atoms have the same order from the first partitioning, then they are renumbered arbitrarily. The numbering of subsequent atoms that have the same order (larger than the first) is generated according their connectivity distances to the first atoms, the shorter the connectivity distance to the first small numbering atom, the lower the numbering. The rule is illustrated by structure 1, a simple example. The process of generating canonical numbering of structure 1 proceeds thus. The canonical numbering for atoms 3 and 9 becomes 1 and 2, respectively. The first pair of atoms with the same order (3) are atom 13 and atom 7. Atom 13 is defined as number 3 and atom 7 number 4. The order of atom 14 is now 4. After atom 7 is thus redefined as 4, atom 4 is renumbered as 5. Next, atoms 5 and 16 have the same order (5), because the connectivity distance of atom 16 to atom 13 is 2, shorter than that to atom 5 (4); atom 16 is therefore assigned the number 6 and atom 5 number 7. Atoms 10, 2, and 15 are renumbered as 8, 9, and 10 respectively. Atoms 6 and 15 (order 8), atoms 8 and 12 (order 9), and atoms 4 and 17 (order 10) are dealt with in the same way. Finally, atoms 4, 1, and 11 are renumbered as 15, 16, and 17 respectively. The result of canonically numbering structure 1 is shown in Table 4.

Table 5. Connection Table and New Structure for Structure 1



order	atom	free radical	charge	configuration	conformation	connections
1	C	0	0	0	0	[[9,2][14,2][5,2][15,2]]
2	C	0	0	0	0	[[13,2][8,2][12,2]]
3	C	0	0	0	0	[[12,2][5,2][10,2]]
4	C	0	0	0	0	[[11,2][5,2][13,2]]
5	C	0	0	0	0	[[1,2][4,2][3,2]]
6	C	.	0	0	0	[[10,2][14,2]]
7	C	.	0	0	0	[[15,2][11,2]]
8	O	0	0	0	0	[[2,2][17,2]]
9	O	0	0	0	0	[[16,2][1,2]]
10	C	0	0	0	0	[[3,2][6,2]]
11	C	0	0	0	0	[[7,2][4,2]]
12	C	0	0	0	0	[[2,2][3,2]]
13	C	0	0	0	0	[[4,2][2,2]]
14	C	0	0	0	0	[[1,2][6,2]]
15	C	0	0	0	0	[[1,2][7,2]]
16	C	0	0	0	0	[[9,2]]
17	C	0	0	0	0	[[8,2]]

Table 6. Result of Partitioning Structure 2

				1		2	
atom	label	atomic character	initial order	value	order	value	order
C	1	400042500	3	9	8	48	17
C	2	400042500	3	27	6	336	7
C	3	400042500	3	18	7	210	11
C	4	600042500	2	12	5	105	6
C	5	800042500	1	72	1	2880	1
C	6	200042500	4	4	9	9	19
C	7	400042500	3	9	8	56	16
C	8	400042500	3	18	7	280	10
C	9	600042500	2	12	5	210	5
C	10	600042500	2	24	3	525	3
C	11	400042500	3	18	7	147	14
C	12	400042500	3	18	7	196	12
C	13	600042500	2	18	4	392	4
C	14	800042500	1	48	2	2880	2
C	15	200042500	4	4	9	18	18
C	16	400042500	3	9	8	96	15
C	17	400042500	3	27	6	288	8
C	18	400042500	3	27	6	252	9
C	19	400042500	3	18	7	168	13

CANONICAL CONNECTION TABLE

Using canonical numbering, the arbitrary connection table is transformed to the canonical connection table. The canonical connection table of structure **1** and the corresponding structure are shown in Table 5.

LNSCS TO CANONICAL CONNECTION TABLE

In ESSESA the LNSCS is translated to the corresponding canonical connection table by a program, the compiler of LNSCS. The structure of the compiler of LNSCS is shown in Figures 2 and 3.

DISCUSSION

Representation of chemical structures in the computer is very important in connection with structure search and analysis. The only structural property used in the Morgan algorithm, one of the earliest to be studied, is the atomic connectivity. Partitioning by the Morgan algorithm is incomplete because, in the connectivity matrix, all the atoms are indistinguishable—the same color in graph theory. Many workers have studied this partition and derived improved algorithms, most of which however are only modifications of the Morgan algorithm. Weininger[14] used atomic invariants to partition atoms by the product of primes method, and the result is better than that of the extended sums method.

Our method generates a unique description for a chemical structure by the partitioning of atoms in the structure according

Table 7. Result of Partitioning Structure 3

				1		2		3	
atom	label	atomic character	initial order	value	order	value	order	value	order
O	1	200072500	3	6	6	24	7	35	8
C	2	400042500	2	6	4	24	5	70	5
C	3	600042500	1	8	1	40	2	180	2
C	4	400042500	2	10	2	16	3	54	3
C	5	200042500	5	10	8	16	9	27	1
C	6	400042500	2	4	5	25	6	72	6
C	7	400042500	2	4	5	25	6	36	7
C	8	600042500	1	8	1	45	1	96	1
C	9	400042500	2	8	3	21	4	32	4
N	10	200042500	4	8	7	21	8	32	9
C	11	400042500	2	8	3	21	4	32	4
N	12	200042500	4	8	7	21	8	32	9

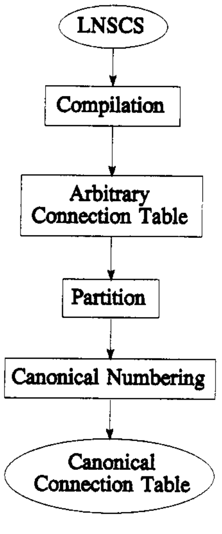


Figure 2. Overview of LNSCS to canonical connection table process.

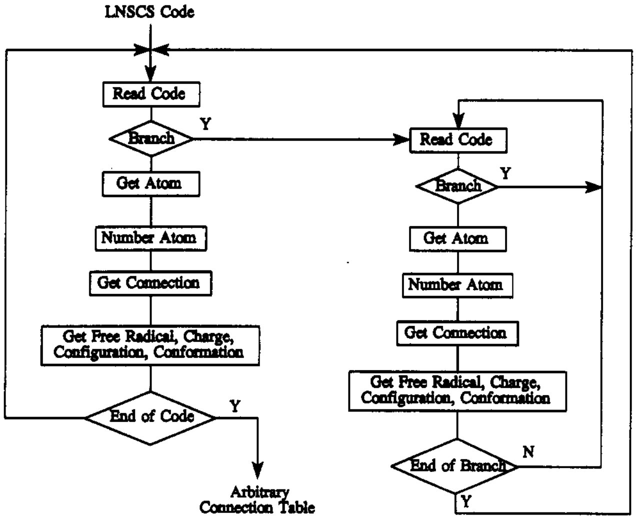
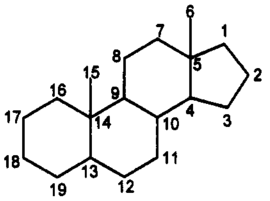


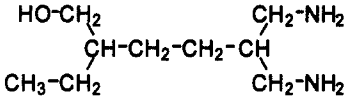
Figure 3. Compilation of LNSCS.

to their structural properties. In this system a chemical structure is treated as a color graph in graph theory, many graphic invariants are combined by ordering function into atomic characters, then the partition is carried out by the product scores method using atomic characters. This algorithm is more effective than the extended sums method, the extended connectivity scores method, or the product of primes method. For structure **2**, for example, the partitioning by the extended sums method and the extended connectivity scores method require five iterative computations,[13] whereas the partition can be completed by only two iterative computations with our algorithm. For structure **3** the partition by the products of primes method requires three iterative computations, and every iteration must get primes and products.[14] By our algorithm the partition is also completed with three iterative computations, but the iterations only are required to get the products. The partition results of structures **2** and **3** by our algorithm are shown in Tables 6 and 7.



Structure 2

LNSCS: C1CCC2C1(C)CCC3C2CCC4C3(C)CCCC4



Structure 3

LNSCS: OCC(CC)CCC(CN)2

ACKNOWLEDGMENT

We owe a deep debt of gratitude to the Editor, G. W. A. Milne, for encouraging the writing of this paper and the redrafting the manuscript.

REFERENCES AND NOTES

(1) Hong, H.; Xin, X. ESSESA: An Expert System for Structure Elucidation from Spectra. 3. LNSCS for Chemical Knowledge Representation. *J. Chem. Inf. Comput. Sci. 1992*, *32*, 116.

(2) Hong, H.; and Xin, X. ESSESA: An Expert System for Elucidation of Structures from Spectra. 1. The Knowledge Base of Infrared Spectra and Analysis and Interpretation Program. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 203.

(3) Hong, H.; Xin, X. ESSESA, an expert system for structure elucidation from spectral analysis Part II. Novel algorithm of perception of the linear independent smallest set of smallest rings. *Anal. Chim. Acta* **1992**, *262*, 179.

(4) Balaban, A. T. *Chemical Application of Graph Theory*; Academic Press: New York, 1976.

(5) Morgan, H. L. Generation of Unique Machine Description for Chemical Structures, a Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5*, 107.

(6) Wipke, W. T.; Dyott, T. M. Stereochemically Unique Naming Algorithm. *J. Am. Chem. Soc.* **1974**, *96*, 4834.

(7) O'Korn, L. J. Algorithms in the Computer Handling of Chemical Information. In *Algorithms for Chemical Computations*; Christofferen, R. E., Ed.; American Chemical Society: New York, 1972; p 122.

(8) Uchino, M. Algorithms for Unique and Unambiguous Coding and Symmetry Perception of Molecular Structure Diagrams. II. Basic Algorithm for Unique Coding and Computation of Group. *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 116.

(9) Bersohn, M.; Esack, A. A Canonical Connection Table Representation of Molecular Structure. *Chim. Scr.* **1974**, *6*, 122.

(10) Shelley, C. A.; Munk, M. E.; Roman, R. V. A Unique Computer Representation for Molecular Structures. *Anal. Chim. Acta* **1978**, *103*, 245.

(11) Shelley, C. A.; Munk, M. E. An Approach to Assignment of Canonical Connection Table and Topological Symmetry Perception. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 247.

(12) Jochum, C.; Gasteiger, J. Canonical Numbering and Constitutional Symmetry. *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 113.

(13) Gray, N. A. B. *Computer-Assisted Structure Elucidation*; Wiley: New York, 1986; pp 287–292.

(14) Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97.