(18) Ritter, G. L.; Isenhour, T. L. Minimal Spanning Tree Clustering of Gas Chromatographic Liquid Phases. *Comput. Chem.* **1977**, *1*, 145–153. Everitt, B. *Cluster Analysis*; Halsted: New York, 1974.

(19) Balaban, A. T. Chemical Graphs. XXXIV. Five New Topological Indices for the Branching of Tree-Like Graphs. *Theor. Chim. Acta* **1979**, *53*, 355–375.

(20) Balaban, A. T.; Motoc, I. Chemical Graphs. XXXVI. Correlations between Octane Number and Topological Indices of Alkanes. *MATCH* **1979**, *5*, 197–218.

(21) Read, R. C.; Corneil, D. G. The Graph Isomorphism Disease. *J. Graph Theor.* **1977**, *1*, 339–363.

(22) Stobaugh, R. E. Chemical Substructure Searching. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 271–275.

(23) Lindsay, R. K.; Buchanan, B. G.; Feigenbaum, E. A.; Lederberg, J. *Application of Artificial Intelligence for Organic Chemistry*; McGraw-Hill: New York, 1980.

(24) Balaban, A. T.; Harary, F. Chemical Graphs. 4. Enumeration and Proposed Nomenclature of Benzenoid Catacondensed Polycyclic Aromatic Hydrocarbons. *Tetrahedron* **1968**, *24*, 2505–2516. Polansky, O. E.; Rouvray, D. H. Graph-Theoretical Treatment of Aromatic Hydrocarbons. I. The Formal Graph-Theoretical Description. *MATCH* **1976**, *2*, 63–90.

(25) Entriger, R. C.; Jackson, D. E.; Snyder, D. A. Distance in Graphs. *Czech. Math. J.* **1976**, *26*, 283–296.

(26) Skorobogatov, V. A.; Khvorostov, P. V. Analiz Metricheskikh svoistv grafov. *Vychisl. Sist.* **1981**, *91*, 1–20.

(27) Balaban, A. T.; Mekenyan, O.; Bonchev, D. Unique Description of Chemical Structures Based on Hierarchical Ordered Extended Connectivities (HOC Procedures). I. Algorithms for Finding Graph Orbits and Canonical Numbering of Atoms. *J. Comput. Chem.* **1985**, *6*, 538–551.

(28) Skorobogatov, V. A.; Dobrynin, A. A. Metric Analysis of Graphs. *MATCH* **1988**, *23*, 105–151.

(29) Polansky, O. E.; Bonchev, D. The Wiener Number of Graphs. I. General Theory and Changes Due to Graph Operations. *MATCH* **1987**, *21*, 133–186.

(30) Bonchev, D.; Mekenyan, O.; Karabunarliev, S. The IVEC Algorithm for Coding of Chemical Compounds and Centric Ordering of Their Atoms and Bonds. Unpublished data.

# SMILES. 2. Algorithm for Generation of Unique SMILES Notation

DAVID WEININGER, ARTHUR WEININGER, and JOSEPH L. WEININGER*

Daylight Chemical Information Systems, Irvine, California 92714

Received May 4, 1988

The chemical notation language SMILES is designed for the conversion of an arbitrarily chosen description of a chemical structure to one unique notation. This is accomplished in a two-stage algorithm, CANGEN. The first stage involves CANonicalization of structure, whereby the molecule is treated as a graph with nodes (atoms) and edges (bonds). Each atom is canonically ordered and labeled. In the second stage, starting with the lowest labeled atom, a molecular graph is GENerated, which is the unique SMILES structure.

## INTRODUCTION

The SMILES chemical notation language was introduced in the first paper of this series.[1] Processing chemical information with greater efficiency than conventional methods, it represents a new approach to computerized chemical nomenclature. SMILES is simple to write because rules and hierarchical procedures, which are inherently difficult for the chemist, are relegated to computer algorithms. For a given chemical structure, arbitrary SMILES notation can take many equally valid forms. One must emerge as "unique" to serve as the identifier of the structure for database and other computer applications.

This is accomplished by a method called CANGEN that combines two separate algorithms, CANON and GENES. The first stage, CANON, labels a molecular structure with canonical labels. The structure is treated as a graph with nodes (atoms) and edges (bonds). Each atom is given a numerical label on the basis of its topology. In the second stage, GENES generates the unique SMILES notation as a tree representation of the molecular graph. GENES selects the starting atom and makes branching decisions by referring to the canonical labels as needed.

The combined procedure designates a unique SMILES notation for each chemical structure regardless of the many possible equivalent descriptions of the structure that might be input.

## THEORETICAL BACKGROUND

Generally, graph theory has become important in applications to chemical information because it provides the basis for codification of nomenclature in chemical computer programs.[2] The classification and ordering of nodes in a graph is here applied to chemical structure notation. With an initial set of node properties and a given connectivity for a two-dimensional, nondirected graph (with $N$ nodes and $E$ edges), each node is assigned a rank. In CANGEN this ranking completely discriminates each node environment with respect to all initial mode properties. Aside from node and edge properties, the classification algorithm must recognize constitutionally symmetric nodes, i.e., nodes that are topologically equivalent in all respects. This step and the generation of unique node order, breaking all ties, graph construction, and identification are all essential parts of the CANGEN process.

Combinatorial and extended sums methods are two different approaches for characterization of graph nodes and their environments. The combinatorial process is suitable for analyses of small graphs (simple chemical structures) but becomes too cumbersome for more complex ones because of the need to characterize each node environment completely. Simple, exhaustive solutions that have orders of max $(N,E)!$ become impractical as $N$ increases beyond 15. Partial characterization is therefore often attempted and is adequate for most symmetry perception problems. Such algorithms use a general approach of breadth-first optimization of a tree.[3] Nodes are characterized successively deeper into the total graph until the combined characterization is adequate. This usually reduces the base of the algorithmic order of $N$ or $E$ to the number of edges in the shortest path between the most distant nodes. However, these algorithms do not avoid the problem of factorial order for the general case.[4]

The sums method achieves greater efficiency by limiting the use of a combined description of connected nodes while ignoring all path-specific topological information. A sum vector **S** is modified iteratively by summing over the **S** elements of

* Address correspondence to this author at 809 Karenwald Lane, Schenectady, NY 12309.

0095-2338/89/1629-0097$01.50/0

**Table I.** Atomic Invariants

| |
|---|
| (1) number of connections |
| (2) number of non-hydrogen bonds |
| (3) atomic number |
| (4) sign of charge |
| (5) absolute charge |
| (6) number of attached hydrogens |

neighboring nodes only one edge away. On iteration, the difference in the developed sums of two nodes indicates nonequivalence. This sums method is an intrinsically low-order process but contains inherent difficulties. The original Morgan algorithm[5] uses only the local degree (number of nearest neighbors) in S and relies on subsequent combinatorial identification of other zeroeth order node properties, *P* (for example, the atomic numbers). It is possible to improve the basic sums method by incorporating node properties in a two-dimensional matrix (known as the extended sums method). Even so, combinations of nodes must be produced and compared.

There is an inherent ambiguity in sums with respect to addends of rational numbers. Hence, identical extended S values do not necessarily assure symmetry of nodes for properties *P* in a general case. This is the reason for eventually invoking combinatorial procedures. It will be shown below, however, that use of an unambiguous function can eliminate the necessity for a subsequent combinatorial procedure. Furthermore, when ambiguity is eliminated, all zeroeth order node properties may be expressed in a single vector S. Tracking individual properties as matrices of "extended sums" becomes unnecessary and redundant.

Node ordering for generation of unique SMILES notation is obtained by developing topological symmetry classes in the manner of extended sums, but using the product of corresponding primes in the extension process. The method of using an "unambiguous function" will be illustrated below with examples of labeling, ranking, and unique ordering of structural notations. It guarantees canonicalization over originally specified graph theoretical invariant properties.

## CANON: CANONICALIZATION OF MOLECULAR GRAPHS USING AN UNAMBIGUOUS FUNCTION

**(a) Initial Graph Invariant Order.** Graph theoretical invariants are properties of graphs that are independent of the way a graph is ordered. Examples are the atomic invariants of Table I. A unique linear combination of these invariants represents their initial vector in the CANGEN algorithms. For example, the methyl carbon in pentane (CCCCC) is represented by invariants 1 (number of connections), 01 (number of non-hydrogen valence bonds), 06 (atomic number), 0 (sign of charge), 0 (absolute charge), and 3 (number of attached hydrogens). The combination of these invariants is given by the linear description 10106003 (see also Table II). This set of six variables is sufficient for the purpose of obtaining unique notation for simple SMILES, but it is not necessarily a "complete" set. No "perfect" set of invariants is known that will distinguish all possible graph asymmetries. However, for any given set of structures, a set of invariants can be devised to provide the necessary discrimination. The list shown in Table I is used by CANGEN for the construction of simple molecular graphs. Other graph properties may be added as needed. When more information is required, for example, in the case of isomeric SMILES, invariants are added to denote isotopic mass, bond directionality, and local chirality. Conversely, one or more invariants may be eliminated in less rigorous operations than CANGEN conversion of SMILES notation.

The set of invariants in Table I have indicated priorities (1 is first, 6 has last priority). This set conforms to the fundamental assumption, made throughout the CANGEN process, that 1:1 mapping represents any set of invariants equally well. As an example, Table II gives the two invariant sets of the methyl and methylene carbons of the pentane molecule, CCCCC.

**Table II.** Invariants for Pentane

| atom type | individual invariant | combined invariant |
|---|---|---|
| methyl carbon | 1, 1, 6, 0, 0, 3 | 10106003 |
| methylene carbon | 2, 2, 6, 0, 0, 2 | 20206002 |

**(b) Rank Equivalence.** Although the different values in an invariant set must be ordered by their priority, there is nothing intrinsically meaningful in their specific values. To avoid numerical overflow of the computer system, these values are replaced by small numbers; the rank of each invariant retains the desired properties. The initial invariants for pentane are

10106003-20206002-202060002-20206002-10106003

Their ranks are

1-2-2-2-1

giving a new invariant set that is just as usable in the CANGEN process, and more suitable for machine processing, than the original set.

**(c) Simple Extended Connectivity.** While there are only two types of carbon atoms in pentane (methyl and methylene), there are three carbon symmetry classes. Morgan[5] and Bersohn[6] view sets of invariants in terms of the sums of the atoms' invariants one-away, two-away, etc. The test of symmetry classes is whether or not the "extended connectivity sums" are different. For pentane summing the neighbors one away reveals the three symmetry classes: the first ranking, from the initial invariant, is

1-2-2-2-1

Operation of a given function over all nearest (i.e., one-away) neighbors yields a set representing extended connectivity. Traditionally, addition is used, which in this example leads to

2-3-4-3-2

which can be replaced with a new rank for another set of symmetry classes

1-2-3-2-1

This process can be repeated until the set no longer changes. To avoid storing an indeterminate number of such sets, the CANGEN method retains the last set of ranks and uses the extended connectivity function only to break ties in each iteration. Thus, the previous ranks are used in the iteration process to maintain rank stability (each iteration only breaks ties). When the rank vector is not changed by an iteration, there is no need to continue because once an iteration fails to differentiate the equivalence of any node pair, all subsequent iterations will also fail.[7] The extended connectivity method is then complete, and an invariant partitioning is presumed to have been developed.

**(d) Extended Connectivity Using an Unambiguous Function.** One pitfall of the extended sums method, as pointed out above, is the inherent ambiguity of the sum of integers, which can lead to false symmetry perceptions. At this point earlier methods[5,6,8] revert to higher order, combinatorial algorithms that are very slow. The CANGEN process avoids this problem by using an unambiguous function rather than simple addition. One possibility is to include functions such as simple products and sum of squares in the extended connectivity evaluation. If the maximum connectivity is *K*, any set of *K* linearly independent functions will be unambiguous over integral input. A simple and elegant function, suggested by Freed,[9] is the product of corresponding primes. In this method, each rank

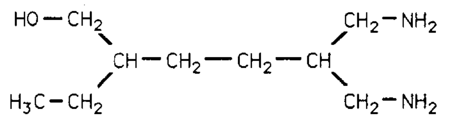

Figure 1. CANON example: OCC(CC)CCC(CN)CN.

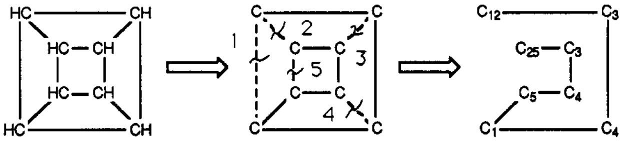

Figure 2. Generation of SMILES for cubane: C12C3C4C1C5C4C3C25.

is replaced by its corresponding prime (starting with 2) and then replaced by the product of its neighbors.

To illustrate the method of product of corresponding primes, take the case of two three-connected atoms whose neighbors' ranks are 1, 4, 4 and 2, 2, 5. The difference would not be distinguished by either simple sums (9 for each) or sum of squares (33 for each). Taking the product of their corresponding primes reveals the difference: 2 × 7 × 7 = 98 differs from 3 × 3 × 11 = 99. It is easily shown from the prime factorization theorem that this procedure will always provide an unambiguous result for any set of input ranks. Furthermore, it can be seen that this function is commutative and that every output number can be used (viz., every number is either prime or the unambiguous product of other prime factors). The only disadvantage is that, for very large molecules, these computations may use very large numbers, so 64-bit arithmetic is used (rather than 32 bit).

As an example, the generation of a unique SMILES notation for the compound 6-amino-2-ethyl-5-(aminomethyl)-1-hexanol with the molecular structure of Figure 1 will be considered. An initial arbitrary SMILES for this structure, OCC(CC)CCC(CN)CN, has the original ranking of invariants

3-4-5-(4-1)-4-4-5-(4-2)-4-2

In this case, the fact that the sum of the atomic numbers of the two nitrogens equals the sum of the atomic numbers of the oxygen and methyl carbon leads to unresolved, false symmetries if only summed invariants were considered:

3-6-8-(4-1)-7-7-8-(5-2)-5-2

In contrast, using the product of corresponding primes on the same initial set yields

3-6-9-(4-1)-7-8-10(5-2)-5-2

Details of this procedure are shown in the Appendix.

**(e) Breaking Ties.** If there are no constitutionally symmetric node classes in the graph, the problem of ordering nodes is not difficult. Simple bilateral symmetries, such as in the previous example, would not require further analysis for the purpose of unique nomenclature generation. Difficulties arise when more symmetric graphs, such as that of cubane with the notation C12C3C4C1C5C4C3C25 (see reference 1), are ordered (Figure 2).

The cubane molecule has eight identical carbon atoms. Iteration shows that all eight nodes are identical, so the starting point must be arbitrary. Once a starting point is chosen, however, the remaining seven nodes are no longer identical; three atoms are one-away, three atoms are two-away, and one atom is three-away. To avoid an arbitrary decision among these later (which would lead to a nonunique final notation), a complete canonical labeling of all nodes is needed. The algorithm proceeds by doubling all ranks and reducing the value of the first (lowest valued) atom, which is tied, by one. The set is then treated as a new invariant set, and the previous algorithm for generating an invariant partitioning is repeated. For cubane the final ranking (for the above input order) is

1-2-5-3-7-8-6-4

Note that there are several equivalent labelings for a symmetrical graph such as cubane. With respect to the initial atom, all of the correct labelings will assign 2,3,4 to one-away atoms, 5,6,7 to the two-away atoms, and 8 to the three-away atom. The "double-and-tie-break" step does not introduce ambiguity into the ordering since only otherwise equivalent atoms will be tied at any point. This step is required to assure that an ordering which is not equivalent to the correct labeling will not be generated (e.g., labeling the three-away atom something other than 8).

**Table III.** CANON Algorithm

(1) Set atomic vector to initial invariants. Go to step 3.
(2) Set vector to product of primes corresponding to neighbors' ranks.
(3) Sort vector, maintaining stability over previous ranks.
(4) Rank atomic vector.
(5) If not invariant partitioning, go to step 2.
(6) On first pass, save partitioning as symmetry classes.
(7) If highest rank is smaller than number of nodes, break ties, go to step 2.
(8) ... else done.

**(f) CANON Algorithm Summary.** Table III lists the eight steps of the CANON algorithm for the canonicalization of chemical structure. For an $N$-atom structure, this algorithm requires, at most, $N$ sorts of $N$ integers, with the tie-breaking step (7) introducing a maximum constant factor of 2. Therefore, the order of this algorithm is

$$N^2 \log_2 (N)$$

## GENES: GENERATION OF UNIQUE SMILES

With symmetrical classes established, the structure is treated as a tree and a SMILES string is generated that corresponds to a depth-first search (DFS) of that tree. The only required decisions are where to start, i.e., at which node of the tree, and which branch to follow at each branching point. Finally, the unique SMILES string per se is generated. Symbols for branch termination and ring closure are included, and a second DFS search is performed for polycyclic structures to ensure proper ordering of ring closure labels.

**(a) Initial Node Selection.** The lowest canonically numbered atom is chosen as the starting point of the SMILES notation. This atom becomes the root of a tree for a subsequent depth-first search. For the example of 6-amino-2-ethyl-5-(aminomethyl)-1-hexanol the final ranking has the terminal carbon of the ethyl group as the starting point (root) of the graph (see above). As a rule, this selection implies that a terminal atom is chosen if one exists. This is desirable for efficiency, because a pair of parentheses is eliminated, and also for aesthetics.

If the chemical structure consists of separate entities, such as ions or ligands, it is considered a disconnected compound, denoted by a period as the disconnection symbol. Repeated selection of starting atoms, using the same criterion of the lowest remaining canonical label, ultimately produces a disconnected SMILES (a forest).

**(b) Branching Decisions.** Branching decisions could be as simple as the selection of a starting atom because the algorithm directs branching toward the lowest labeled atom at the fork in the branch. For example, acetone has the combined invariants

```
   C    -    C    ( =O     )-    C
10106003–30406000–(10208000)–10106003
```

resulting in the canonical labeling 1-4-(3)-2. Starting with the methyl group labeled 1, the direction of branching at the

Table IV. Perception of Topological Symmetry Classes

| A | O | C | C | (C | C) | C | C | C | (C | N) | C | N |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| (a) Structure and Initial Invariants | | | | | | | | | | | | |
| B | 118001 | 226002 | 336001 | 226002 | 116003 | 226002 | 226002 | 336001 | 226002 | 117002 | 226002 | 117002 |
| C | 3 | 4 | 5 | 4 | 1 | 4 | 4 | 5 | 4 | 2 | 4 | 2 |
| (b) Classification by Extended Sums Method | | | | | | | | | | | | |
| D | 3, 4 | 4, 8 | 5, 12 | 4, 6 | 1, 4 | 4, 9 | 4, 9 | 5, 12 | 4, 7 | 2, 4 | 4, 7 | 2, 4 |
| E | 3 | 6 | 8 | 4 | 1 | 7 | 7 | 8 | 5 | 2 | 5 | 2 |
| F | 3, 6 | 6, 11 | 8, 17 | 4, 9 | 1, 4 | 7, 15 | 7, 15 | 8, 17 | 5, 10 | 2, 5 | 5, 10 | 2, 5 |
| G | 3 | 6 | 8 | 4 | 1 | 7 | 7 | 8 | 5 | 2 | 5 | 2 |
| (c) Classification by Product of Primes Method | | | | | | | | | | | | |
| C | 3 | 4 | 5 | 4 | 1 | 4 | 4 | 5 | 4 | 2 | 4 | 2 |
| C′ | 5 | 7 | 11 | 7 | 2 | 7 | 7 | 11 | 7 | 3 | 7 | 3 |
| D | 7 | 55 | 343 | 22 | 7 | 77 | 77 | 343 | 33 | 7 | 33 | 7 |
| E | 3 | 6 | 8 | 4 | 1 | 7 | 7 | 8 | 5 | 2 | 5 | 2 |
| E′ | 5 | 13 | 19 | 7 | 2 | 17 | 17 | 19 | 11 | 3 | 11 | 3 |
| F | 13 | 95 | 1547 | 38 | 7 | 323 | 323 | 2057 | 57 | 11 | 57 | 11 |
| G | 3 | 6 | 8 | 4 | 1 | 7 | 7 | 9 | 5 | 2 | 5 | 2 |
| G′ | 5 | 13 | 19 | 7 | 2 | 17 | 17 | 23 | 11 | 3 | 11 | 3 |
| H | 13 | 95 | 1547 | 38 | 7 | 323 | 391 | 2057 | 69 | 11 | 69 | 11 |
| I | 3 | 6 | 9 | 4 | 1 | 7 | 8 | 10 | 5 | 2 | 5 | 2 |
| I′ | 5 | 13 | 23 | 7 | 2 | 17 | 19 | 29 | 11 | 3 | 11 | 3 |
| J | 13 | 115 | 1547 | 46 | 7 | 437 | 493 | 2299 | 87 | 11 | 87 | 11 |
| K | 3 | 6 | 9 | 4 | 1 | 7 | 8 | 10 | 5 | 2 | 5 | 2 |

central carbon atom will be toward the second methyl group, which has a lower rank than the oxygen of the carbonyl. Consequently, the unique SMILES for acetone is CC(C)=O, not CC(=O)C.

In cyclic structures, at branches with multiple bonds, it would be preferable not to select a multiple bond for a SMILES ring closure. This is avoided by branching in a ring toward the multiple bond rather than toward the single bond. The following two rules apply: (1) Branch to a double or triple bond in the ring if one exists, or (2) branch to the lower canonically numbered atom. Rule 2 is the same as the one that applies to linear structures [cf. CC(C)=O and CCC(CO)CCC(CN)CN above].

**(c) Two-Pass Method: Treatment of Cyclic and Polycyclic Structures.** There are several algorithms for SMILES generation available that are based on the DFS. A two-pass method is chosen because it produces unique SMILES for complex polycyclic structures in an intuitively correct manner. It starts with a simple DFS, appending nodes (atomic) and edge (bond) symbols to the output SMILES as the search progresses. Each time a branch is taken, a left parenthesis is added to the output string; each time a dead-end is reached, a right parenthesis is added. The first pass terminates when all nodes have been reached. If the structure is linear, the SMILES is complete and unique; there is no need for a second pass. For cyclic structures, however, the search will encounter a node that has already been visited. At this point the ring closure nodes are known (the last node and the already visited one) so that the SMILES ring closure indicators (digits) can now be appended to the node symbols in preparation for the second DFS pass. This second DFS enables the two-pass method to cope with the following problems that are specific to polycyclic systems:[10] (i) searching around a ring where an errant left parenthesis may be dangling; (ii) sorting the digits on nodes with multiple ring closures; (iii) ordering the digits in "opening" order since their assignment is not determined by the closing order.

## SUMMARY

The CANGEN process consists of a two-stage algorithm. The first stage involves canonicalization of structure, whereby the molecular structure is treated as a graph with nodes (atoms) and edges (bonds). All atoms are canonically ranked on the basis of a suitable set of invariant node properties and are labeled numerically. In the second stage, starting with the lowest ranked atom, a tree (molecular graph) is constructed that is the unique SMILES notation regardless of which of various valid original linear SMILES notations was originally specified. The generation of unique SMILES by this process provides the key to solving the basic problem of chemical nomenclature, namely, that a single chemical compound may have many different names. When a unique notation for a structure is obtained, chemical nomenclature is amenable to many applications for databases where SMILES can be associated with any number of synonyms, identifiers, and structure keys, such as common names, Collective Index names, IUPAC names, and CAS numbers. A unique SMILES notation serves extremely well as an identifier for a chemical database. This will be illustrated in one of the next publications in this series which describes a SMILES-oriented, extremely efficient database.

## ACKNOWLEDGMENT

We thank Prof. Robert S. Pearlman for a critical reading of the manuscript.

## APPENDIX: COMPUTATION OF EXTENDED SUMS AND PRODUCT OF CORRESPONDING PRIMES FOR 6-AMINO-2-ETHYL-5-(AMINOMETHYL)-1-HEXANOL

The methods of extended sums and product of corresponding primes are illustrated in Table IV as an application of extended connectivity. Starting with the nonunique SMILES OCC(CC)CCC(CN)CN for 6-amino-2-ethyl-5-(aminomethyl)-1-hexanol, a stable set of symmetry classes is developed by using each method from a common set of graph theoretical invariants.

Table IV shows the CANGEN operation as individual steps in sequence and in parallel. Table IVa shows the initial structure input (in arbitrary SMILES input order), original invariants, and their ranks in line A–C, respectively.

The inherent ambiguity of extended sums is shown by the stable classification developed following iterations, Table IVb, where the identity of atoms in classes 7 (central methylenes) and 8 (tertiary carbons) are erroneous. In each iteration, ranks are generated from (last rank, sum-of-neighbors), with the last rank having higher priority.

The unambiguous function method is shown in Table IVc, where the primes corresponding to ranks are shown in rows with primed labels, and the products of adjacent primes (which are used only to break rank ties) are listed in rows D, F, H, and J. Note that the symmetry is now correctly perceived due to the 1547 ≠ 2057 tiebreak in row F.

The symmetry classification in row K is stable (recognized by being identical with previous classification in row I).

As described in the text, CANON continues by breaking the lowest tie (symmetry class 2, nitrogens) to produce 12 distinct labelings. Starting with the lowest labeled atom and branching to lower labeled atoms at forks in the structure, the unique SMILES, CCC(CO)CCC(CN)CN, is established by GENES.

## REFERENCES AND NOTES

(1) Weininger, D. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31.
(2) Balaban, A. T. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 334.
(3) Joachim, C.; Gasteiger, J. *Top. Curr. Chem.* **1987**, *74*, 93.
(4) Wipke, W. T.; Dyott, T. M. *J. Am. Chem. Soc.* **1974**, *96*, 4834.
(5) Morgan, H. L. *J. Chem. Doc.* **1965**, *5*, 107.
(6) Bersohn, M. *Comput. Chem.* **1987**, *2*, 113.
(7) Hagadone, T. R.; Howe, W. J. *J. Chem. Inf. Comput. Sci.* **1982**, *22*, 182.
(8) Uchino, M. *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 116.
(9) Freed, E. E. Harvey Mudd College, personal communication.
(10) Wenger, J. C.; Smith, D. H. *J. Chem. Inf. Comput. Sci.* **1982**, *22*, 29.

# Computer Translation of IUPAC Systematic Organic Chemical Nomenclature. 1. Introduction and Background to a Grammar-Based Approach

D. I. COOKE-FOX, G. H. KIRBY,* and J. D. RAYNER

Department of Computer Science, University of Hull, Hull HU6 7RX, England

Received May 6, 1988

Since Garfield's pioneering work over 25 years ago in the linguistic aspects of systematic chemical nomenclature, leading to an algorithm for translating chemical names to formulas, very few reports of grammar-based analysis of systematic chemical nomenclatures have appeared in the literature. These have applied only to a few specific classes of names. While the major abstracting services use automated methods to process chemical nomenclatures, the limited details that have been published point to ad hoc approaches based on dictionaries of morphemes. This paper introduces a series that covers in detail the various aspects of the application of grammar-based techniques to the recognition of IUPAC systematic chemical nomenclature and hence the translation of chemical names to structure diagrams. Some necessary elements of language and grammar are discussed here in the context of the automatic recognition of chemical nomenclature.

## INTRODUCTION

There are three broad categories of chemical language by which structural information is represented and communicated. These are the nomenclatures used to name compounds, formulas and line notations used as shorthand representations of compounds, and structure diagrams used as the primary means of communication of structural information and compounds. Chemical structures are also represented by connection tables, which are used internally by most computer-based transformation techniques as a topological description of molecular structure. However, connection tables are rarely used for communication between people and are not regarded as languages. The translation or interconversion of these languages by automatic means is an important application of computer science to chemical structure representation and processing. A review with references to those interconversions that have been reported is given by Rush.[1]

Computer translation from and to a systematic nomenclature has received little attention, and a recent book[2] has said that existing programs are very large and complicated and will be successful in this translation in considerably less than 100% of cases. This situation is associated with the slowness with which systematic nomenclature, as typified by the schemes devised by the International Union of Pure and Applied Chemistry (IUPAC), is accepted and used, with the continuing use of much semisystematic and trivial nomenclature, and with the questionable need for fully systematic nomenclature as perceived by the chemical industry.[3] In the U.K., the Chemical Nomenclature Advisory Service of the Laboratory of the Government Chemist encourages the use of systematic nomenclature following the principles set by IUPAC and is prominent in advising European Commision Services on these matters. Egan[4] and Egan and Godly[5] have discussed some of the benefits of using IUPAC systematic nomenclature, while the issues and problems associated with the use of chemical nomenclature are covered in the book edited by Lees and Smith.[6]

Work supported by the Laboratory of the Government Chemist has been in progress in this department for some years to investigate the application of grammar-based techniques, as developed for compiling computer programming languages, to automatic name recognition and translation into structure diagrams. In this project attention has been paid to certain classes of compounds of industrial importance, including some cases of semisystematic and trivial nomenclature. A particular feature of the project has been the use of inexpensive and readily available computing facilities as exemplified by the IBM PC and compatible microcomputers. An outline of the project in its early stages has been published.[7]

The first step in the translation of chemical nomenclature by grammar-based techniques is to develop a grammar that formally describes the syntax of the nomenclature. From the grammar, a parser can be produced to recognize names that satisfy the grammar and to check the semantics, or meaning, of the names. Names that are syntactically correct may nevertheless be chemical nonsense. Only after satisfying semantic checking is an intermediate form of a name constructed, the concise connection table.[8] Further processing leads to representations suited to communication to other computer software or to the display of a structure diagram.

## CHEMICAL NOMENCLATURES

**Overview of Nomenclature Styles.** An excellent review of the development of chemical nomenclature is given by Cahn and Dermer.[9] Following the Geneva Congress of 1892, the maintenance of the rules of chemical nomenclature was taken on by the International Union of Pure and Applied Chemistry (IUPAC), who published revisions to the rules of organic

0095-2338/89/1629-0101$01.50/0