Table I. Number of Different Types of Unbranched Catacondensed Benzenoids

h	d	m	c	d	total unbranched
1	1	0	0	0	1
2	1	0	0	0	1
3	1	1	0	0	2
4	1	1	1	1	4
5	1	4	1	4	10
6	1	3	4	16	24
7	1	12	4	50	67
8	1	10	13	158	182
9	1	34	13	472	520
10	1	28	39	1406	1474
11	1	97	39	4111	4248
12	1	81	116	11998	12196
13	1	271	115	34781	35168
14	1	226	339	100660	101226
15	1	764	336	290464	291565
16	1	638	988	837137	838764
17	1	2141	977	2408914	2412033
18	1	1787	2866	6925100	6929754
19	1	6025	2832	19888057	19896915
20	1	5030	8298	57071610	57084939

according to their symmetries into dihedral, d, D_{2h} (regular hexagonal, D_{6h}, for $h = 1$); mirror-symmetrical, m, C_{2v}; centrosymmetrical, c, C_{2h}; and unsymmetrical, u, C_s.

A listing of the computer program in PASCAL is available on request to the authors.

REFERENCES AND NOTES

(1) Klarner, A. D. "Some Results Concerning Polyominoes". *Fibonacci Q.* **1965**, *3*(1), 9-20.
(2) Golomb, S. W. *Polyominoes*; Scribner, New York, 1965.
(3) Harary, F.; Read, R. C. "The Enumeration of Tree-like Polyhexes". *Proc. Edinburgh Math. Soc.* **1970**, *17*, 1-14.
(4) Lunnon, W. F. "Counting Polyominoes" in *Computers in Number Theory*; Academic: London, 1971; pp 347-372.
(5) Lunnon, W. F. "Counting Hexagonal and Triangular Polyominoes". *Graph Theory Comput.* **1972**, 87-100.
(6) Brunvoll, J.; Cyvin, S. J.; Cyvin, B. N. "Enumeration and Classification of Benzenoid Hydrocarbons". *J. Comput. Chem.* **1987**, *8*, 189-197.
(7) Balaban, A. T., et al. "Enumeration of Benzenoid and Coronoid Hydrocarbons". *Z. Naturforsch., A: Phys., Phys. Chem., Kosmophys.* **1987**, *42A*, 863-870.
(8) Gutman, I. "Topological Properties of Benzenoid Systems". *Bull. Soc. Chim., Beograd* **1982**, *47*, 453-471.
(9) Gutman, I.; Polansky, O. E. *Mathematical Concepts in Organic Chemistry*; Springer: Berlin, 1986.
(10) Tošić, R.; Doroslovački, R.; Gutman, I. "Topological Properties of Benzenoid Systems—The Boundary Code". *MATCH* **1986**, No. 19, 219-228.
(11) Doroslovački, R.; Tošić, R. "A Characterization of Hexagonal Systems". *Rev. Res. Fac. Sci.-Univ. Novi Sad, Math. Ser.* **1984**, *14*(2) 201-209.
(12) Knop, J. V.; Szymanski, K.; Trinajstić, N. "Computer Enumeration of Substituted Polyhexes". *Comput. Chem.* **1984**, *8*(2), 107-115.
(13) Stojmenović, I.; Tošić, R.; Doroslovački, R. "Generating and Counting Hexagonal Systems". *Proc. Yugosl. Semin. Graph Theory, 6th, Dubrovnik 1985*; pp 189-198.
(14) Doroslovački, R.; Stojmenović, I.; Tošić, R. "Generating and Counting Triangular Systems". *BIT* **1987**, *27*, 18-24.
(15) Knop, J. V.; Müler, W. R.; Szymanski, K.; Trinajstić, N. *Computer Generation of Certain Classes of Molecules*; Association of Chemists and Technologists of Croatia: Zagreb, 1985.

Downloaded by UNIV NAC AUT DE MEXICO UNAM on August 12, 2009
Published on February 1, 1988 on http://pubs.acs.org | doi: 10.1021/ci00057a005

SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules

DAVID WEININGER

Medicinal Chemistry Project, Pomona College, Claremont, California 91711

Received June 17, 1987

SMILES (Simplified Molecular Input Line Entry System) is a chemical notation system designed for modern chemical information processing. Based on principles of molecular graph theory, SMILES allows rigorous structure specification by use of a very small and natural grammar. The SMILES notation system is also well suited for high-speed machine processing. The resulting ease of usage by the chemist and machine compatability allow many highly efficient chemical computer applications to be designed including generation of a unique notation, constant-speed (zeroeth order) database retrieval, flexible substructure searching, and property prediction models.

INTRODUCTION

The first step in the formalization of chemistry is to name a chemical compound. This requires an unambiguous and reproducible notation for the simplest atom to the most complicated structure. All other chemical information procedures follow from the fundamental process of chemical nomenclature. Consequently, the improvement of chemical notation has been an ongoing endeavor, amply documented in the pages of this journal.

Wiswesser[1] described the historical development of chemical nomenclature from the beginning of chemistry as a rudimentary science to the start of the computer era. When computers opened up the processing and storage of chemical information, this depended on the description of chemical structure. Morgan[2] developed a technique for generation of unique machine description from which followed the CAS (Chemical Abstracts Service) ONLINE search system.[3] Other important advances were the application of graph theory to chemical notation[4] and chemical substructure search systems.[5-7] The uniqueness of chemical information has also been considered from a theoretical point of view.[8]

With the introduction of computers, line notation[9] became widely used in chemical nomenclature because computers can process linear strings of data with relative ease. Line notation serves as the basis of the International Union of Pure and Applied Chemistry (IUPAC) notation system.[10] It is described by Read[11] in general terms in relation to graph-theoretical concepts. Read lists 12 attributes considered desirable in a chemical coding system. In 1983 many of these attributes were incompatible with each other. In the few intervening years, however, advances in computer technology have accelerated so that they are overcoming the incremental increase of chemical information. Computer technology and chemical knowledge are now at a point where it is possible to store all of the extant chemical information on existing hardware. The chemical information problem of just getting a machine to store information is largely historical. Current and future efforts must be directed to building highly efficient systems

0095-2338/88/1628-0031$01.50/0 © 1988 American Chemical Society

Downloaded by UNIV NAC AUT DE MEXICO UNAM on August 12, 2009
Published on February 1, 1988 on http://pubs.acs.org | doi: 10.1021/ci00057a005

that provide chemically relevant information as needed. For this purpose, a new chemical language and ancillary computer programs are being developed, which are based on the new information system SMILES (Simplified Molecular Input Line System).[12] This paper introduces the methodology and encoding rules used by the SMILES system.

SMILES: OBJECTIVES AND APPROACH

SMILES is a chemical notation language specifically designed for computer use by chemists. It is easily accessible to chemists, yet flexible enough to allow interpretation and generation of chemical notation independent of the specific computer system in use. Similar to conventional chemical notation, it improves on conventional software methods by greater speed and better use of computer capacity. Molecular structures are uniquely and accurately specified and can be used with chemical databases. Among several approaches to computerized chemical notation, line notation is popular because it represents molecular structure by a linear string of symbols, similar to natural language. The Wiswesser Line Notation[9] is the most widely used representative of this method. It meets the essential requirements for a deterministic chemical notation, but it is difficult to use because many rules must be followed to generate the correct notation of a complex structure. To overcome this and other difficulties, the SMILES system was designed to be truly computer interactive. The simplicity of its use is based on computer programs that rigorously recode the chemical user's input. It is the result of achieving the following original objectives:

(1) The graph of a chemical structure was to be uniquely described, including but not limiting it to the molecular graph comprising nodes (atoms) and edges (bonds).

(2) A user-friendly structure specification was to be provided, so that all input rules could be learned quickly and naturally.

(3) A machine-friendly and machine-independent system was to be designed for interpretation and generation of a unique notation.

Unlike other chemical notation systems, brevity of notation and economy of alphabet were not primary objectives. Many of the pitfalls in other line notations can be attributed to overuse of symbols and hierarchical rules based on the length of the final notation. Advances in computer hardware have made these restrictions obsolete. The present approach separates the unambiguous but general description of a chemical structure (by the chemist-user) from the generation of the unique structural description (by the computer). The latter requires rules and hierarchies that are inherently difficult for the chemist. This task is, therefore, relegated to computer algorithms.

SMILES: SPECIFICATION RULES

SMILES denotes a molecular structure as a graph that is essentially the two-dimensional valence-oriented picture chemists draw to describe a molecule. This is an important simplification of chemical structure. No attempt is made to represent any particular three-dimensional arrangement of atoms.

SMILES notation is a series of characters that ends with a space. Hydrogen atoms may be omitted (hydrogen-suppressed graphs) or included (hydrogen-complete graphs). Aromatic structures are specified directly in preference to the Kekulé form.

Rules for generating SMILES for virtually any chemical structure are given in the following sections. The discussion will be limited to rules for specifying SMILES for chemical structures; specification of isomerisms, substructures, and unique SMILES generation are the subjects of following papers.

(1) Atoms. Atoms are represented by their atomic symbols; this is the only required use of letters in SMILES. Each non-hydrogen atom is specified independently by its atomic symbol enclosed in square brackets. The second letter of two-character symbols must be entered in lower case. Elements in the "organic subset", B, C, N, O, P, S, F, Cl, Br, and I, may be written without brackets if the number of attached hydrogens conforms to the lowest normal valence consistent with explicit bonds. Atoms in aromatic rings are specified by lower case letters; e.g., normal carbon is represented by the letter C, aromatic carbon by c. Since attached hydrogens are implied in the absence of brackets, the following atomic symbols are valid SMILES notations.

C	methane (CH_4)
N	ammonia (NH_3)
O	water (H_2O)
P	phosphine (PH_3)
S	hydrogen sulfide (H_2S)
Cl	hydrogen chloride (HCl)

Elements not in the organic subset must be described in brackets, e.g.

[Au]	elemental gold

Attached hydrogens and formal charges are always specified inside brackets. The number of attached hydrogens is shown by the symbol H followed by an optional digit. Similarly, a formal charge is shown by one of the symbols + or −, followed by an optional digit. If unspecified, the number of attached hydrogens and charges is assumed to be zero for an atom inside the bracket. Examples are

[H+]	proton
[OH−]	hydroxyl anion
[OH3+]	hydronium cation
[Fe+2]	iron(II) cation
[NH4+]	ammonium cation

The SMILES program also recognizes constructions of the form [Fe+++] as being synonymous with the form [Fe+3].

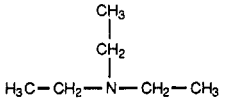
(2) Bonds. Single, double, triple, and aromatic bonds are represented by the symbols —, =, #, and :, respectively. Single and aromatic bonds may be, and usually are, omitted. Examples are

CC	ethane (CH_3CH_3)
C=C	ethylene ($CH_2=CH_2$)
COC	dimethyl ether (CH_3OCH_3)
CCO	ethanol (CH_3CH_2OH)
C=O	formaldehyde (CH_2O)
O=C=O	carbon dioxide (CO_2)
O=CO	formic acid (HCOOH)
C#N	hydrogen cyanide (HCN)
[H][H]	molecular hydrogen (H_2)

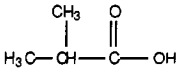
For linear structures, SMILES notation corresponds to conventional diagrammatic notation except that hydrogens can be omitted. For example, 6-hydroxy-1,4-hexadiene can be represented by three equally valid SMILES:

structure	valid SMILES
CH_2=CH-CH_2-CH=CH-CH_2-OH	C=CCC=CCO C=C-C-C=C-C-O OCC=CCC=C

(3) Branches. Branches are specified by enclosures in parentheses. Examples are



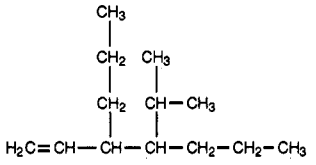
CCN(CC)CC
Triethylamine



CC(C)C(=O)O
Isobutyric acid

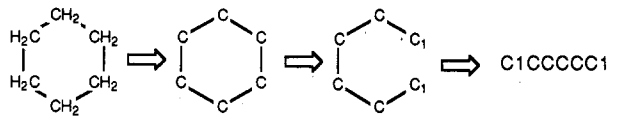
Downloaded by UNIV NAC AUT DE MEXICO UNAM on August 12, 2009
Published on February 1, 1988 on http://pubs.acs.org | doi: 10.1021/ci00057a005

Branches can be nested or stacked, as shown for 3-propyl-4-isopropyl-1-heptene:



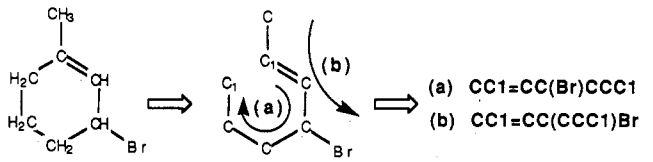
C=CC(CCC)C(C(C)C)CCC

(4) Cyclic Structures. Cyclic structures are represented by breaking one single (or aromatic) bond in each ring. The bonds are numbered in any order, designating ring-opening (or ring-closure) bonds by a digit immediately following the atomic symbol at each ring closure. This leaves a connected noncyclic graph, which is written as a noncyclic structure by using the three rules described above. Cyclohexane is a typical example:



C1CCCCC1

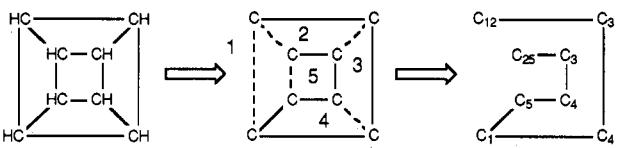
There are usually many different but equally valid descriptions of the same structure, e.g., the following SMILES notations for 1-methyl-3-bromo-cyclohexene:



(a) CC1=CC(Br)CCC1
(b) CC1=CC(CCC1)Br

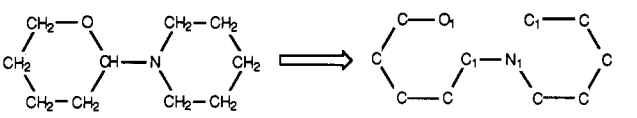
Many other notations may be written for the same structure, deriving from different ring closures. SMILES does not have a preferred entry on input; although (a) above may be simplest, others are just as valid.

A single atom may have more than one ring closure. This is illustrated by the structure of cubane, in which two atoms have more than two ring closures:



Generation of SMILES for cubane: C12C3C4C1C5C4C3C25

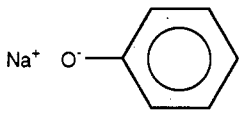
If desired, digits denoting ring closures can be reused. Reading left to right, a ring is closed at the first matching digit and may be reused without ambiguity. As an example, the digit 1 is used twice in the following specification:



O1CCCCC1N1CCCCC1

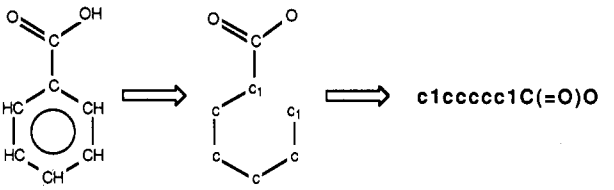
The ability to reuse ring-closure digits makes it possible to specify structures with 10 or more rings. Structures that require more than 10 ring closures to be open at once are exceedingly rare. If necessary or desired, higher numbered ring closures may be specified by prefacing a two-digit number with a percent sign. For example, a carbon with ring closures 2, 13, and 24 would be written C2%13%24.

(5) Disconnected Structures. Disconnected compounds are written as individual structures separated by a period. The order in which ions or ligands are listed is arbitrary. There is no implied pairing of one charge with another, nor is it necessary to have a net zero charge. If desired, the SMILES of one ion may be imbedded within another, as shown in the example of sodium phenoxide.



[Na+].[O-]c1ccccc1
or
c1cc([O-].[Na+])ccc1

(6) Aromaticity. Aromatic structures may be distinguished by writing the atoms in the aromatic ring in lower case letters, for example, benzoic acid.

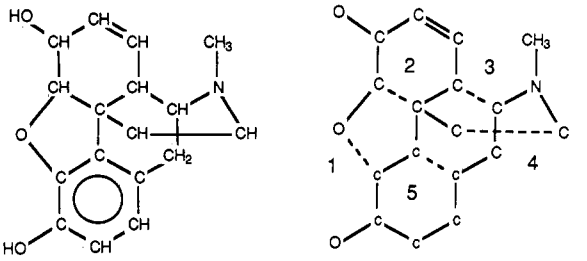


c1ccccc1C(=O)O

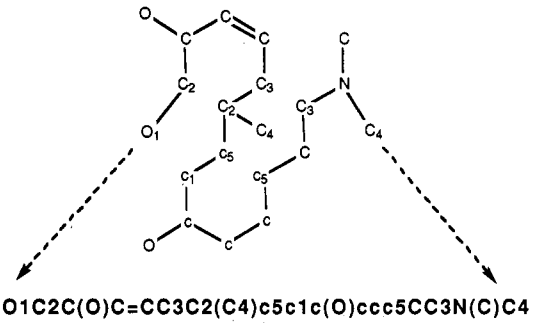
As will be discussed in more detail, the SMILES system automatically detects aromaticity, so input of an equivalent nonaromatic structure such as O=C(O)C1=CC=CC=C1 would be internally converted to the SMILES standard form.

With the above simple rules almost all organic structures can be described in line notation. An example of a more complex structure is that of morphine in Figure 1. It contains five rings, of which one is aromatic. Breaking of rings and designation of ring closures by means of digits attached to the symbols of ring atoms are shown. At ring closures 1 and 5 aromatic carbon atoms are shown in lower case letters c.

Morphine: Break & number 5 ring closures:



Generate SMILES for resulting non-cyclic structure:



O1C2C(O)C=CC3C2(C4)c5c1c(O)ccc5CC3N(C)C4

Figure 1. Evolution of SMILES for morphine.

BASIC SMILES

As simple as the SMILES rules are, an even simpler four-rule subset suffices for the vast majority of organic compounds. Although SMILES allows direct specification of charges, attached hydrogens, and aromaticity, often they are not required. This subset uses only the symbols H, C, N, O, P, S, F, Cl, Br, I, and (,) and digits, with the following four rules: (1) Atoms are represented by atomic symbols. (2) Double and triple bonds are represented by = and #, respectively. (3) Branching is indicated by parentheses. (4) Ring

closures are indicated by matching digits appended to symbols.

SMILES NOTATION CONVENTIONS

The process of establishing SMILES notation for morphine (Figure 1) shows the simplicity, flexibility, and consistency of SMILES language for specifying chemical structure. The flexibility of SMILES mirrors the variety found in chemistry. In an effort to provide a uniform notation system, SMILES rules are refined by stipulating some conventions that are to be followed for writing SMILES of certain classes of chemical compounds. These involve bond specification, hydrogen specification in certain nitrogen ring compounds, and additional factors involving aromaticity.

(1) Hydrogen Specification. Hydrogen atoms do not normally need to be specified when SMILES are written. Except for special purposes, the SMILES system treats hydrogen attachment as a property of non-hydrogen atoms. The number of attached hydrogens may be specified in three ways: (1) implicitly, for atoms specified without brackets, from normal valence assumptions; (2) explicitly by count, for atoms specified inside brackets, by the hydrogen count supplied (zero if unspecified); and (3) as explicit atoms, as [H] atoms in the SMILES. The SMILES convention regarding implicit hydrogen attachment assumes that hydrogens make up the remainder of an atom's lowest normal valence, consistent with explicit bond specification. The single normal valences of B, C, N, O, and the halogens are 3, 4, 3, 2, and 1, respectively. "Lowest normal valence" refers to 3 or 5 for phosphorus and 2, 4, or 6 for aliphatic sulfur. This ensures, for example, that OS(=O)(=O)O is interpreted as H_2SO_4 (sulfuric acid), while S is interpreted as H_2S (hydrogen sulfide). Aromatic sulfur donating a lone pair is assigned a formal valence of 3 or 5.

Even when specified as explicit atoms, the SMILES system removes all hydrogen atoms and just retains the attached hydrogen count. From this computer programs can generate a hydrogen-complete graph whenever needed. Eliminating hydrogens from the molecular graph makes tasks easier for both the chemist and the machine, mostly because there are fewer atoms to deal with.

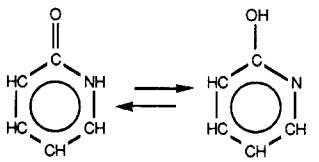
There are few exceptions to the hydrogen-suppression convention, the most obvious being specification of a proton, [H+], and molecular hydrogen, [H][H]. There are also some applications that require general specification of more than one bond to hydrogen, such as in crystallographic databases. The rule used in the SMILES system is to eliminate all hydrogen atoms except in the following three cases: (1) hydrogens connected to other hydrogens; (2) hydrogens connected to zero or more than one other atom; and (3) in isomeric SMILES, isotopic hydrogen specifications, e.g., [2H]. In these cases, hydrogens are retained and are treated like any other atom except that their hydrogen count is always zero. Case 3 is included here for completeness; isomeric SMILES is not otherwise covered in this paper.

(2) Bonds. The four fundamental bonds in SMILES are single, double, triple, and aromatic bonds. Ionic "bonds" are not specified directly; separate parts with formal charges are written as a disconnected structure. When needed, atoms may be connected and still show charge separation. There are, however, a few types of bonds that do not fit easily into the above categories. Occasionally, bonds have a hybrid character with both covalent and ionic characteristics. Such bonds can be denoted in several ways, but for database applications it is necessary to choose one description and adhere to it.

In most organic compounds the bonds are covalent, so the choice of bond type is simple for organic and also for covalent inorganic compounds. Delocalized bonds, such as in the nitro moiety, are also best written as covalent bonds to both uncharged oxygens, as in nitromethane, CN(=O)=O. This is a matter of convention because nitromethane could also be written as a valid charge-separated structure, C[N+](=O)[O-]. The advantage of the uncharged form is that it preserves correct topological symmetry. When symmetry is not an issue, charge-separated structures are preferred if they avoid representing atoms in unusual valence states. For instance, diazomethane is written as C=[N+]=[N-] in preference to C=N=[N].

There is no distinction between "organic" and "inorganic" SMILES nomenclature. One may specify explicitly the number of attached hydrogens for any atom in any SMILES. The hydrogen count, however, must then be specified inside the brackets; the default value is zero. For example, propane may be entered as [CH3][CH2][CH3] instead of as CCC.

(3) Tautomers. Tautomeric structures are explicitly specified in SMILES. There are no "tautomeric bond" or "mobile hydrogen" specifications. Selection of one or all tautomeric structures is left to the user and will depend on the application. For database and indexing purposes many authors prefer the enol over the keto form. For example, for the structure



keto enol

the enol form, Oc1nccc1 (2-pyridinol) is usually preferred over O=c1[nH]cccc1 (2-pyridone), but this is a matter of convention. If a formal representation must be chosen for modeling, then the more "stable" form is generally preferred. In actual practice the only effective approach for predictive modeling purposes is to generate all reasonable tautomeric forms and model each one.

(4) Aromaticity Detection. Aromaticity must be detected in a system that generates an unambiguous chemical nomenclature. As will be discussed in following papers, this is needed both for the generation of a unique nomenclature and for effective substructure recognition. There can be no definition of "aromaticity" that is both rigorous and all-encompassing; the word implies something about "reactivity" to a synthetic chemist, "ring current" to a NMR spectroscopist, "symmetry" to a crystallographer, and presumably "odor" to the original user of the word. Our objective in defining aromaticity is to provide an automatic and rigorous definition for the purposes of generating an unambiguous chemical nomenclature. Although the SMILES algorithm produces results that most chemists find natural, nothing is implied by this definition about physical properties.

Given effective aromaticity-detection algorithms, it is not necessary to enter any structure as aromatic if the user prefers to enter an aliphatic (Kekulé-like) structure. Entering structures as aromatic provides a shortcut to accurate chemical specification and is closer to the mental molecular model that most chemists use. One advantage of a rigorous algorithmic redefinition is that any valid structure which suits the user can be automatically converted to a standard form.

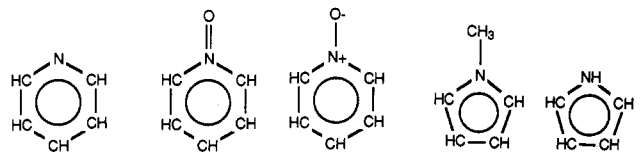
SMILES algorithms detect accurately the vast majority of aromatic compounds and ions. The system will accept either aromatic or nonaromatic input specifications; it will detect aromaticity and will convert the input structure accordingly. This is accomplished with an extended version of Hückel's rule to identify aromatic molecules and ions.[13] To qualify as aromatic, all atoms in the ring must be sp2 hybridized and the number of available "excess" π electrons must satisfy Hückel's $4N + 2$ criterion. As an example, benzene is written c1ccccc1, but an entry of C1=CC=CC=C1 (cyclohexatriene)—the Kekulé form—leads to detection of aro-

Downloaded by UNIV NAC AUT DE MEXICO UNAM on August 12, 2009
Published on February 1, 1988 on http://pubs.acs.org | doi: 10.1021/ci00057a005

maticity and results in an internal structural conversion to aromatic representation. Entries of c1ccc1 and c1cccccc1 will produce the correct antiaromatic structures for cyclobutadiene and cyclooctatetraene, C1=CC=C1 and C1=CC=CC=CC=C1, respectively. In such cases the SMILES system looks for a structure that preserves the implied sp2 hybridization, the implied hydrogen count, and the specified formal charge if any. Some inputs, however, may be not only formally incorrect but also nonsensical such as c1cccc1. Here c1cccc1 is not the same as C1=CCC=C1 (which is a valid SMILES for cyclopentadiene) since one of the carbon atoms is sp3 with two attached hydrogens. In such a structure, alternating single- and double-bond assignments cannot be made. The SMILES system will flag this as an "impossible" input.

One of the features of the SMILES interpreter is that all structures that are denoted as aromatic (for purposes of unique notation) may be automatically converted to nonaromatic form (for modeling, compatability with other systems, or other purposes).

(5) Compounds Containing Aromatic Nitrogen. To avoid confusion, aromatic nitrogens require special attention. There are two types of aromatic nitrogens that are distinguished within the SMILES system; both may be specified with the aromatic nitrogen symbol n. Archetypical examples are pyridine and pyrrole:

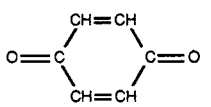


n1ccccc1 Pyridine **O=n1ccccc1** Pyridine-*N*-oxide **[O-][N+]c1ccccc1** **Cn1cccc1** methyl and 1H-pyrrole **[nH]1cccc1**

Pyridine is best written as n1ccccc1; SMILES correctly deduces that no hydrogens are attached to the nitrogen in pyridine because two aromatic bonds satisfy the normal valence of nitrogen. Leaving the problem of aromaticity detection aside, neither the nitrogen in O=n1ccccc1 (pyridine N-oxide) nor the nitrogen in Cn1cccc1 (methylpyrrole) has a valency left for an extra hydrogen at the normal valence state. In [nH]1cccc1 (1*H*-pyrrole), however, the nitrogen is both aromatic and two-connected, but has one attached hydrogen atom. This is indicated in SMILES by writing the aromatic n symbol in brackets, where an attached hydrogen can be specified. Alternative valid input SMILES for 1*H*-pyrrole include Hn1cccc1, [H]n1cccc1, and of course the aliphatic form N1C=CC=C1; each of these will be converted to the correct internal representation.

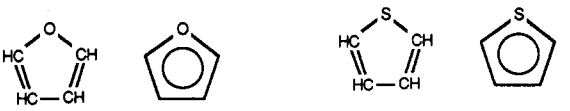
(6) Examples of Aromatic and Nonaromatic Compounds. The rules for the SMILES aromaticity detection algorithm as given above are quite simple; rings of sp2-hybridized atoms that have $4N + 2$ π electrons are classified as aromatic. The operation of this algorithm is discussed below with reference to example structures.

Neutral unsaturated carbon donates one π electron to the ring except when it is double bonded to an electronegative atom outside the ring. In that case the carbon retains its sp2 orbitals, but the excess electron is not available for π orbital sharing. For example, quinone is nonaromatic, with only four excess electrons:



O=C1CCC(=O)CC1
quinone

In-ring oxygen and sulfur atoms donate a lone pair, so furan and thiophene

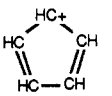


o1cccc1 furan **s1cccc1** thiophene

are both treated as aromatic ($N = 1$, i.e., $6 = 4N + 2$). In-ring sulfone is assumed to donate a lone pair, but sulfonyl cannot participate in normal π electron sharing.

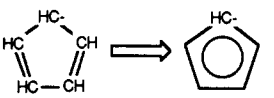
Uncharged in-ring nitrogen can donate one or two electrons to the π cloud in its pyridyl and pyrrolyl forms, respectively. In this respect at least, aromatic nitrogen is well-behaved. For instance, the nitrogen in pyridine N-oxide can be thought of as a pyridyl nitrogen donating its normal single electron to the ring or as a pyrrolyl nitrogen losing one electron to oxygen (of the normally contributed two electrons, still leaving one for the ring).

SMILES also correctly handles aromaticity of charged structures. Charged ring structures are identical with neutral ones except that the number of "excess" π electrons is reduced by the positive charge or increased by the negative in-ring charge. For example, the cyclopentadienyl cation



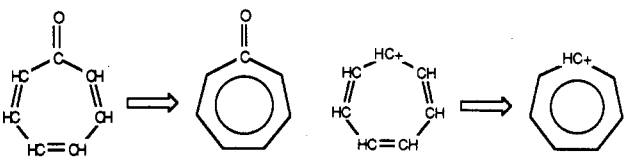
[CH+]1C=CC=C1

is nonaromatic, while the cyclopentadienate anion is aromatic:



[cH-]1cccc1

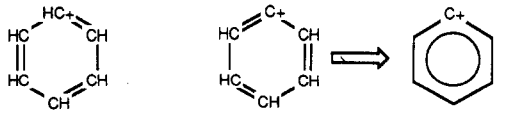
Tropone and tropylium cations are both aromatic



O=c1ccccc1 tropone **[cH+]1ccccc1** tropylium

because the carbonyl carbon "loses" its electron to the oxygen in one case and to cation formation in the other.

Special care must be taken to specify the number of attached hydrogens on each charged carbon. For example, if one of the benzene ring's electrons is removed to form c1ccc[cH+]1, this ion is not aromatic because there are only five π electrons. But if one of the hydrogens and a lone pair of electrons are removed, the aromatic phenyl cation is formed (which still has six π electrons):



[CH+]1=CC=CC=C1 non-aromatic **[c+]1ccccc1** aromatic

Phosphorus is treated like nitrogen, although there are few known aromatic pyrrole-like phosphorus compounds. Selenium and arsenic are treated similarly to phosphorus and sulfur. Elements other than C, N, O, P, S, As, and Se are not yet dealt with in an aromatic context.

Downloaded by UNIV NAC AUT DE MEXICO UNAM on August 12, 2009
Published on February 1, 1988 on http://pubs.acs.org | doi: 10.1021/ci00057a005

DISCUSSION

SMILES is believed to represent the best compromise to date between the human and machine aspects of chemical notation. It is easily understood by the chemist because it has a minimum number of simple rules. Computationally, SMILES is interpreted in a very fast, compact manner, thereby satisfying the machine objectives of time and space savings. It is based on a computer approach to language parsing so that the machine part follows algorithms consistent with rigorous hierarchical nomenclature rules. This results in a great improvement in the efficiency of information processing as compared to conventional methods.

In computer terms, SMILES notation represents a tree that can be interpreted in a single pass. The increase in efficiency derives from the language syntax. For example, the bonding arrangements are implied by the position of atoms without having to be defined specifically. Thus, where previously 1000–2000 characters may have been needed to store a connection table, depending on the method used, to describe a compound such as morphine (Figure 1), SMILES stores the same information in 40 characters. It can also be shown that the necessary computer processing time is reduced 100-fold over conventional procedures using connection table format, e.g., those used in CAS or MOL connection tables.

Originally, SMILES was developed to provide a human/machine language interface. Beyond this objective, it has been valuable for the implementation of a wide variety of machine-oriented chemical information functions. Successful applications include data storage, structural display, modeling new structures, and substructure searches and recognition. An example is the use of SMILES in computation of partition coefficients and molecular refractivity in model compounds. These two properties are widely used in biochemical research. The logarithm of the partition coefficient is the hydrophobic parameter in Hammett methodology and has been applied to quantitative structural activity (QSAR).[14] On the basis of structural considerations, fragments of a modeled compound, designated and processed in SMILES terminology, serve to estimate log P accurately.

This paper is intended to be an introduction to SMILES methodology and to cover the fundamental rules needed to enter a structure into the SMILES system. Subsequent publications will present the method of obtaining "unique" SMILES, generating a structural depiction for any SMILES, generating a SMILES-oriented database that retrieves information at a speed independent of how many structures are stored, and SMILES methods for fast and powerful substructure searching.

ACKNOWLEDGMENT

Research on SMILES was initiated by the author at the Environmental Research Laboratory, U.S.E.P.A., Duluth, MN, and was completed at Pomona College. The author thanks Arthur Weininger for assistance in programming the SMILES system and Dr. Joseph L. Weininger for editorial assistance.

REFERENCES AND NOTES

(1) Wiswesser, W. J. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 258–263.
(2) Morgan, H. L. *J. Chem. Doc.* **1965**, *5*, 107–113.
(3) Dittmar, P. G.; Farmer, N. A.; Fisanick, W.; Haines, R. C.; Mockus, J. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 93–102.
(4) Balaban, A. T. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 334–343.
(5) Wipke, W. T.; Rogers D. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 255–262.
(6) Attias, R. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 102–108.
(7) Stobaugh, R. E. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 271–275.
(8) Fugman, R. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 174–180.
(9) Wiswesser, W. J. *A Line-Formula Chemical Notation*; Crowell: New York, 1954.
(10) *IUPAC Nomenclature of Organic Chemistry*, Sections A–F, and H; Pergamon: Oxford, England, 1979.
(11) Read, R. C. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 135–149.
(12) Weininger, D.; Weininger, A.; Weininger, J. L. *Chem. Des. Autom. News* **1986**, *1*(8), 2–15.
(13) Weininger, S. J.; Stermitz, F. R. *Organic Chemistry*; Academic: Orlando, FL, 1984.
(14) Leo, A. J. *J. Chem. Soc., Perkin Trans. 2* **1983**, 825–838.

Downloaded by UNIV NAC AUT DE MEXICO UNAM on August 12, 2009
Published on February 1, 1988 on http://pubs.acs.org | doi: 10.1021/ci00057a005

COMPUTER SOFTWARE REVIEWS

Chemtext

YECHESKEL WOLMAN

Department of Organic Chemistry, The Hebrew University of Jerusalem, Jerusalem, Israel

Received September 9, 1987

There has been a need for a good chemical word-processing program that would enable the chemist to create documents which include complex chemical and mathematical formulas, documents in which one would be able to place chemical structures, molecules, and/or reactions directly into the text, manipulate them, store them, and finally print them on various printers.

CHEMTEXT (Version 1.1.) is an excellent answer to the above need. It is part of the Molecular Design Ltd (MDL) Chemist Personal Software Series (CPSS). CHEMTEXT is a word-processing program that consists of a main menu together with four different editors (Document, Molecule, Reaction, and Form editors). In order to use the program one needs an IBM personal computer (PC, PC XT, PC AT) or compatibles with 640K memory, color or monochrome monitor, graphic board (Hercules, IBM CGA, IBM EGA, or compatibles), a mouse (Mouse System or Microsoft Mouse), at least one floppy disk drive, a hard disk (at least 10 MB), and a dot-matrix or a laser printer.

The main menu is the entry point into CHEMTEXT. It is used to insert images into documents from the editors; it is as well a drawing editor on its own. It is used to enhance images