Pharmaceutical Chemistry

Simplified molecular-input line-entry system
SMILES

SAPIENZA
UNIVERSITÀ DI ROMA

Definition

The simplified molecular-input line-entry system (SMILES) is a specification in form of a line notation for describing the structure of chemical species using short ASCII strings. SMILES strings can be imported by most molecule editors for conversion back into two-dimensional drawings or three-dimensional models of the molecules.

The original SMILES specification was initiated in the 1980s. It has since been modified and extended. In 2007, an open standard called OpenSMILES was developed in the open-source chemistry community.

History

The original SMILES specification was initiated by David Weininger at the USEPA Mid-Continent Ecology Division Laboratory in Duluth in the 1980s. Acknowledged for their parts in the early development were "Gilman Veith and Rose Russo (USEPA) and Albert Leo and Corwin Hansch (Pomona College) for supporting the work, and Arthur Weininger (Pomona; Daylight CIS) and Jeremy Scofield (Cedar River Software, Renton, WA) for assistance in programming the system.

It has since been modified and extended by others, most notably by Daylight Chemical Information Systems. In 2007, an open standard called "OpenSMILES" was developed by the Blue Obelisk open-source chemistry community.

In July 2006, the IUPAC introduced the InChI as a standard for formula representation. SMILES is generally considered to have the advantage of being slightly more human-readable than InChI; it also has a wide base of software support with extensive theoretical (e.g., graph theory) backing.

Terminology

The term SMILES refers to a line notation for encoding molecular structures and specific instances should strictly be called SMILES strings. However, the term SMILES is also commonly used to refer to both a single SMILES string and a number of SMILES strings; the exact meaning is usually apparent from the context.

Typically, a number of equally valid SMILES strings can be written for a molecule. For example, CCO, OCC and C(O)C all specify the structure of ethanol. Algorithms have been developed to generate the same SMILES string for a given molecule; of the many possible strings, these algorithms choose only one of them. This SMILES is unique for each structure, although dependent on the canonicalization algorithm used to generate it, and is termed the canonical SMILES.

SMILES notation allows the specification of configuration at tetrahedral centers, and double bond geometry. These are structural features that cannot be specified by connectivity alone and SMILES which encode this information are termed isomeric SMILES. A notable feature of these rules is that they allow rigorous partial specification of chirality. The term isomeric SMILES is also applied to SMILES in which isotopes are specified.

Graph-based definition

In terms of a graph-based computational procedure, SMILES is a string obtained by printing the symbol nodes encountered in a depth-first tree traversal of a chemical graph. The chemical graph is first trimmed to remove hydrogen atoms and cycles are broken to turn it into a spanning tree. Where cycles have been broken, numeric suffix labels are included to indicate the connected nodes. Parentheses are used to indicate points of branching on the tree.

The resultant SMILES form depends on the choices:

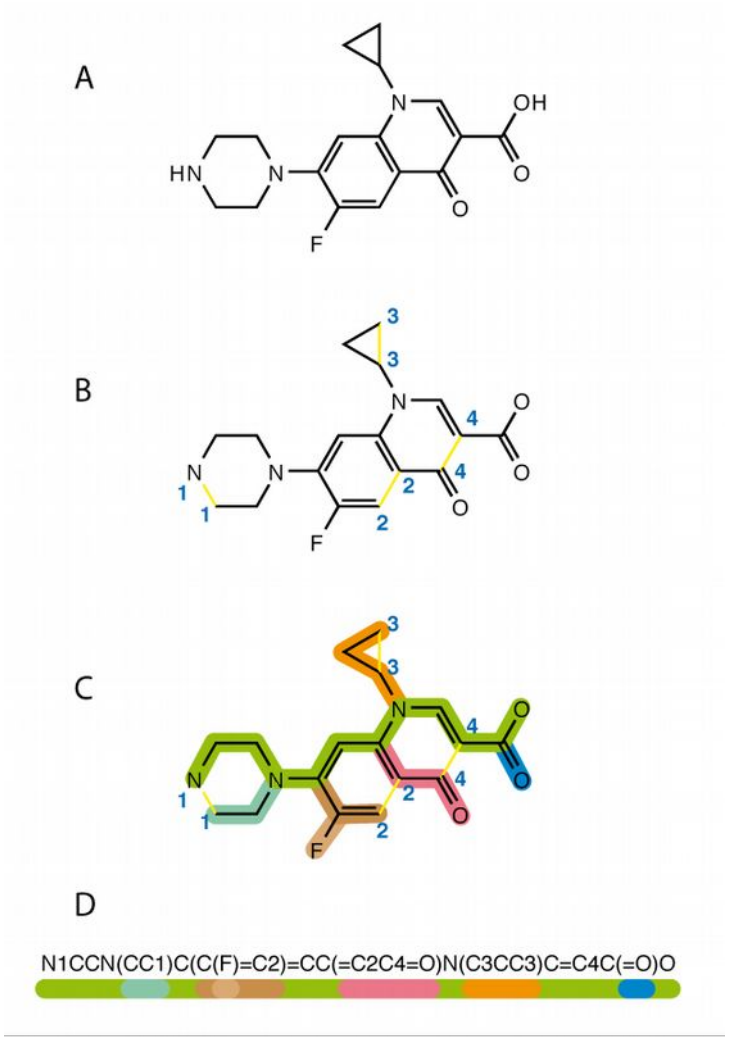
- of the bonds chosen to break cycles,
- of the starting atom used for the depth-first traversal, and
- of the order in which branches are listed when encountered.

Graph-based definition

In terms of a graph-based computational procedure, SMILES is a string obtained by printing the symbol nodes encountered in a depth-first tree traversal of a chemical graph. The chemical graph is first trimmed to remove hydrogen atoms and cycles are broken to turn it into a spanning tree. Where cycles have been broken, numeric suffix labels are included to indicate the connected nodes. Parentheses are used to indicate points of branching on the tree.

The resultant SMILES form depends on the choices:

- of the bonds chosen to break cycles,
- of the starting atom used for the depth-first traversal, and
- of the order in which branches are listed when encountered.



Description: Atoms

Atoms are represented by the standard abbreviation of the chemical elements, in square brackets, such as [Au] for gold. Brackets may be omitted in the common case of atoms which:

are in the "organic subset" of B, C, N, O, P, S, F, Cl, Br, or I, and

- have no formal charge, and
- have the number of hydrogens attached implied by the SMILES valence model (typically their normal valence, but for N and P it is 3 or 5, and for S it is 2, 4 or 6), and
- are the normal isotopes, and
- are not chiral centers.

All other elements must be enclosed in brackets, and have charges and hydrogens shown explicitly. For instance, the SMILES for water may be written as either O or [OH2]. Hydrogen may also be written as a separate atom; water may also be written as [H]O[H].

When brackets are used, the symbol H is added if the atom in brackets is bonded to one or more hydrogen, followed by the number of hydrogen atoms if greater than 1, then by the sign '+' for a positive charge or by '-' for a negative charge. For example, [NH4+] for ammonium. If there is more than one charge, it is normally written as digit; however, it is also possible to repeat the sign as many times as the ion has charges: one may write either [Ti+4] or [Ti++++] for Titanium IV (Ti4+). Thus, the hydroxide anion is represented by [OH-], the hydronium cation is [OH3+] and the cobalt III cation (Co3+) is either [Co+3] or [Co+++].