Representation of Chemical Structures with the IUPAC International Chemical Identifier (InChI)

Stephen R. Heller and Alan D. McNaught

The IUPAC International Chemical Identifier (InChI) is a freely available, non-proprietary identifier for chemical substances that can be used in both printed and electronic data sources. It is generated from a computerized representation of a molecular structure diagram, which can be produced by chemical structure-drawing software. Its use enables linking of diverse data compilations and unambiguous identification of chemical substances. A full description of the Identifier and software for its generation are available from the IUPAC website (Ref. 1), and a helpful compilation of answers to frequently asked questions has been put together at the Unilever Centre for Molecular Science Informatics (Ref. 2). Commercial structure-drawing software that will generate the Identifier is available from several organizations, listed on the IUPAC website.

The conversion of structural information to the Identifier is based on a set of IUPAC structure conventions, and rules for normalization and canonicalization (conversion to a single, predictable sequence) of an input structure representation. The resulting InChI is simply a series of characters that serve to uniquely identify the structure from which it was derived. The InChI uses a layered format to represent all available structural information relevant to compound identity. InChI layers are listed below. Each layer in an InChI representation contains a specific type of structural information. These layers, automatically extracted from the input structure, are designed so that each successive layer adds additional detail to the Identifier. The specific layers generated depend on the level of structural detail available and whether or not allowance is made for tautomerism. Of course, any ambiguities or uncertainties in the original structure will remain in the InChI.

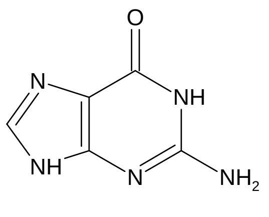
This layered structure design offers a number of advantages. If two structures for the same substance are drawn at different levels of detail, the one with the lower level of detail will, in effect, be contained within the other. Specifically, if one substance is drawn with stereo-bonds and the other without, the layers in the latter will be a subset of the former. The same will hold for compounds treated by one author as tautomers and by another as exact structures with all H-atoms fixed. This can work at a finer level. For example, if one author includes double bond and tetrahedral stereochemistry, but another omits stereochemistry, the latter InChI will be contained in the former.

The InChI layers are:

1. Formula
2. Connectivity (no formal bond orders)
    a. disconnected metals
    b. connected metals
3. Isotopes
4. Stereochemistry
    a. double bond (*Z/E*)
    b. tetrahedral (sp^3)
5. Tautomers (on or off)

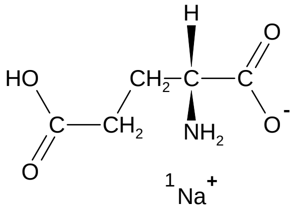
Charges are not part of the basic InChI, but rather are added at the end of the InChI string.

Two examples of InChI representations are given below. It is important to recognize, however, that InChI strings are intended for use by computers and end users need not understand any of their details. In fact, the open nature of InChI and its flexibility of representation, after implementation into software systems, may allow chemists to be even less concerned with the details of structure representation by computers.



guanine

InChI=1/C5H5N5O/c6-5-9-3-2(4(11)10-5)7-1-8-3/h1H,(H4,6,7,8,9,10,11)/f/h8,10H,6H2



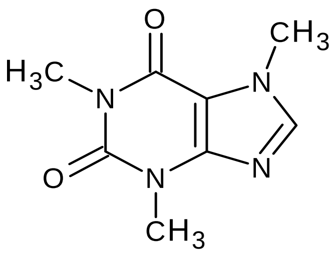
monosodium glutamate

InChI=1/C5H9NO4.Na/c6-3(5(9)10)1-2-4(7)8;/h3H,1-2,6H2,(H,7,8)(H,9,10);/q;+1/p-1/t3-;/m1./s1/fC5H8NO4.Na/h7H;/q-1;m

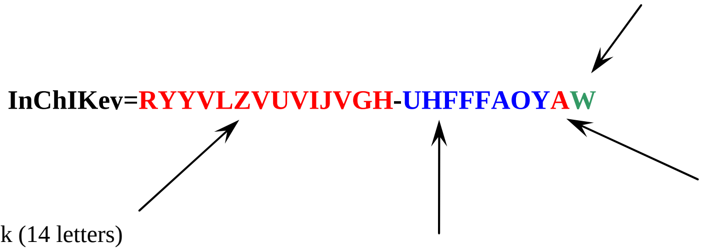
The layers in the InChI string are separated by the '/' character followed by a lowercase letter (except for the first layer, the chemical formula), with the layers arranged in predefined order. In the examples the following segments are included:

InChI version number
/ chemical formula
/c connectivity-1.1 (excluding terminal H)
/h connectivity-1.2 (locations of terminal H, including mobile H attachment points)
/q charge
/p proton balance
/t sp^3 (tetrahedral) parity
/m parity inverted to obtain relative stereo (1 = inverted, 0 = not inverted)
/s stereo type (1 = absolute, 2 = relative, 3 = racemic)
/f chemical formula of the fixed-H structure if it is different
/h connectivity-2 (locations of fixed mobile H)
/q charge
/t sp^3 (tetrahedral) parity
/m parity inverted to obtain relative stereo (1 = inverted, 0 = not inverted, **.** = inversion does not affect the parity)
/s stereo type (1 = absolute, 2 = relative, 3 = racemic)

One of the most important applications of InChI is the facility to locate mention of a chemical substance using internet-based search engines. This is made easier by using a shorter (compressed) form of InChI, known as InChIKey. The InChIKey is a 25-character representation that, because it is compressed, cannot be reconverted into the original structure, but it is not subject to the undesirable and unpredictable breaking of longer character strings by some search engines. An example is shown below.



InChI=1/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3 (caffeine)



Check character

InChIKev=RYYVLZVUVIJVGH-UHFFFAOYAW

Flag character

Indicates InChI version, presence/absence of fixed H layer, isotopes, and stereochemistry

First block (14 letters)

Encodes molecular skeleton (connectivity)

Second block (8 letters)

Encodes proton positions (tautomers), stereochemistry, isotopes, reconnected layer

Use of InChIKey also allows searches based solely on atomic connectivity (first 14 characters). Software for generating InChIKey is available from the IUPAC website (Ref. 1).

The enormous databases compiled by organisations such as PubChem (Ref. 4), the US National Cancer Institute (NCI), and ChemSpider (Ref. 5) contain millions of InChIs and InChIKeys, which allow sophisticated searching of these collections. PubChem provides InChI-based structure-search facilities for both identical and similar structures (Ref. 6), and ChemSpider offers both search facilities and web services enabling a variety of InChI and InChIKey conversions (Ref. 7). The NCI Chemical Structure Lookup Service (Ref. 8) provides InChI-based search access to over 39 million chemical structures from over 80 different public and commercial data sources.

References

1. http://www.iupac.org/inchi
2. http://wwmm.ch.cam.ac.uk/inchifaq/
3. *Pure Appl. Chem.*, in preparation.
4. http://pubchem.ncbi.nlm.nih.gov
5. http://www.chemspider.com
6. http://pubchem.ncbi.nlm.nih.gov/search
7. http://www.chemspider.com/InChI.asmx
8. http://cholla.chemnavigator.com/cgi-bin/lookup/new/search