CHAPTER 14 THE IUPAC INTERNATIONAL CHEMICAL IDENTIFIER (InChI)

14.1 IUPAC INTERNATIONAL CHEMICAL IDENTIFIERS

Chemists use diagrammatic representations to convey structural information, and these are sometimes supplemented by verbal descriptions of structure. Conventional chemical nomenclature is a means of specifying a chemical structure in words, and systematic nomenclature provides an unambiguous description of a structure, a diagram of which can be reconstructed from its systematic name. The **IUPAC International Chemical Identifier**, or **InChI,** which is currently being developed, is a machine-readable string of symbols which enables a computer to represent the compound in a completely unequivocal manner. InChIs are produced by computer from structures drawn on-screen, and the original structure can be regenerated from an InChI with appropriate software. An InChI is not directly intelligible to the normal human reader, but InChIs will in time form the basis of an unequivocal and unique data base of all chemical compounds.

There is more than one way to specify molecular structures, and those based on 'connection tables' (specifications of atomic connectivities) are more suitable for processing by computer than conventional nomenclature, as they are matrix representations of molecular graphs, readily governed and handled by graph theory. This does not imply that traditional IUPAC nomenclature will eventually be displaced by computer methods and the continued development of verbal nomenclature has run in parallel with the development of InChIs.

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 computerised representation of a molecular structure diagram 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 InChI and of the software for its generation are available from the IUPAC website (reference 1), and further information can be obtained from the website of the InChI Trust(reference 2), a consortium of journal publishers, database providers and chemical software developers constituted in 2009 to provide direction and funding for ongoing development of the InChI standard.

A full account of the InChI project is in preparation (reference 3). Commercial structure-drawing software that will generate the Identifier is available from several organisations, which are listed on the IUPAC website (reference 1).

14.2 SHORT DESCRIPTION OF THE InChI

The conversion of structural information to its InChI is based on a set of IUPAC structure conventions and the rules for normalisation and canonicalisation (conversion to a single, predictable sequence) of a structure representation. The resulting InChI is simply a series of characters that serve to identify uniquely the structure from which it was derived. This conversion of a graphical representation of a chemical substance into the unique InChI character string can be carried out automatically by anyone using the freely available programs, and the facility can be built into any program dealing with chemical structures. The InChI uses a layered format to represent all the 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, if there are any ambiguities or uncertainties in the original structure representation, these will remain in the InChI.

This layered structure design of an InChI 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 hydrogen atoms fixed. This can work at a finer level. For example, if one author includes a double bond and tetrahedral stereochemistry, but another omits stereochemistry, the InChI for the latter description will be contained within that for the former.

14.3 THE STRUCTURE OF InChIs

The successive layers of an InChI are characterised as follows.

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

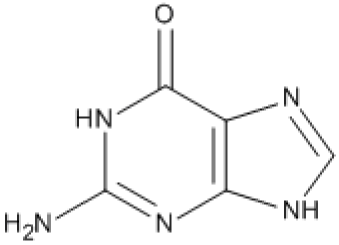
Note that charges are not considered within the basic InChI, but are added at the end of the InChI string, as in Example 2 below.

Two examples of InChI representations are given below. However, it is important to recognise 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.

Examples

1.

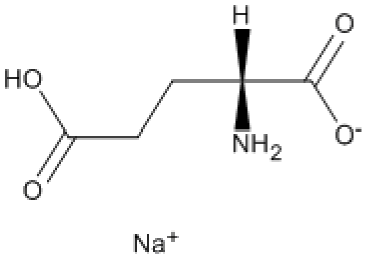
Name: guanine



The InChI for this structure is:
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

2.

Name: monosodium glutamate



The InChI for this structure is:
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 slash, /, followed by a lower-case letter (except for the first layer, the chemical formula), with the layers arranged in a predefined order. In the Examples above the following segments are included:

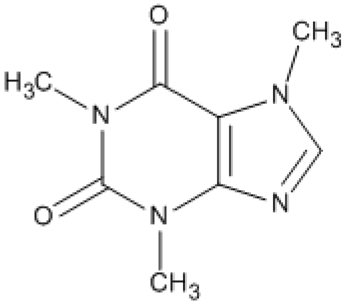
InChI version number
/chemical formula
/c connectivity-1.1 (excluding terminal H)
/h connectivity-1.2 (locations of terminal H, including mobile H attachment points)
/q charge
/p proton balance
/t tetrahedral parity
/m parity inverted to obtain relative stereo (1 = inverted, 0 = not inverted)
/s stereo 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)

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 an InChIKey. The InChIKey is a 27-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. The usefulness of the InChIKey as a search tool is enhanced if it is derived from a 'standard' InChI, *i.e.*, an InChI produced with standard option settings for features such as tautomerism and stereochemistry.

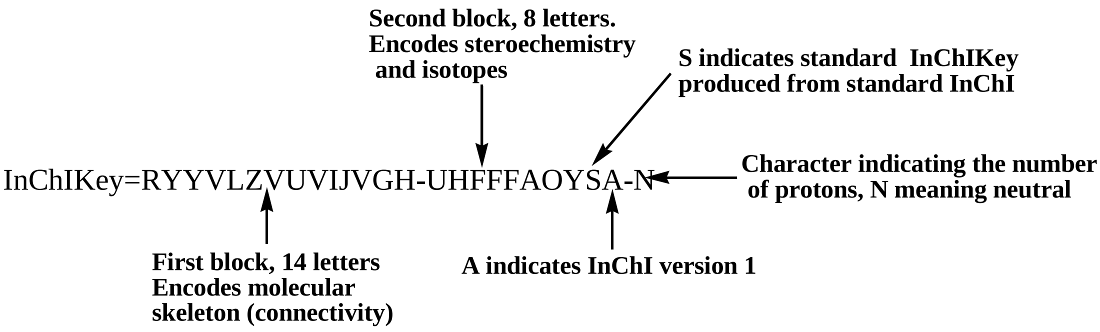
An Example is shown below; the standard InChI is denoted by the letter S after the version number. Use of the InChIKey also allows searches based solely on atomic connectivity (first 14 characters). The software for generating InChIKey is also available from the IUPAC website (reference 1).

Example 3:

Name: caffeine



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


Second block, 8 letters.
Encodes stereochemistry and isotopes
S indicates standard InChIKey produced from standard InChI
InChIKey=RYYVLZVUVIJVGH-UHFFFAOYSA-N
Character indicating the number of protons, N meaning neutral
First block, 14 letters Encodes molecular skeleton (connectivity)
A indicates InChI version 1

The enormous databases compiled by organisations such as PubChem (reference 4), the US National Cancer Institute, and ChemSpider (reference 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) (reference 6), and ChemSpider offers both search facilities and web services enabling a variety of InChI and InChIKey conversions (reference 7). The NCI Chemical Structure Lookup Service (reference 8) provides InChI-based search access to over 39 million chemical structures from over 80 different public and commercial data sources. Current (January 2010) members of the InChI Trust include Elsevier, FIZ-Chemie Berlin, IUPAC, Nature, Royal Society of Chemistry, Thomson

Reuters, Wiley, Taylor & Francis, ACD Labs, Chemaxon, OpenEye, and Symyx, and this list is expected to grow.

In the age of the computer, the IUPAC International Chemical Identifier is an essential component of the chemist's armoury of information tools, allowing location and manipulation of chemical data with unprecedented ease and precision.

REFERENCES

1. http://www.iupac.org/inchi
2. http://www.inchi-trust.org
3. *Pure and Applied Chemistry*, 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://cactus.nci.nih.gov/cgi-bin/lookup/search