The IUPAC International Chemical Identifier (InChI)

by Stephen R. Heller and Alan D. McNaught

The properties and behaviors of chemical substances are generally interpreted and discussed in terms of their molecular structures, and to convey structural information, chemists use diagrammatic representations supplemented by verbal descriptions. In order to have a means of specifying or describing a chemical structure in words, conventional chemical nomenclature was developed.

Systematic nomenclature provides an unambiguous description of a structure; a diagram of which can be reconstructed from its systematic name. However, there are other means of specifying molecular structures. Those based on "connection tables" (coded specifications of atomic connectivities) are more suitable than conventional nomenclature for processing by computer, as they are matrix representations of molecular graphs readily governed and handled by graph theory. In parallel with its continued development of conventional nomenclature, IUPAC has developed a structural identifier that can be readily interpreted by computers, or more precisely, by computer algorithms.

The IUPAC International Chemical Identifier (InChI) is a freely available, nonproprietary 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, 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.[1] In addition, an unofficial, but helpful compilation of answers to frequently asked questions has been compiled by Nick Day of the Unilever Centre for Molecular Science Informatics as part of his Ph.D. project on the Chemical Semantic Web.[2] A full account of the InChI project is in preparation.[3] Commercial structure-drawing software that generates the Identifier is available from several organizations, listed on the IUPAC website.[1]

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. This conversion of a graphical representation of a chemical substance into the unique InChI character string can be carried out automatically by any organization, and the facility can be built into any program dealing with chemical structures.

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:

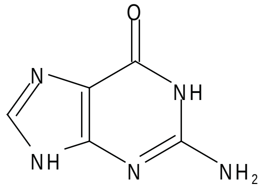
1. Formula
2. Connectivity (no formal bond orders)
   a. disconnected metals
   b. connected metals
3. Isotopes
4. Stereochemistry
   a. double bond (*Z*/*E*)
   b. tetrahedral (sp^3)
5. Tautomers (on or off)

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

The IUPAC International Chemical Identifier (InChI)

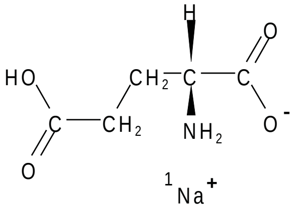
Two examples of InChI representations are given below. It is important to recognize, however, that InChI strings are intended for use by computers and end users need not understand any of their details. In fact, the open nature of InChI and its flexibility of representation, after implementation into 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



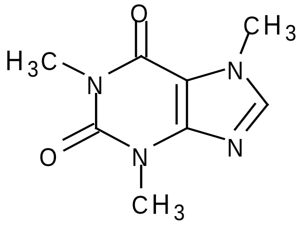
InChI=1/C5H9NO4.Na/c6-3(5(9)10)1-2-4(7)8;/h3H,1-2,6H2,(H,7,8)(H,9,10);/q;+1/p-1/t3-;/m1./s1/fC5H8NO4.Na/h7H;/q-1;m

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:

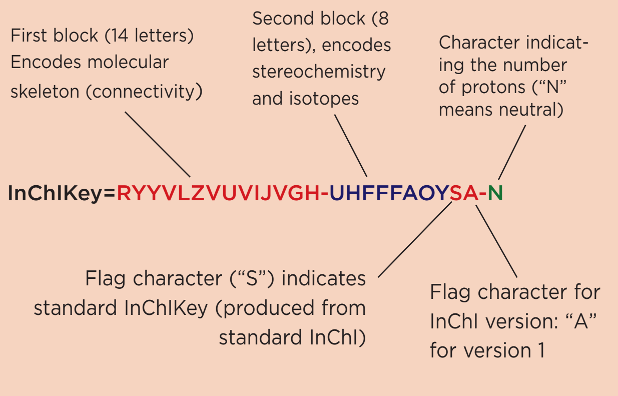
InChI version number
/ chemical formula
/c connectivity-1.1 (excluding terminal H)
/h connectivity-1.2 (locations of terminal H, including mobile H attachment points)
/q charge
/p proton balance
/t sp^3 (tetrahedral) parity
/m parity inverted to obtain relative stereo (1 = inverted, 0 = not inverted)
/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 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 by its derivation 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.

caffeine



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



First block (14 letters) Encodes molecular skeleton (connectivity)

Second block (8 letters), encodes stereochemistry and isotopes

Character indicating the number of protons ("N" means neutral)

InChIKey=RYYVLZVUVIJVGH-UHFFFAOYSA-N

Flag character ("S") indicates standard InChIKey (produced from standard InChI)

Flag character for InChI version: "A" for version 1

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

The enormous databases compiled by organizations such as PubChem,[4] the U.S. National Cancer Institute, and ChemSpider[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),[6] and ChemSpider offers both search facilities and web services enabling a variety of InChI and InChIKey conversions.[7] The NCI Chemical Structure Lookup Service[8] provides InChI-based search access to over 39 million chemical structures from over 80 different public and commercial data sources.

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

References

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

Alan McNaught <mcnaught@ntlworld.com>, retired from RSC, is one of the InChI's fathers; with broad expertise in publication and nomenclature, he has been involved in IUPAC activities for many years (including ICTNS, CPEP, and Div VIII) and with InChI since day one. Steve Heller <steve@hellers.com>, from NIST, is also a father of InChI, stimulating development and making the identifier known to the community.

www.iupac.org/inchi/

Use of InChI and InChIKey in the XML Gold Book

The *IUPAC Compendium of Chemical Terminology* (aka Gold Book) is a valuable resource to all chemists. It contains definitions of many chemistry-related terms and, thus, drawings of many chemical structures. In producing the XML version of the Gold Book we use InChI both internally and as meta-data on Gold Book pages to enable search engines to index this information.

As early adopters of InChI, we started to include InChIs of molecules in pages of individual terms in 2006. The InChIs are hidden from users, but are visible to search engines.

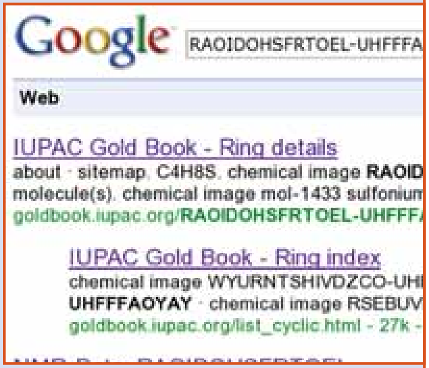
Thus, it is possible to reach appropriate Gold Book pages by searching for InChI or InChIKey code using any popular search engine.

To provide chemists with as many ways to navigate the website as possible, we created a few chemistry-related indexes. For this task, InChI was an invaluable tool that saved us much time and effort because it enabled us to compare structures in different entries using a simple text comparison.

Maybe the most interesting of the available chemical indexes in the Gold Book is the ring index. In this case, we not only used InChI to compare rings extracted from individual molecules using our in-house tools, but we used InChIKey to name files of individual rings. In this way, we solved a problem of giving the files useful and unique names while creating yet another way to make our structures visible to the outside.

For questions/comments, please contact Bedrich Košata <Bedrich.Kosata@vscht.cz>.

http://goldbook.iupac.org/



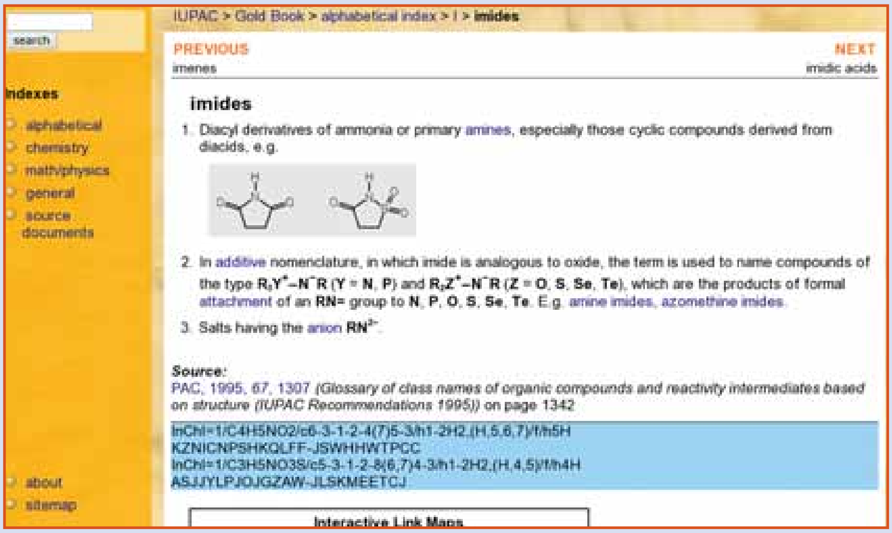
Google RAOIDOHSFRTOEL-UHFFFA

Web

IUPAC Gold Book - Ring details
about · sitemap, C4H8S, chemical image RAOID
molecule(s), chemical image mol-1433 sulfonium
goldbook.iupac.org/RAOIDOHSFRTOEL-UHFFF

IUPAC Gold Book - Ring index
chemical image WYURNTSHIVDZCO-UHI
UHFFFAOYAY · chemical image RSEBUV
goldbook.iupac.org/list_cyclic.html - 27k -

Google search for the InChIKey of thiolane, Gold Book ring index takes the first position.



IUPAC > Gold Book > alphabetical index > I > **imides**

PREVIOUS
imenes

NEXT
imidic acids

imides

1. Diacyl derivatives of ammonia or primary amines, especially those cyclic compounds derived from diacids, e.g.
2. In additive nomenclature, in which imide is analogous to oxide, the term is used to name compounds of the type $R_3Y^+–N^-R$ (Y = N, P) and $R_2Z^+–N^-R$ (Z = O, S, Se, Te), which are the products of formal attachment of an RN= group to N, P, O, S, Se, Te. E.g. amine imides, azomethine imides.
3. Salts having the anion RN^{2-}.

Source:
PAC, 1995, 67, 1307 *(Glossary of class names of organic compounds and reactivity intermediates based on structure (IUPAC Recommendations 1995))* on page 1342

InChI=1/C4H5NO2/c6-3-1-2-4(7)5-3/h1-2H2,(H,5,6,7)/f/h5H
KZNICNPSHKQLFF-JSWHHWTPCC
InChI=1/C3H5NO3S/c5-3-1-2-8(6,7)4-3/h1-2H2,(H,4,5)/f/h4H
ASJJYLPJOJGZAW-JLSKMEETCJ

Interactive Link Maps

The blue rectangle highlights InChIs and InChIKeys (shown here in the beta-test version of the 25 characters) that are normally invisible to the user.