The IUPAC International Chemical Identifier

The chemical structure of a compound is its true "identifier" but structure representations are not unique or convenient for computers. The International Union of Pure and Applied Chemistry (IUPAC) has thus developed a method for generating a freely available, non-proprietary identifier for chemical substances that can be used in printed and electronic data sources, thus enabling easier linking of diverse data compilations and unambiguous identification of chemical substances.

The project to develop the IUPAC International Chemical Identifier (InChI)[1-7] was proposed in 2000 and approved in 2002.[8] Version 1 of the InChI system was launched in 2005. IUPAC decided to tackle this problem because the increasing complexity of molecular structures was making conventional naming procedures inconvenient, and because there was no suitable, openly available electronic format for exchanging chemical structure information over the Internet.

In a digital world structures are not ideal "names": there are too many ways to draw them, they are non-linear, and they are inconvenient. What was needed was an openly available, electronic format for exchanging chemical structure information over the Internet: a unique, linear identifier, or "digital signature". The InChI algorithm converts a chemical structure (in the form of its connection table) into a unique, alphanumeric string of characters. The program can also convert an InChI label back into a molecular structure. Two requirements must be fulfilled in doing this: different compounds must have different identifiers, with all the information needed to distinguish the structures; and any one compound must have only one identifier, including only the necessary information to identify that compound.

InChI is free, open source software, sponsored by IUPAC, implemented by the US National Institute of Standards and Technology (NIST), and distributed under the terms of the GNU Lesser General Public License. Any organization can use it, in either public or private databases. The source code and associated software, documentation, and licensing conditions can be downloaded free from the IUPAC website.[9] InChI is written in C and can be compiled on most systems. It can be packaged into a dll for Windows or a library for UNIX.[10]

Creation of an InChI

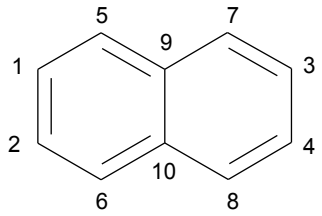
An InChI identifier is created from an input connection table (in molfile,[11] SDfile,[11] or CML[7] format) in three steps: normalization, canonicalization, and serialization. In the normalization step, electron density is ignored; salts and metal atoms in organometallic compounds are disconnected; and mobile hydrogens, and variable protonation and charge are normalized. The step is needed, for example, to remove variations in the ways of representing a nitro group.

In the canonicalization step, a set of atom labels is algorithmically generated that does not depend on how the structure was initially drawn; equivalent atoms get the same label. Bond orders and charge positions are ignored: connectivity alone is used. This does not introduce ambiguity as long as all hydrogen atoms are accounted for. Dmitrii Tchekhovskoi of NIST wrote the canonical numbering algorithm[12] by modifying a more recent version[13] of the well known Morgan algorithm.[14] In the final step the labeled structure is serialized and the InChI character string is output.

The identifier is hierarchically "layered"; each layer holds a distinct and separable class of structural information, with the layers ordered to provide successive structural refinement. There are currently six InChI layer types, each representing a different class of structural information: the main layer, a charge layer, a stereochemical layer, an isotopic layer, a fixed-H layer, and a reconnected layer. Except for the main layer (atoms and their bonds), the presence of a layer is not required and appears only when corresponding input information has been provided. Layers and sublayers are separated by the forward slash (/) delimiter. Except in the case of the chemical

formula sublayer of the main layer, each layer starts (after the slash mark) with a lower-case letter to indicate the type of information held.

Take, for example, naphthalene:



InChI=1/C10H8/c1-2-6-10-8-4-3-7-9(10)5-1/h1-8H

In the InChI, the first “1” refers to the version of the InChI software. (Note that this will actually be “1S” in the “standard InChI” version to be released soon with version 1.02.) The next segment of the string, C10H8, provides the molecular formula. The third segment is the connection table, which indicates how the atoms are connected. The last segment provides information about the placement of hydrogen atoms. Note that the identifier does not contain any information on the double bond positions.

Where relevant, stereochemical sublayers include sp^2, double bond stereochemistry, and sp^3, tetrahedral stereochemistry. Relative, absolute and racemic stereoisomers are distinguished. Stereochemistry can also be entered as “unknown” or as “unspecified”. Tautomers are dealt with by hydrogen atom migration between 1,3 heteroatoms.

Extension

Currently, the InChI algorithm can handle neutral and ionic organic molecules, radicals, and inorganic, organometallic, and coordination compounds. Since InChI is composed of hierarchical layers, new layers could be added to extend the scope of the identifier. Work is currently underway to extend InChI to include polymer representation. Extensions for other forms of stereochemistry, complex organometallics, (including coordinate bonds), other compound classes such as Markush structures, macromolecules, and conformations, and other attributes such as phases and excited states may be considered later. A project to allow this work to be carried out in an open source context has been registered with SourceForge.net.[15] Users are encouraged to report their experiences and any problems through the SourceForge website.

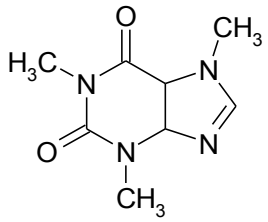
InChIKey

A beta-release of InChI version 1.02 was issued in September 2007. The principal new feature of this version was the introduction of a fixed length (25-character) condensed digital representation of the identifier known as InChIKey.[16] This key will facilitate web searching, previously complicated by unpredictable breaking of InChI character strings by search engines. It will also allow development of a web-based InChI lookup service; permit an InChI representation to be stored in fixed length fields; and make chemical structure database indexing easier.

In the formal release of version 1.02, due very soon, the InChIKey will be slightly modified and will actually be 27 characters long. The first part is 14 characters long and encodes the molecular skeleton (connectivity). After a hyphen, there is a second string of 10 characters, the first eight of which encode stereochemistry and isotopes. The first 23 characters of both versions of InChIKey are the same. In the post-beta version of the InChIKey, the 10-character block ends with a flag character indicating that this is a standard InChIKey (produced out of standard InChI) and a version character indicating the version number of InChI. The key ends with a hyphen followed by a character indicating [de]protonation state.

Both parts of the InChIKey are based on a truncated SHA-256 hash[17] of the corresponding InChI layers. For encoding of the data, only uppercase ASCII letters are used which ensures that the indexing engines will not split the data and also avoids case-sensitivity problems. There is a finite, but extremely small probability of finding two structures with the same InChIKey.

An example will make the structure of the key clearer. The “standard InChI” and InChIKey for caffeine are shown below. The first block of 14 letters (RYYVLZVUVIJVGH) encodes the molecular skeleton (connectivity). The first eight letters of the second block (UHFFFAOY) encode stereochemistry and isotopes. After that, “S” indicates that the key was produced from standard InChI and “A” indicates that version 1 of InChI was used. The final character, “N”, means “neutral”.



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

Use of InChIKey allows searches based solely on atom connectivity (the first 14 characters). For example, the stereoisomers D-fructose and L-fructose both have the same first block of 14 characters, BJHIKXHVCXFQLS.

Generating InChI

The PubChem Server Side Structure Editor v1.8 includes a facility for generating InChIs as the user draws the structure.[18] ACD/Labs’ freely available structure-drawing program ChemSketch[19] includes the facility to generate InChIs from drawn structures. Other structure drawing packages (MDL Draw, BKChem, ChemDraw, and Marvin) also allow an input chemical structure to be cut and pasted into the InChI Generator. ChemSpider provides methods to manipulate InChI strings and InChIKeys, including conversion to and from the molfile format, checking validity of the InChI identifiers, and searching ChemSpider using an input InChI.[20]

Some Other Identifiers

Readers will no doubt be familiar with CAS Registry Numbers.[21] InChI is not a registry system; it does not depend on the existence of a database of unique substance records to establish the next available sequence number for any new chemical substance being assigned an InChI. Registry systems which index the literature are complementary to any InChI databases that anyone creates. The Simplified Molecular Input Line Entry System (SMILES) language[22] is another well known way of representing a chemical structure by a string of characters. Like InChI, SMILES allows canonicalization of a structure, but SMILES is proprietary and not an open project. This has led to the use of different generation algorithms, and thus, different SMILES versions of the same compound have been found.[10]

Use of InChI and InChIKey

InChI is useful for communication between databases, merging data collections developed using different systems and protocols, maintaining a laboratory chemical inventory, and passing the “identity” of a substance to a colleague. The program could also be useful for chemical suppliers, by giving greatly increased exposure to their catalogs. The scientific and medical community can

use the InChIKey as a universal chemical identifier: InChI can be freely searched using Internet search engines. The InChIKey will allow commercial chemical information providers to have a free structure and number linking system. Millions of InChIs have already been created; at least 21 databases now incorporate them. The Royal Society of Chemistry (RSC) uses InChI in Project Prospect,[23] the aim of which is to make the science within RSC journal articles machine-readable through semantic enrichment, the integration of metadata into text. Text mining is used to attach structural information (InChI, SMILES and CML) to chemical names. A list of software developers, database providers, and journal publishers incorporating InChI in their products is maintained on the IUPAC website.[24]

Acknowledgment

I am grateful to the InChI project team, Steve Heller, Alan McNaught, Igor Pletnev, Steve Stein, and Dmitrii Tchekhovskoi, for their helpful comments and suggestions, also to Beda Kosata for his advice on the IUPAC website.

References

1. The IUPAC International Chemical Identifier (InChI). http://www.iupac.org/inchi/ (accessed January 7, 2009).

2. Stein, S. E.; Heller, S. R.; Tchekhovskoi, D. An Open Standard for Chemical Structure Representation: The IUPAC Chemical Identifier. In *Proceedings of the 2003 International Chemical Information Conference (Nîmes)*, Infonortics: Tetbury, UK, 2003; pp. 131-143.

3. Kosata, B. A Website Dedicated to the International Chemical Identifier. http://www.inchi.info/ (accessed December 31, 2008).

4. Heller, S. R.; Pletnev I. InChI and the InChIKey – A Status Report. http://www.hellers.com/steve/pub-talks/google-1007/frame.htm (accessed December 31, 2008).

5. Heller, S. R.; Stein, S. IUPAC InChI. http://www.hellers.com/steve/pub-talks/google1-1106/frame.htm (accessed December 31, 2008).

6. Heller, S. R.; Stein, S. Video about InChI. http://video.google.com/videoplay?docid=-6653695245776470969&q=heller+chemical (accessed December 31, 2008).

7. Coles, S. J.; Day, N. E.; Murray-Rust, P.; Rzepa, H. S.; Zhang, Y. Enhancement of the Chemical Semantic Web Through the Use of InChI Identifiers. *Org. Biomol. Chem.* **2005**, *3*(10), 1832-1834.

8. InChI Project Website. http://iupac.org/projects/2000/2000-025-1-800.html (accessed January 7, 2009).

9. InChI can be downloaded from http://iupac.org/inchi/download/index.html (accessed January 7, 2009).

10. Day, N. Unofficial InChI FAQ. http://wwmm.ch.cam.ac.uk/inchifaq/ (accessed December 31, 2008).

11. MDL CTfile formats. http://www.mdl.com/solutions/white_papers/ctfile_formats.jsp (accessed December 31, 2008).

12. IUPAC Project Meetings: Extensible Markup Language (XML) Data Dictionaries and Chemical Identifier. http://warr.com/inchi.pdf (accessed December 31, 2008).

13. McKay, B. D. Practical Graph Isomorphism. *Congressus Numerantium* **1981**, *30*, 45–87.

14. Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures - a Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5*, 107-113.

15. InChI project. http://sourceforge.net/projects/inchi/ (accessed December 31, 2008).

16. 16. InChIKey Overview. http://www.inchi.info/inchikey_overview_en.html (accessed December 31, 2008).

17. Secure Hash Standard. http://csrc.nist.gov/publications/fips/fips180-2/fips180-2withchangenotice.pdf (accessed December 31, 2008).

18. PubChem Server Side Structure Editor. http://pubchem.ncbi.nlm.nih.gov/edit/ (accessed December 31, 2008).

19. *ChemSketch*. http://www.acdlabs.com/download/chemsk.html (accessed December 31, 2008).

20. ChemSpider InChI services. http://www.chemspider.com/inchi.asmx (accessed December 31, 2008).

21. CAS REGISTRY and CAS Registry Numbers. http://www.cas.org/expertise/cascontent/registry/regsys.html (accessed December 31, 2008).

22. SMILES - A Simplified Chemical Language. http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html (accessed December 31, 2008).

23. RSC Project Prospect http://www.rsc.org/publishing/journals/projectprospect/faq.asp (accessed December 31, 2008).

24. InChI Use by Software Developers, Database Providers, and Journal Publishers. http://old.iupac.org/inchi/adopters.html. (accessed December 31, 2008).

Dr. Wendy A. Warr, Wendy Warr & Associates (wendy@warr.com, http://www.warr.com),
December 2008