The IUPAC International Chemical Identifier:

InChI—A New Standard for Molecular Informatics

by Alan McNaught

The emergence of computerized information-handling systems has had an enormous impact on chemistry and chemists. The ease with which chemical information can be shuttled around the world is phenomenal. Nevertheless, we are only just beginning to exploit the huge potential of the computer for sharing and processing such information. A major stumbling block has been the lack of agreement on standard ways of structuring and encoding molecular information (i.e., chemical structures and properties). Progress in this area has been disappointingly slow. Although work towards a standard format for chemical structure files has been discussed extensively during the past decade, it has been inhibited by various technical and political factors. However, the widespread availability of the Internet and IUPAC's increasing interest in these problems have now helped create an environment where progress can be made.

There are many ways of specifying the identity of a chemical compound. Chemical identifiers can be information poor, carrying no information about molecular structure (e.g., a registry number), or information rich, allowing the structure to be deduced (e.g., a systematic name or a computerized representation of bonding). Naming systems are internationally agreed (through IUPAC), but hitherto there has been no successful attempt to establish an agreed unique computerized representation for any molecular structure. There are several file formats in common use offering various approaches to uniqueness, but these are proprietary, and generally geared to specific applications for their owners. Furthermore, as molecular structures of interest to researchers in chemistry become more and more complex, our ability to devise nomenclature systems giving compact and intelligible names is being severely challenged.

An IUPAC strategy meeting in March 2000 at the National Academy of Sciences in Washington, D.C., USA, brought together a broad spectrum of providers and users of chemical information to discuss future requirements for nomenclature and other ways of designating chemical compounds. The need for a computerized equivalent of an IUPAC name (i.e., a standard chemical identifier) was recognized, and after some exploratory studies, including a September 2000 consultative meeting in Cambridge, UK, with representatives from a number of interested organizations, a project to develop such an identifier was launched early in 2001. The project is described in detail on the IUPAC website.[1]

The work on the Chemical Identifier was carried out under IUPAC auspices by Dmitrii Tchekhovskoi, Steve Stein, and Steve Heller at the US National Institute of Standards and Technology (NIST). Their approach is to express a chemical structure in terms of five layers of information (connectivity, tautomeric, isotopic, stereochemical, and electronic). In the final representation the unique connectivity layer is essential, but the user can choose which other layers to keep. The InChI algorithm converts input structural information into the identifier in a three-step process: normalization (to remove redundant information), canonicalization (to generate a unique set of atom labels), and serialization (to give a string of characters). The procedure generates a different Identifier for every compound, but always gives the same identifier for a particular compound regardless of how the structure is input. Of course, the procedure is equally applicable to both known and as yet unknown compounds.

A PC-based, executable version of an InChI test algorithm was released in March 2002. This version was developed to deal with well-defined, covalently-bonded organic molecules (both neutral and ionic). It was given to testers in a form that would accept structure input in a commonly used format, and deliver data as tagged text. No problems were reported, and the InChI was received enthusiastically when presented at the Chemical Abstracts Service/IUPAC Conference on Chemical Identifiers and XML for Chemistry, held in Columbus, Ohio, USA, in July 2002. A further version of the software, with applicability expanded to deal with inorganic, organometallic, and coordination compounds, was presented at a meeting with potential users at NIST in November 2003. The meeting was intended to obtain further comments on desirable output formats, and in light of the feedback, version 1 of the InChI software was released in April 2005. An updated version was released in August 2006 (see IUPAC Wire, p. 23), along with a validation protocol for software developers to check the validity of output from applications incorporating the InChI algorithm.

Figure 1 shows the InChI strings for two examples:

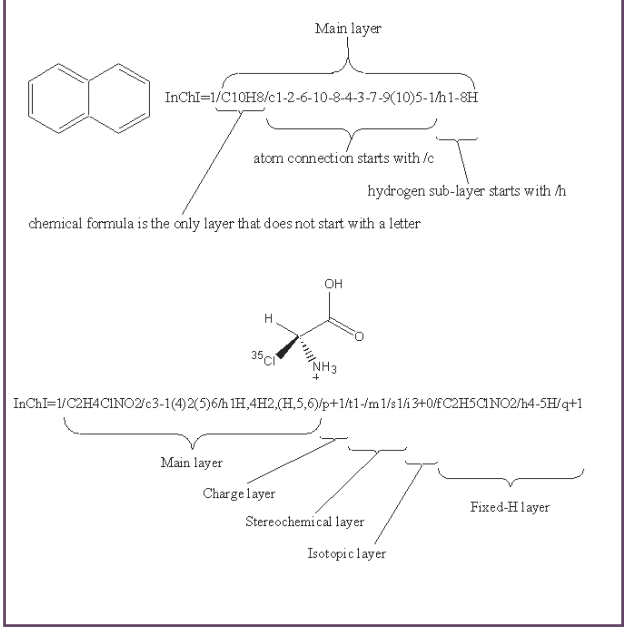

Main layer
InChI=1/C10H8/c1-2-6-10-8-4-3-7-9(10)5-1/h1-8H
atom connection starts with /c
hydrogen sub-layer starts with /h
chemical formula is the only layer that does not start with a letter
OH
H
O
^{35}Cl
NH$_3$
+
InChI=1/C2H4ClNO2/c3-1(4)2(5)6/h1H,4H2,(H,5,6)/p+1/t1-/m1/s1/i3+0/fC2H5ClNO2/h4-5H/q+1
Main layer
Charge layer
Stereochemical layer
Isotopic layer
Fixed-H layer

Figure 1. InChI strings for naphthalene and [^{35}Cl]chloro-L-glycinium.

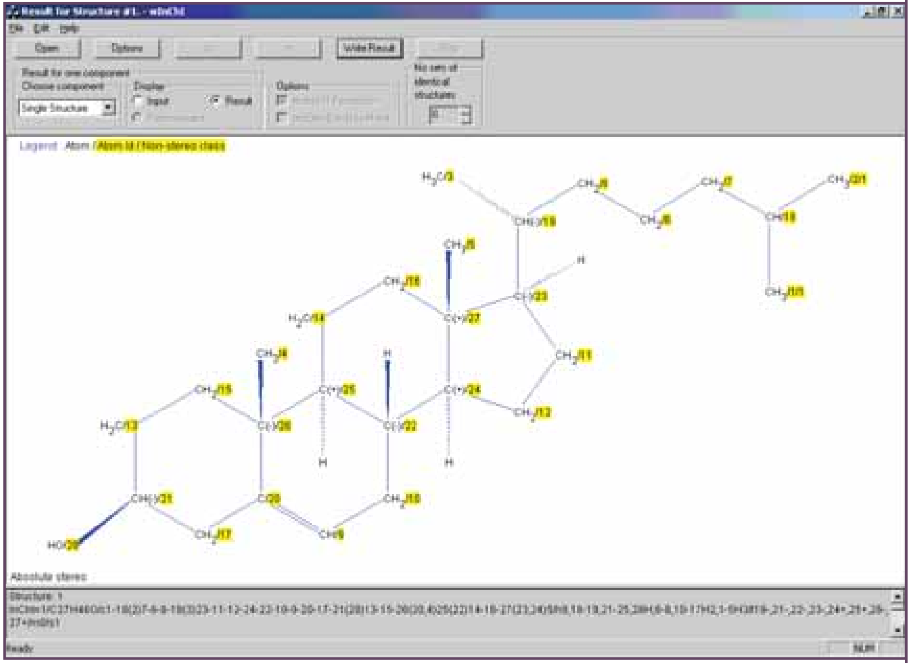
the unsubstituted naphthalene molecule and the isotopically substituted cation [^{35}Cl]chloro-L-glycinium, with the component layers indicated. A full explanation of the way in which layers are specified is given in the *InChI Technical Manual* distributed with the InChI software and on the InChI website at the University of Cambridge (UK).[2] The manual is an invaluable source of answers to InChI-related questions. Figure 2 shows the software's InChI display window, containing the structure, canonical numbering, and identifier for cholesterol.

For the International Chemical Identifier to fulfill its potential, software developers need to incorporate it into their products. InChI files can already be generated easily by using the freely available structure-drawing program ChemSketch,[3] and the PubChem database of the US National Institutes of Health offers an online "InChI-generation-as-you-draw" facility.[4] The Identifier has also been included as an integral component of Chemical Markup Language.[5] The potential for using InChI in Internet searching is highlighted in a recent article,[6] and other InChI-related articles are listed on the IUPAC website.[7] Anyone can easily obtain an InChI file at the desktop, or convert an InChI file back into a displayed structure.

The availability of this new standard will enable a wide variety of applications, such as:

- ordering chemicals from suppliers
- finding compounds in the chemical/patent/general literature via text-based search engines
- communication between databases
- merging data collections developed using different systems/protocols
- maintaining a laboratory chemical inventory or any broad-based local chemical collection
- passing the "identity" of a substance to a colleague for use in any of the above

Database providers have been among the first to recognize the enormous potential of InChI, and a list of these early adopters is provided on the IUPAC website.[8] As a result, millions of identifiers are available for searching on the web. At present, the largest collections are in the NIH/NCI database (~26 million), the NIH/PubChem database (~8 million), the Thomson/ISI database (~2 million), and the MDL/Elsevier database (>2 million). The freely accessible PubChem database of the US National Institutes of Health[9] also demonstrates the utility of InChI in structure searching, including both similarity and substructure searching[10]



Structure: 1
InChI=1/C27H46O/c1-18(2)7-6-8-19(3)23-11-12-24-22-10-9-20-17-21(28)13-15-26(20,4)25(22)14-16-27(23,24)5/h9,18-19,21-25,28H,6-8,10-17H2,1-5H3/t19-,21-,22-,23+,24+,25-,26-,27+/m0/s1

Figure 2. IUPAC International Chemical Identifier for cholesterol.

and a similar facility is provided by the National Cancer Institute's Chemical Structure Lookup Service, allowing the use of InChIs to search 78 databases containing a total of ~31 million entries.[11]

Publishers of all varieties of chemical information are recognizing the identifier as an essential way of "labelling" molecular data. We will all reap the benefits of a generally accepted convention for uniquely representing and communicating electronically the identity of any chemical substance.

References

1. www.iupac.org/projects/2000/2000-025-1-800.html
2. http://wwmm.ch.cam.ac.uk/inchifaq
3. www.acdlabs.com/products/chem_dsn_lab/chemsketch
4. http://pubchem.ncbi.nlm.nih.gov/edit
5. www.xml-cml.org
6. "Enhancement of the Chemical Semantic Web through the Use of InChI Identifiers," Simon J Coles, Nick E Day, Peter Murray-Rust, Henry S Rzepa and Yong Zhang, *Org. Biomol. Chem.*, 2005, 3, 1832–1834. (doi: 10.1039/b502828k)
7. www.iupac.org/inchi/articles.html
8. www.iupac.org/inchi/adopters.html
9. http://pubchem.ncbi.nlm.nih.gov
10. http://pubchem.ncbi.nlm.nih.gov/search
11. http://cactus.nci.nih.gov/lookup

Alan McNaught <adm@rsc.org> has been involved with InChI since the beginning. He is past president of the IUPAC Division on Chemical Nomenclature and Structure Representation (Division VIII) and formerly was general manager, production, at the Royal Society of Chemistry in Cambridge, UK.

www.iupac.org/inchi

Using InChI

by Jeremy G. Frey

The Southampton group has recently published several papers that make use of the IUPAC International Chemical Identifier (InChI). The InChI came along at a very convenient time for this group's research and became a key part of its e-Science Project[1] on computers to support the undertaking of chemical research[2] and new methodologies for dissemination of that research; bringing the Semantic Web or Web2.0 to the chemistry laboratory.

One of the major problems in chemistry is ensuring that chemical information is fully annotated to allow computers to facilitate the processing of this information. This is especially difficult when chemistry researchers are confronted with data overload, an increasingly common issue. Because of rapid advances in high-throughput chemistry and analysis, traditional approaches to the dissemination of data, or even the wide range of chemical databases available now, can not keep pace with the rate at which new data is generated. Therefore, it is proving ever more difficult to assess the validity of the information.

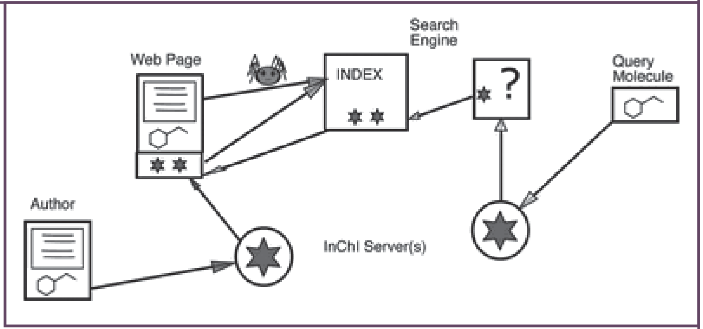
The InChI provides an excellent way of calculating a unique computer-readable identifier from a structure file, admittedly it is not an identifier that a person would wish to employ, but we have IUPAC names for that. It is even possible to use the InChI in a Google search to locate articles pertaining to specific molecules.[3]

Increasingly, the value of depositing data along with publications is understood as a way to promote its subsequent re-use and the information, supporting the provenance and enabling re-analysis. Some of this material can be stored as supplementary data at a journal site, but this does not usually support a rich enough description to ensure that the data can be found and accessed in a digital form.[4] The InChI works well in providing a link to chemical information stored in a repository. At Southampton, we initially experimented with using data repositories for crystallography data,[5] but now we are using a greater range of experimental data within the Repositories for the Laboratory (R4L) project.[6] The ability to correlate information on the same molecular species via the InChI makes for a very powerful approach. The National Crystallography Service deposits structures in a local version of the ecrystals archive and routinely provides the InChI.[7]

We have been investigating ways in which to store semantically annotated chemical information, describing the data items as fully as possible. For example, describing that a molecule has a melting point, recorded by a given method, reported by specified people, measured with a given uncertainty, and all recorded in a computer readable form using RDF, which is an XML-like approach that fully incorporates the ideas of unique identifiers to link together related information.[8,9] This approach enables us to automatically link items recorded in the electronic laboratory notebook to properties information about this entity. A similar underlying technology is used to record the information on both processes and properties.[10] This rich labelling is carried forward to model building, undertaken using the annotated data. One use of this approach is to track the impact of information subsequently discovered to be inaccurate.

We have also found the InChI useful in more local contexts. The e-Malaria project[11] is a system we have developed to teach chemical concepts to 16-18 year-old students by allowing them to use drug-design software running on a cycle steeling computational grid. They sketch a molecule, which is then converted into 3D so they can test its suitability as an anti-malarial drug by looking up its docking score with Gold software.[12] One of the aspects that interests the students is to know if someone else has run their molecule before. We can simply compare the InChI of the new molecule with all the ones stored in the database. Interesting issues concerning different stereoisomers can then arise as the molecular mechanics and quantum calculations that turn a 2D sketch into a 3D molecule do not always lead to the same 3D stereochemistry. As the InChI is a structured URI, a more complex comparison between two InChIs can be made by determining to what degrees they may match.

The InChI may still have a few problems. One which has caused some concern is that it is defined by the InChI program rather than an explicit algorithm. However, this program is widely available and the InChI has proved extremely valuable in enabling the linking up of annotated chemical data, providing a very good example of the "network effect," and potentially increasing the usefulness of any single data item added to the web.



Molecules, as defined by connectivity specified via InChI, are precisely indexed by major web search engines so that Internet tools can be transparently used for unique structure searches. Reprinted from reference 3, by permission of the Royal Society of Chemistry.

References

1. www.rcuk.ac.uk/escience
2. www.combechem.org
3. Coles, S.J., Day, N.E., Murray-Rust, P., Rzepa, H.S. and Zhang, Y. (2005) "Enhancement of the Chemical Semantic Web through the Use of InChI Identifiers." *Org. Biomol. Chem.* 3(10), 1832–1834. (doi:10.1039/b502828k)
4. Rousay, E.R., Fu, H., Robinson, J.M., Essex, J.W. and Frey, J.G. (2005) "Grid-Based Dynamic Electronic Publication: A Case Study Using Combined Experiment and Simulation Studies of Crown Ethers at the Air/Water Interface." *Philosophical Transactions of the Royal Society A: Mathematical Physical and Engineering Sciences* 363, (1833), 2075–2095. (doi: 10.1098/rsta.2005.1630)
5. http://ecrystals.chem.soton.ac.uk
6. http://r4l.eprints.org/about.html
7. Coles, S.J., Frey, J.G., Hursthouse, M.B., Light, M.E., Milsted, A.J., Carr, L.A., De Roure, D., Gutteridge, C.J., Mills, H.R., Meacham, K.E., Surridge, M., Lyon, E., Heery, R., Duke, M. and Day, M. (2006) "An E-Science Environment for Service Crystallographys from Submission to Dissemination." *J. Chem. Inf. Model.* 46(3), 1006–1016 (doi:10.1021/ci050362w); Coles, S., Frey, J.G., Hursthouse, M.B., Light, M.E., Meacham, K.E., Marvin, D.J. and Surridge, M. (2005) "ECSES—Examining Crystal Structures Using `e-science': A Demonstrator Employing Web and Grid Services to Enhance User Participation in Crystallographic Experiments." *J. Appl. Cryst.* 38(5), 819–826 (doi:10.1107/ S0021889805025197)
8. Taylor, K.R., Gledhill, R., Essex, J.W., Harris, S.W., De Roure, D.C. and Frey, J.G. (2006) "Bringing Chemical Data Onto the Semantic Web." *J. Chem. Inf. Model.* 46(3), 939–952. (doi:10.1021/ci050378m)
9. Taylor, K., Essex, J.W., Frey, J.G., Mills, H.R., Hughes, G. and Zaluska, E.J. (2006) "The Semantic Grid and Chemistry: Experiences with CombeChem." *Web Semantics* 4(2), 84–101. (doi:10.1016/j.websem.2006.03.003)
10. Hughes, G., Mills, H., De roure, D., Frey, J.G., Moreau, L., Schraefel, M.C., Smith, G. and Zaluska, E. (2004) "The Semantic Smart Laboratory: A System for Supporting the Chemical eScientist." *Org. Biomol. Chem.* 2(22), 3284–3293. (doi:10.1039/B410075A)
11. Gledhill, R., Kent, S., Hudson, B., Richards, W.G., Essex, J.W. and Frey, J.G. (2006) "A computer-Aided Drug Discovery System for Chemistry Teaching." *J. Chem. Inf. Model.* 46(3), 960–970. (doi:10.1021/ci050383q)
12. The Gold software was provided by Cambridge Crystallographic Data Centre (CCDC) for use with this teaching project.

Jeremy G. Frey <j.g.frey@soton.ac.uk> is a professor at the School of Chemistry, University of Southampton, Southampton, SO17 1BJ, UK. He is a member of the IUPAC Physical and Biophysical Chemistry Division and is chairman of the Commission on Physicochemical Symbols, Terminology, and Units.