WIKIPEDIA

International Chemical Identifier

The IUPAC **International Chemical Identifier** (**InChI** /ˈɪntʃiː/ *IN-chee* or /ˈɪŋkiː/ *ING-kee*) is a textual identifier for chemical substances, designed to provide a standard way to encode molecular information and to facilitate the search for such information in databases and on the web. Initially developed by IUPAC (International Union of Pure and Applied Chemistry) and NIST (National Institute of Standards and Technology) from 2000 to 2005, the format and algorithms are non-proprietary.

The continuing development of the standard has been supported since 2010 by the not-for-profit **InChI Trust**, of which IUPAC is a member. The current software version is 1.06 and was released in December 2020.

Prior to 1.04, the software was freely available under the open-source LGPL license,[3] but it now uses a custom license called IUPAC-InChI Trust License.[4]

InChI

Developer(s)	InChI Trust
Initial release	April 15, 2005[1][2]
Stable release	1.06 / December 2020
Operating system	Microsoft Windows and Unix-like
Platform	IA-32 and x86-64
Available in	English
License	IUPAC / InChI Trust Licence
Website	https://www.inchi-trust.org/ (https://www.inchi-trust.org/)

Contents

Overview

The identifiers describe chemical substances in terms of *layers* of information — the atoms and their bond connectivity, tautomeric information, isotope information, stereochemistry, and electronic charge information.[5] Not all layers have to be provided; for instance, the tautomer layer can be omitted if that type of information is not relevant to the particular application.

InChIs differ from the widely used CAS registry numbers in three respects: firstly, they are freely usable and non-proprietary; secondly, they can be computed from structural information and do not have to be assigned by some organization; and thirdly, most of the information in an InChI is human readable (with practice).

InChIs can thus be seen as akin to a general and extremely formalized version of IUPAC names. They can express more information than the simpler SMILES notation and differ in that every structure has a unique InChI string, which is important in database applications. Information about the 3-dimensional coordinates of atoms is not represented in InChI; for this purpose a format such as PDB can be used.

The InChI algorithm converts input structural information into a unique InChI identifier in a three-step process: normalization (to remove redundant information), canonicalization (to generate a unique number label for each atom), and serialization (to give a string of characters).

The InChIKey, sometimes referred to as a hashed InChI, is a fixed length (27 character) condensed digital representation of the InChI that is not human-understandable. The InChIKey specification was released in September 2007 in order to facilitate web searches for chemical compounds, since these were problematic with the full-length InChI.[6] Unlike the InChI, the InChIKey is not unique: though collisions can be calculated to be very rare, they happen.[7]

In January 2009 the 1.02 version of the InChI software was released. This provided a means to generate so called standard InChI, which does not allow for user selectable options in dealing with the stereochemistry and tautomeric layers of the InChI string. The standard InChIKey is then the hashed version of the standard InChI string. The standard InChI will simplify comparison of InChI strings and keys generated by different groups, and subsequently accessed via diverse sources such as databases and web resources.

Generation

In order to avoid generating different InChIs for tautomeric structures, before generating the InChI, an input chemical structure is normalized to reduce it to its so-called core parent structure. This may involve changing bond orders, rearranging formal charges and possibly adding and removing protons. Different input structures may give the same result; for example, acetic acid and acetate would both give the same core parent structure, that of acetic acid. A core parent structure may be disconnected, consisting of more than one component, in which case the sublayers in the InChI usually consist of sublayers for each component, separated by semicolons (periods for the chemical formula sublayer.) One way this can happen is that all metal atoms are disconnected during normalization; so, for example, the InChI for tetraethyllead will have five components, one for lead and four for the ethyl groups.[5]

The first, main, layer of the InChI refers to this core parent structure, giving its chemical formula, non-hydrogen connectivity without bond order (`/c` sublayer) and hydrogen connectivity (`/h` sublayer.) The `/q` portion of the charge layer gives its charge, and the `/p` portion of the charge layer tells how many protons (hydrogen ions) must be added to or removed from it to regenerate the original structure. If present, the stereochemical layer, with sublayers `/b`, `/t`, `/m` and `/s`, gives stereochemical information, and the isotopic layer `/i` (which may contain sublayers `/h`, `/b`, `/t`, `/m` and `/s`) gives isotopic information. These are the only layers which can occur in a standard InChI.[5]

If the user wants to specify an exact tautomer, a fixed hydrogen layer `/f` can be appended, which may contain various additional sublayers; this cannot be done in standard InChI though, so different tautomers will have the same standard InChI (for example, alanine will give the same standard InChI whether input in a neutral or a zwitterionic form.) Finally, a nonstandard reconnected `/r` layer can be added, which effectively gives a new InChI generated without breaking bonds to metal atoms. This may contain various sublayers, including `/f`.[5]

Format and layers

InChI format

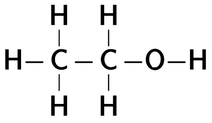
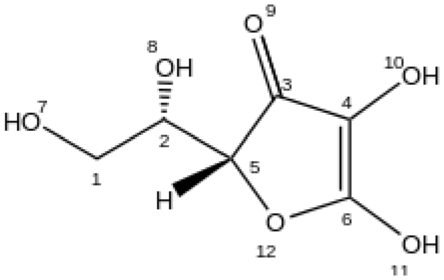
Internet media type	chemical/x-inchi
Type of format	chemical file format

Every InChI starts with the string "InChI=" followed by the version number, currently 1. If the InChI is standard, this is followed by the letter S for **standard InChIs**, which is a fully standardized InChI flavor maintaining the same level of attention to structure details and the same conventions for drawing perception. The remaining information is structured as a sequence of layers and sub-layers, with each layer providing one specific type of information. The layers and sub-layers are separated by the delimiter "/" and start with a characteristic prefix letter (except for the chemical formula sub-layer of the main layer). The six layers with important sublayers are:

1. Main layer
   - Chemical formula (no prefix). This is the only sublayer that must occur in every InChI.
   - Atom connections (prefix: "c"). The atoms in the chemical formula (except for hydrogens) are numbered in sequence; this sublayer describes which atoms are connected by bonds to which other ones.
   - Hydrogen atoms (prefix: "h"). Describes how many hydrogen atoms are connected to each of the other atoms.
2. Charge layer
   - charge sublayer (prefix: "q")
   - proton sublayer (prefix: "p" for "protons")
3. Stereochemical layer
   - double bonds and cumulenes (prefix: "b")
   - tetrahedral stereochemistry of atoms and allenes (prefixes: "t", "m")
   - type of stereochemistry information (prefix: "s")
4. Isotopic layer (prefixes: "i", "h", as well as "b", "t", "m", "s" for isotopic stereochemistry)
5. Fixed-H layer (prefix: "f"); contains some or all of the above types of layers except atom connections; may end with "o" sublayer; never included in standard InChI
6. Reconnected layer (prefix: "r"); contains the whole InChI of a structure with reconnected metal atoms; never included in standard InChI

The delimiter-prefix format has the advantage that a user can easily use a wildcard search to find identifiers that match only in certain layers.

Examples

Structural formula	standard InChI
 ethanol	InChI=1S/C2H6O/c1-2-3/h3H,2H2,1H3
 O 9 8 OH 10 OH HO 7 3 4 2 5 1 H 6 O 12 OH 11 L-ascorbic acid	InChI=1S/C6H8O6/c7-1-2(8)5-3(9)4(10)6(11)12-5/h2,5,7-8,10-11H,1H2/t2-,5+/m0/s1

InChIKey

The condensed, 27 character **InChIKey** is a hashed version of the full InChI (using the SHA-256 algorithm), designed to allow for easy web searches of chemical compounds.[6] The **standard InChIKey** is the hashed counterpart of **standard InChI**. Most chemical structures on the Web up to 2007 have been represented as GIF files, which are not searchable for chemical content. The full InChI turned out to be too lengthy for easy searching, and therefore the InChIKey was developed. There is a very small, but nonzero chance of two different molecules having the same InChIKey, but the probability for duplication of only the first 14 characters has been estimated as only one duplication in 75 databases each containing one billion unique structures. With all databases currently having below 50 million structures, such duplication appears unlikely at present. A recent study more extensively studies the collision rate finding that the experimental collision rate is in agreement with the theoretical expectations.[8]

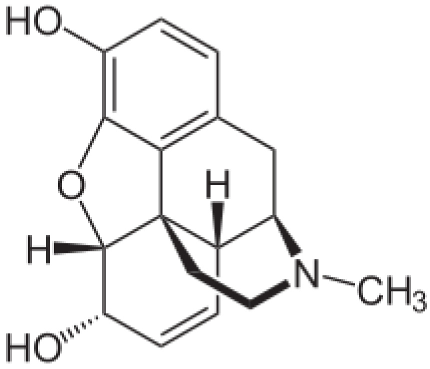
The InChIKey currently consists of three parts separated by hyphens, of 14, 10 and one character(s), respectively, like `XXXXXXXXXXXXXX-YYYYYYYYFV-P`. The first 14 characters result from a SHA-256 hash of the connectivity information (the main layer and `/q` sublayer of the charge layer) of the InChI. The second part consists of 8 characters resulting from a hash of the remaining layers of the InChI, a single character indicating the kind of InChIKey (`S` for standard and `N` for nonstandard), and a character indicating the version of InChI used (currently `A` for version 1.) Finally, the single character at the end indicates the protonation of the core parent structure, corresponding to the `/p` sublayer of the charge layer (`N` for no protonation, `O`, `P`, ... if protons should be added and `M`, `L`, ... if they should be removed.)[9][5]

Example

Morphine has the structure shown on the right. The standard InChI for morphine is `InChI=1S/C17H19NO3/c1-18-7-6-17-10-3-5-13(20)16(17)21-15-12(19)4-2-9(14(15)17)8-11(10)18/h2-5,10-11,13,16,19-20H,6-8H2,1H3/t10-,11+,13-,16-,17-/m0/s1` and the standard InChIKey for morphine is `BQJCRHHNABKAKU-KBQPJGBKSA-N`.[10]

InChI resolvers

As the InChI cannot be reconstructed from the InChIKey, an InChIKey always needs to be linked to the original InChI to get back to the original structure. InChI Resolvers act as a lookup service to make these links, and prototype services are available from National Cancer Institute, the UniChem service (https://www.ebi.ac.uk/unichem/) at the European Bioinformatics Institute, and PubChem. ChemSpider has had a resolver until July 2015 when it was decommissioned.[11]



Morphine structure

Name

The format was originally called IChI (IUPAC Chemical Identifier), then renamed in July 2004 to INChI (IUPAC-NIST Chemical Identifier), and renamed again in November 2004 to InChI (IUPAC International Chemical Identifier), a trademark of IUPAC.

Continuing development

Scientific direction of the InChI standard is carried out by the IUPAC Division VIII Subcommittee, and funding of subgroups investigating and defining the expansion of the standard is carried out by both IUPAC and the InChI Trust. The InChI Trust funds the development, testing and documentation of the InChI. Current extensions are being defined to handle polymers and mixtures, Markush structures, reactions[12] and organometallics, and once accepted by the Division VIII Subcommittee will be added to the algorithm.

Software

The InChI Trust has developed software to generate the InChI, InChIKey and other identifiers. The release history of this software follows.[13]

Software and version	Date	License	Comments
InChI v. 1	April 2005		
InChI v. 1.01	August 2006		
InChI v. 1.02beta	Sep. 2007	LGPL 2.1	Adds InChIKey functionality.
InChI v. 1.02	Jan. 2009	LGPL 2.1	Changed format for InChIKey. Introduces standard InChI.
InChI v. 1.03	June 2010	LGPL 2.1	
InChI v. 1.03 source code docs	March 2011		
InChI v. 1.04	Sep. 2011	IUPAC/InChI Trust InChI Licence 1.0	New license. Support for elements 105-112 added. CML support removed.
InChI v. 1.05	Jan. 2017	IUPAC/InChI Trust InChI Licence 1.0	Support for elements 113-118 added. Experimental polymer support. Experimental large molecule support.
RInChI v. 1.00	March 2017	IUPAC/InChI Trust InChI Licence 1.0, and BSD-style	Computes reaction InChis.[12]
InChI v. 1.06	Dec. 2020	IUPAC/InChI Trust InChI Licence 1.0	Revised polymer support.

Adoption

The InChI has been adopted by many larger and smaller databases, including ChemSpider, ChEMBL, Golm Metabolome Database, OpenPHACTS, and PubChem.[14] However, the adoption is not straightforward, and many databases show a discrepancy between the chemical structures and the InChI they contain, which is a problem for linking databases.[15]

See also

- Molecular Query Language
- Simplified molecular-input line-entry system (SMILES)
- Molecule editor
- SYBYL Line Notation
- Bioclipse generates InChI and InChIKeys for drawn structures or opened files
- the Chemistry Development Kit uses JNI-InChI to generate InChIs, can convert InChIs into structures, and generate tautomers based on the InChI algorithms

Notes and references

1. "IUPAC International Chemical Identifier Project Page" (https://web.archive.org/web/20120527162256/http://www.iupac.org/home/projects/project-db/project-details.html?tx_wfqbe_pi1%5Bpr

oject_nr%5D=2000-025-1-800). *IUPAC*. Archived from the original (http://www.iupac.org/home/projects/project-db/project-details.html?tx_wfqbe_pi1%5bproject_nr%5d=2000-025-1-800) on 27 May 2012. Retrieved 5 December 2012.

2. Heller, S.; McNaught, A.; Stein, S.; Tchekhovskoi, D.; Pletnev, I. (2013). "InChI - the worldwide chemical structure identifier standard" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3599061). *Journal of Cheminformatics*. **5** (1): 7. doi:10.1186/1758-2946-5-7 (https://doi.org/10.1186%2F1758-2946-5-7). PMC 3599061 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3599061). PMID 23343401 (https://pubmed.ncbi.nlm.nih.gov/23343401).
3. McNaught, Alan (2006). "The IUPAC International Chemical Identifier:InChI" (http://www.iupac.org/publications/ci/2006/2806/4_tools.html). *Chemistry International*. **28** (6). IUPAC. Retrieved 2007-09-18.
4. http://www.inchi-trust.org/download/104/LICENCE.pdf
5. Heller, S.R.; McNaught, A.; Pletnev, I.; Stein, S.; Tchekhovskoi, D. (2015). "InChI, the IUPAC International Chemical Identifier" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4486400). *Journal of Cheminformatics*. **7**: 23. doi:10.1186/s13321-015-0068-4 (https://doi.org/10.1186%2Fs13321-015-0068-4). PMC 4486400 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4486400). PMID 26136848 (https://pubmed.ncbi.nlm.nih.gov/26136848).
6. "The IUPAC International Chemical Identifier (InChI)" (https://web.archive.org/web/20071030202540/http://www.iupac.org/inchi/release102.html). IUPAC. 5 September 2007. Archived from the original (http://www.iupac.org/inchi/release102.html) on October 30, 2007. Retrieved 2007-09-18.
7. E.L. Willighagen (17 September 2011). "InChIKey collision: the DIY copy/pastables" (http://chem-bla-ics.blogspot.nl/2011/09/inchikey-collision-diy-copypastables.html). Retrieved 2012-11-06.
8. Pletnev, I.; Erin, A.; McNaught, A.; Blinov, K.; Tchekhovskoi, D.; Heller, S. (2012). "InChIKey collision resistance: An experimental testing" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3558395). *Journal of Cheminformatics*. **4** (1): 39. doi:10.1186/1758-2946-4-39 (https://doi.org/10.1186%2F1758-2946-4-39). PMC 3558395 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3558395). PMID 23256896 (https://pubmed.ncbi.nlm.nih.gov/23256896).
9. "Technical FAQ - InChI Trust" (http://www.inchi-trust.org/technical-faq/#13.1). *inchi-trust.org*. Retrieved 8 Jan 2021.
10. "InChI=1/C17H19NO3/c1-18..." (http://www.chemspider.com/RecordView.aspx?id=5760) Chemspider. Retrieved 2007-09-18.
11. InChI Resolver, 27 July 2015, http://www.chemspider.com/InChiResolverDecommissioned.aspx
12. Grethe, Guenter; Blanke, Gerd; Kraut, Hans; Goodman, Jonathan M. (9 May 2018). "International chemical identifier for reactions (RInChI)" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4015173). *Journal of Cheminformatics*. **10** (1): 45. doi:10.1186/s13321-018-0277-8 (https://doi.org/10.1186%2Fs13321-018-0277-8). PMC 4015173 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4015173). PMID 24152584 (https://pubmed.ncbi.nlm.nih.gov/24152584).
13. Downloads of InChI Software (https://www.inchi-trust.org/downloads/), accessed Jan. 8, 2021.
14. Warr, W.A. (2015). "Many InChIs and quite some feat". *Journal of Computer-Aided Molecular Design*. **29** (8): 681–694. Bibcode:2015JCAMD..29..681W (https://ui.adsabs.harvard.edu/abs/2015JCAMD..29..681W). doi:10.1007/s10822-015-9854-3 (https://doi.org/10.1007%2Fs10822-015-9854-3). PMID 26081259 (https://pubmed.ncbi.nlm.nih.gov/26081259). S2CID 31786997 (https://api.semanticscholar.org/CorpusID:31786997).
15. Akhondi, S. A.; Kors, J. A.; Muresan, S. (2012). "Consistency of systematic chemical identifiers within and between small-molecule databases" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3539895). *Journal of Cheminformatics*. **4** (1): 35. doi:10.1186/1758-2946-4-35 (https://doi.org/10.1186%2F1758-2946-4-35). PMC 3539895 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3539895). PMID 23237381 (https://pubmed.ncbi.nlm.nih.gov/23237381).

External links

- IUPAC InChI site (http://www.iupac.org/inchi/)
- Description of the canonicalization algorithm (http://depth-first.com/articles/2006/08/12/inchi-canonicalization-algorithm)
- Googling for InChIs (http://lists.w3.org/Archives/Public/public-swls-ws/2004Oct/att-0019/) a presentation to the W3C.
- InChI Release 1.02 (https://web.archive.org/web/20100330213717/http://www.iupac.org/inchi/release102final.html) InChI final version 1.02 and explanation of Standard InChI, January 2009
- NCI/CADD Chemical Identifier Resolver (https://cactus.nci.nih.gov/chemical/structure) Generates and resolves InChI/InChIKeys and many other chemical identifiers
- PubChem online molecule editor (https://pubchem.ncbi.nlm.nih.gov/edit/index.html) that supports SMILES/SMARTS and InChI
- ChemSpider Compound APIs (https://developer.rsc.org/compounds-v1/apis) ChemSpider REST API that allows generation of InChI and conversion of InChI to structure (also SMILES and generation of other properties)
- MarvinSketch (https://web.archive.org/web/20070404073952/http://www.chemaxon.com/demosite/marvin/index.html) from ChemAxon, implementation to draw structures (or open other file formats) and output to InChI file format
- BKchem (http://bkchem.zirael.org/inchi_en.html) implements its own InChI parser and uses the IUPAC implementation to generate InChI strings
- CompoundSearch (http://www.compoundsearch.com) implements an InChI and InChI Key search of spectral libraries
- SpectraBase (http://www.spectrabase.com) implements an InChI and InChI Key search of spectral libraries
- JSME (http://peter-ertl.com/jsme/) is a free JavaScript based molecular editor that generates InChI and InChI Key in a web browser, which allows for easy web searches of chemical compounds

Retrieved from "https://en.wikipedia.org/w/index.php?title=International_Chemical_Identifier&oldid=1006804026"

This page was last edited on 14 February 2021, at 22:04 (UTC).