The IUPAC InChI project

Stephen Heller

InChI-Trust Project Director

steve@inchi-trust.org

The main web sites for the IUPAC InChI project are:

http://www.iupac.org/inchi

and

http://www.inchi-trust.org

The slides from this presentation can be found at:

http://www.hellers.com/steve/pub-talks/

11/2010

InChI TRUST

Objective

The IUPAC Chemical Identifier (InChI) is an open source, freely available, non-proprietary identifier for well defined chemical substances.

The InChI enables chemical information in electronic data sources (databases, registries, journals and repositories) to be machine readable.

Enabling easier LINKING of, and working with, diverse data and information compilations.

InChI have some advantages over other chemical identifiers developed before:

(1) They are freely useable and non-proprietary.

(2) They allow a more advanced representation of chemical information than other codes (such as the SMILES code).

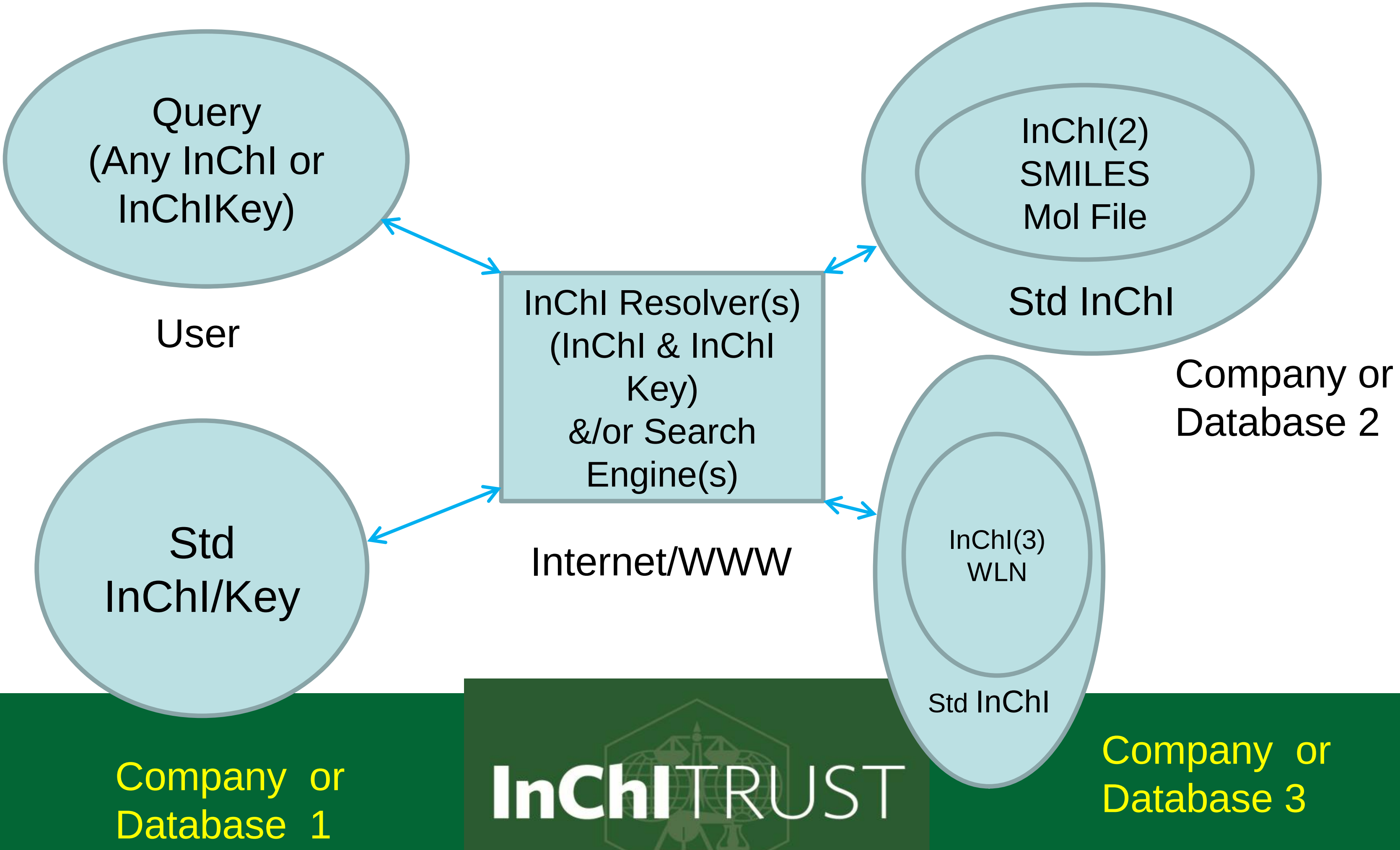
(3) They are unambiguous, i.e. conversion of chemical structures using standardized algorithms only leads to one InChI.

(4) They are precisely indexed by major search engines such as Google.

However, InChI are not applicable to generic formats often disclosed in patent literature, such as Markush structures, since they were rather designed to represent specific chemical structures and compounds. InChI therefore are not yet useful for comprehensive retrieval of patent literature.

Excerpt taken from:
Full-text prior art and chemical structure searching in e-journals and on the internet – A patent information professional's perspective
World Patent Information, Volume 31, Issue 4, December 2009, Pages 278-284
Maik Annies (Syngenta)

The LINKED and Interoperable and Combinable World of InChI



Query (Any InChI or InChIKey)

User

InChI Resolver(s) (InChI & InChI Key) &/or Search Engine(s)

Internet/WWW

InChI(2)
SMILES
Mol File

Std InChI

Company or Database 2

Std InChI/Key

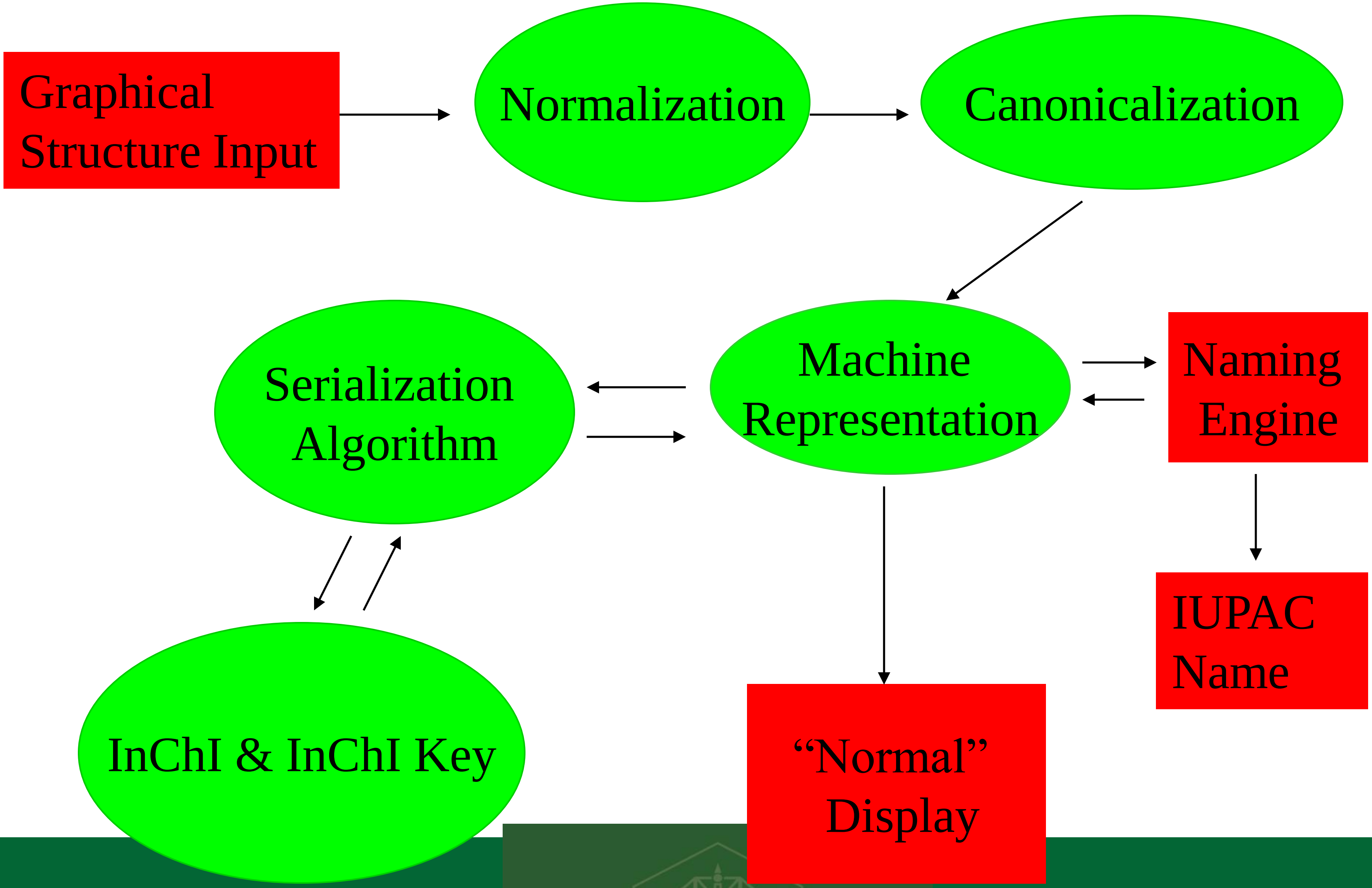
Company or Database 1

InChI(3)
WLN

Std InChI

Company or Database 3

InChI TRUST



Graphical Structure Input

Normalization

Canonicalization

Machine Representation

Naming Engine

IUPAC Name

Serialization Algorithm

InChI & InChI Key

“Normal” Display

InChI layered structure design

The current 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 (sp3)**
5. **Tautomers (on or off)**

Charges are added to end of the string

InChI Characteristics

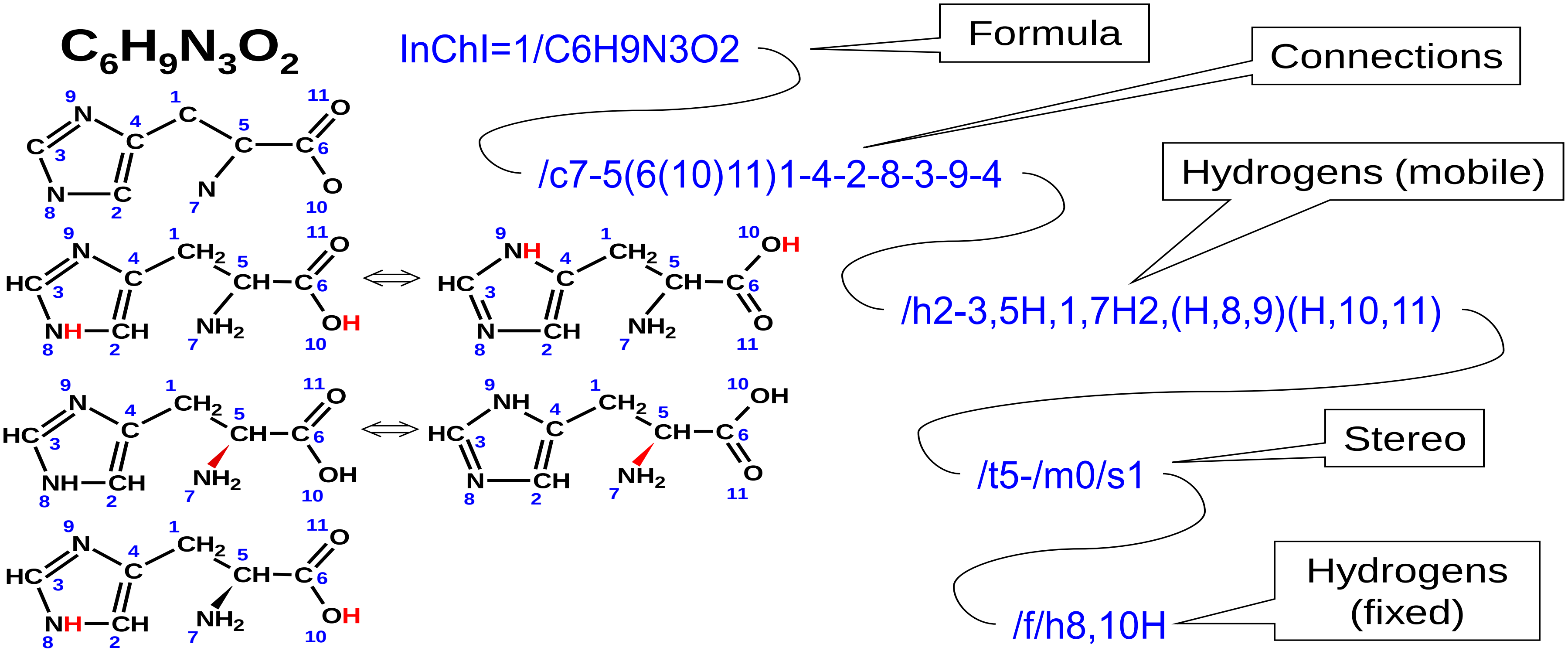
1. Easy to generate (It will use existing software.)

2. Expressive (It will contain structural information.)

3. Unique/Unambiguous

4. Easy to search for structure via Internet search engines (Google, Yahoo, Microsoft Live, etc.) using the InChI (hash) Key.

InChI Layers: L-Histidine



$C_6H_9N_3O_2$

InChI=1/C6H9N3O2

Formula

/c7-5(6(10)11)1-4-2-8-3-9-4

Connections

/h2-3,5H,1,7H2,(H,8,9)(H,10,11)

Hydrogens (mobile)

/t5-/m0/s1

Stereo

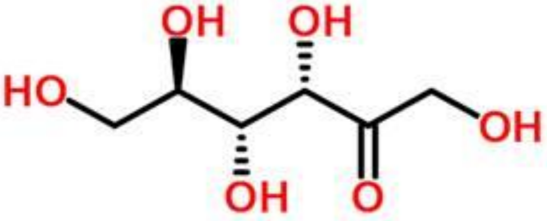
/f/h8,10H

Hydrogens (fixed)

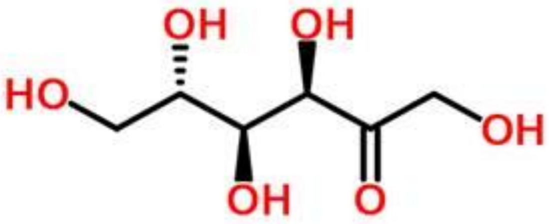
InChI=1/C6H9N3O2/c7-5(6(10)11)1-4-2-8-3-9-4/h2-3,5H,1,7H2,(H,8,9)(H,10,11)/t5-/m0/s1/f/h8,10H

InChIKey=HNDVDQJCIGZPNO-QLMCEAFFNA-N **InChIKey=HNDVDQJCIGZPNO-YFKPBYRVSA-N**

How does the InChI work?



D-Fructose



L-Fructose

D-Fructose (Natural)

InChI=1S/C6H12O6/c7-1-3(9)5(11)6(12)4(10)2-8/h3,5-9,11-12H,1-2H2/t3-,5-,6-/m1/s1

InChIKey: BJHIKXHVCXFQLS-UYFOZJQFSA-N

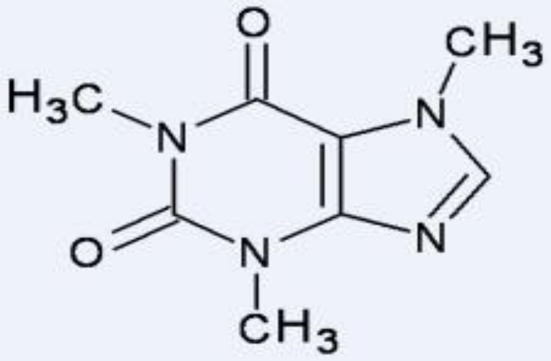
L-Fructose

InChI=1S/C6H12O6/c7-1-3(9)5(11)6(12)4(10)2-8/h3,5-9,11-12H,1-2H2/t3-,5-,6-/m0/s1

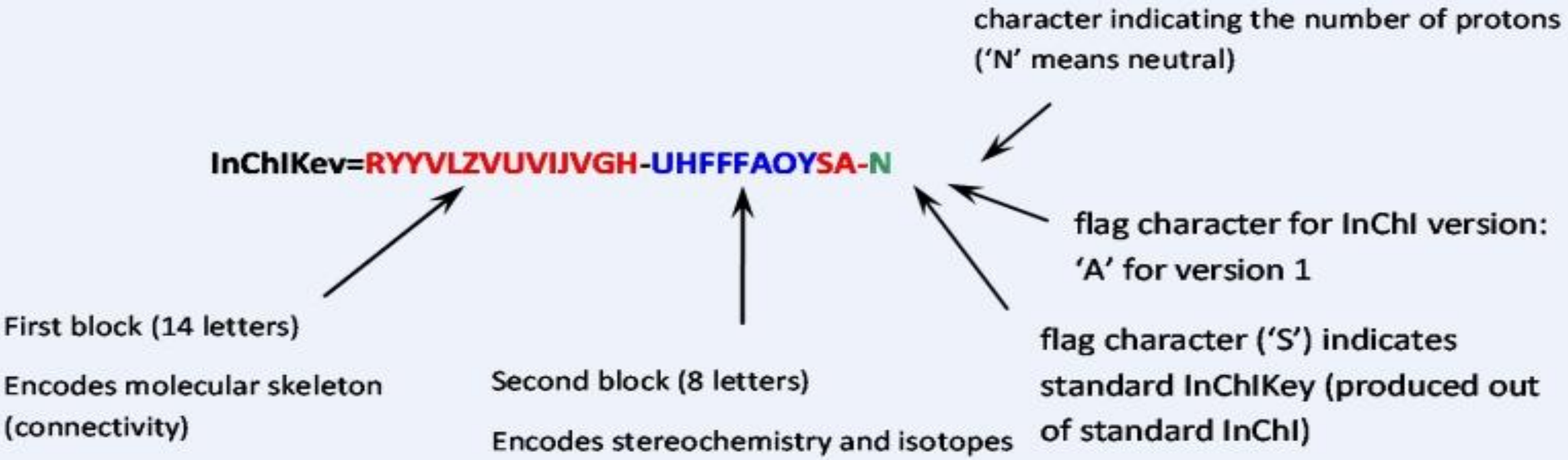
InChIKey: BJHIKXHVCXFQLS-FUTKDDECSA-N

Bar Codes – not designed to be read by humans

InChI – not designed to be read by humans. It is designed to be read by search engines and computer programs



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



character indicating the number of protons ('N' means neutral)

InChIKev=RYYVLZVUVIJVGH-UHFFFAOYSA-N

flag character for InChI version: 'A' for version 1

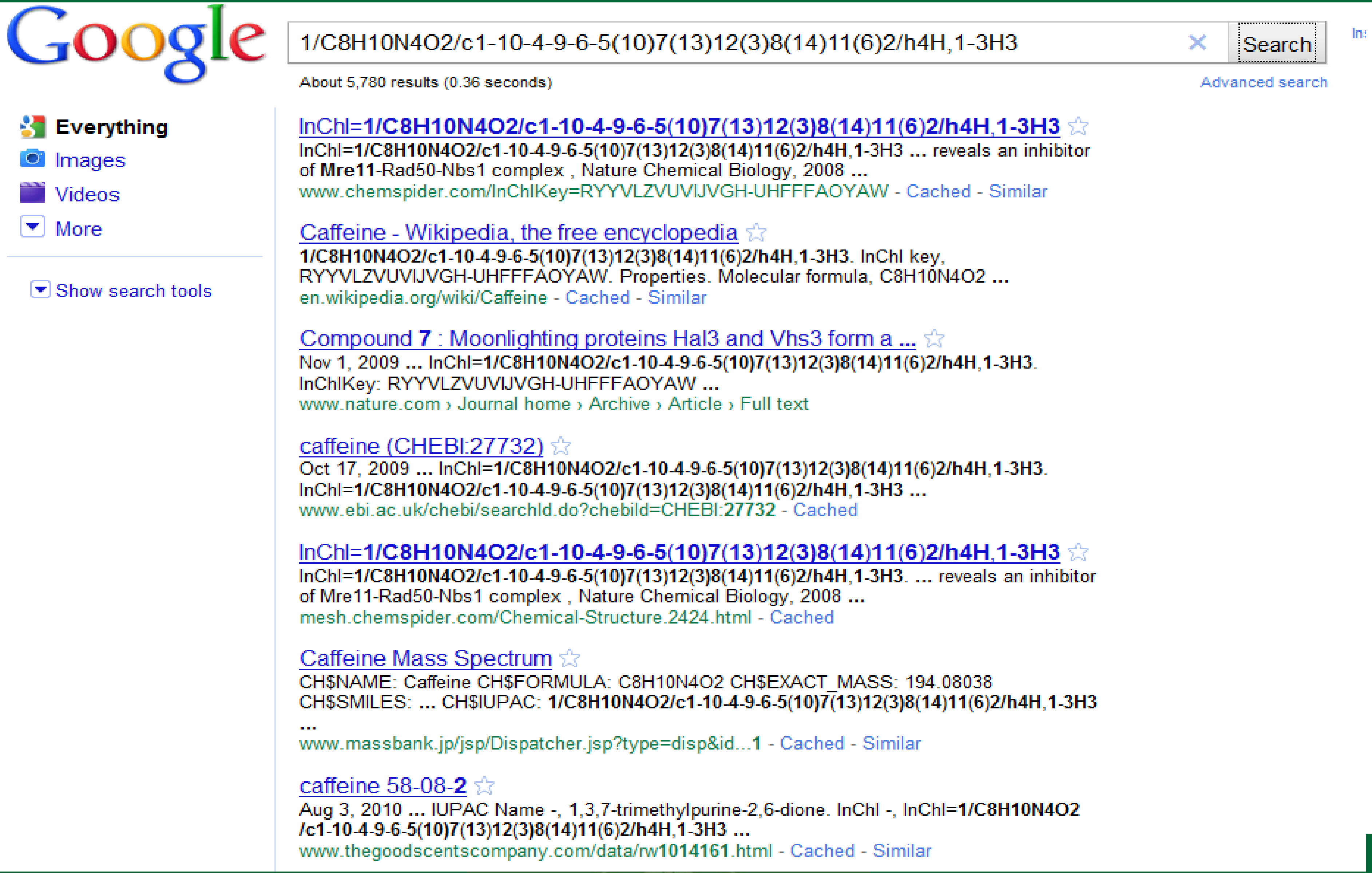
First block (14 letters)

Encodes molecular skeleton (connectivity)

Second block (8 letters)

Encodes stereochemistry and isotopes

flag character ('S') indicates standard InChIKey (produced out of standard InChI)



About 5,780 results (0.36 seconds)

InChI=1/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3
InChI=1/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3 ... reveals an inhibitor of Mre11-Rad50-Nbs1 complex , Nature Chemical Biology, 2008 ...
www.chemspider.com/InChIKey=RYYVLZVUVIJVGH-UHFFFAOYAW - Cached - Similar

Caffeine - Wikipedia, the free encyclopedia
1/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3. InChI key, RYYVLZVUVIJVGH-UHFFFAOYAW. Properties. Molecular formula, C8H10N4O2 ...
en.wikipedia.org/wiki/Caffeine - Cached - Similar

Compound 7 : Moonlighting proteins Hal3 and Vhs3 form a ...
Nov 1, 2009 ... InChI=1/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3. InChIKey: RYYVLZVUVIJVGH-UHFFFAOYAW ...
www.nature.com › Journal home › Archive › Article › Full text

caffeine (CHEBI:27732)
Oct 17, 2009 ... InChI=1/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3. InChI=1/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3 ...
www.ebi.ac.uk/chebi/searchId.do?chebiId=CHEBI:27732 - Cached

InChI=1/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3
InChI=1/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3. ... reveals an inhibitor of Mre11-Rad50-Nbs1 complex , Nature Chemical Biology, 2008 ...
mesh.chemspider.com/Chemical-Structure.2424.html - Cached

Caffeine Mass Spectrum
CH$NAME: Caffeine CH$FORMULA: C8H10N4O2 CH$EXACT_MASS: 194.08038 CH$SMILES: ... CH$IUPAC: 1/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3 ...
www.massbank.jp/jsp/Dispatcher.jsp?type=disp&id...1 - Cached - Similar

caffeine 58-08-2
Aug 3, 2010 ... IUPAC Name -, 1,3,7-trimethylpurine-2,6-dione. InChI -, InChI=1/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3 ...
www.thegoodscentscompany.com/data/rw1014161.html - Cached - Similar

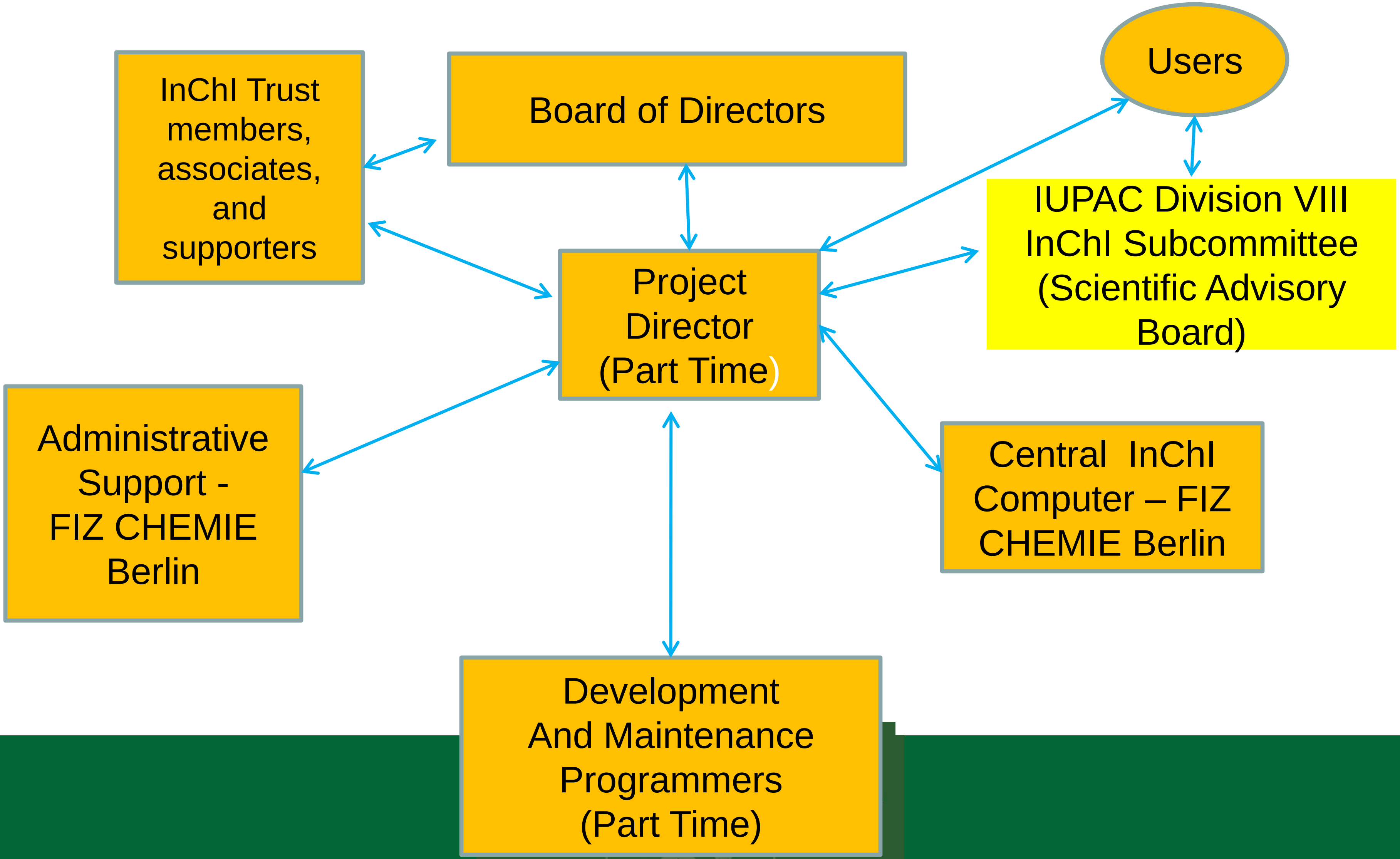
Scientific Articles Mentioning InChI

"The Chemical Translation Service (CTS) - a web-based tool to improve standardization of metabolomic reports"
Gert Wohlgemuth, Pradeep Kumar Haldiya, Egon Willighagen, Tobias Kind, and Oliver Fiehn
Bioinformatics, published 9 September 2010 (Open Access)

"PathwayAccess: CellDesigner plugins for pathway databases"
John L. Van Hemert and Julie A. Dickerson
Bioinformatics **2010**, *26(18)*, 2345-2346 (Open Access)

"Utopia documents: linking scholarly literature with research data"
T. K. Attwood, D. B. Kell, P. McDermott, J. Marsh, S. R. Pettifer, and D. Thorne
Bioinformatics **2010**, *26(18)*, 568-574 (Open Access)

InChI Trust Organization



InChI Trust members, associates, and supporters

Board of Directors

Users

IUPAC Division VIII InChI Subcommittee (Scientific Advisory Board)

Project Director (Part Time)

Administrative Support - FIZ CHEMIE Berlin

Central InChI Computer – FIZ CHEMIE Berlin

Development And Maintenance Programmers (Part Time)

Current InChI Trust Members

ACD Labs
ChemAxon
Elsevier
FIZ CHEMIE – Berlin
IBM
Informa/Taylor & Francis
IUPAC
John Wiley & Sons
Microsoft
Nature Publishing Group
OpenEye
ProQuest/Dialog
Royal Society of Chemistry (RSC)
Springer
Symyx
Thomson-Reuters

16 as of 11/1/2010

Current InChI Trust Supporters

CalTech, CA, USA
Chem21, IL, USA
Indiana University, IN, USA
National Chemical Laboratory, Pune, India
National Institute of Chemistry, Ljubljana, Slovenia
SharePoint, WA, USA
Trinity University, TX ,USA
Unilever Centre for Molecular Science Informatics, Cambridge UK
University of Applied Science, Gelsenkirchen , Germany
University of California – Riverside
University of California – San Francisco
University of North Carolina, NC, USA
University of the West Indies, Mona, Jamaica
Xemistry GmbH, Germany

14 as of 11/1/2010

Future development

There are working groups looking at InChI extensions for:

Markush	(results expected 2011)
Polymers/Mixtures	(results expected 2011)
InChI Resolver protocols	(results expected 2011)
Organometallics	(results expected 2012)
Electronic States	(results expected 2012)
RInChI –InChI for Reactions	(results expected 2012)

Possible Future Enhancements

1. **Transrutherfordium elements**
2. **Electronic States, including Transition states and Excited states.**
3. **Work with IUCr for 3D information**
4. **Proteins, Peptides & Biopolymers**
5. **Mac supported version**
6. **Java version**
7. **VS2010 .NET compilation support**

The Future

InChI has become mainstream for publishers, databases providers, and software developers. Over the next 5-10 years, publishers will use data mining to create both better abstracts, useful indexing, and concept terms. Search engines will be able to search for appropriate text and structures and direct users to the original (fee or free/Open Access/Open Data) sources.

Acknowledgements

(Primarily members for the IUPAC InChI subcommittee and associated InChI working groups)

Steve Bachrach, Colin Batchelor, John Barnard ,Evan Bolton, Steve Boyer, Steve Bryant, Szabolcs Csepregi ,Rene Deplanque, Nicko Goncharoff, Jonathan Goodman, Guenter Grethe, Richard Hartshorn, Jaroslav Kahovec , Richard Kidd, Hans Kraut, Alexander Lawson , Peter Linstrom, Bill Milne, Gerry Moss, Peter Murray-Rust, Heike Nau , Marc Nicklaus, Carmen Nitsche, Matthias Nolte , Igor Pletnev, Josep Prous, Hinnerk Rey, Ulrich Roessler, Roger Schenck , Martin Schmidt, Steve Stein, Peter Shepherd, Markus Sitzmann, Chris Steinbeck, Keith Taylor, Dmitrii Tchekhovskoi, Bill Town, Wendy Warr, Jason Wilde, Tony Williams, Andrey Yerin.

Special Acknowledgement: Ted Becker& Alan McNaught for their vision and leadership of the future of IUPAC nomenclature.