A Short History of the IUPAC InChI Algorithm

Stephen Heller

NIST & InChI Project Director

The main web sites for the IUPAC InChI project are:
http://www.iupac.org/inchi
and
http://www.inchi-trust.org **4/1/2019**

Slides are available at http://www.hellers.com/steve/orlando-4-19.pdf

This is a green talk –

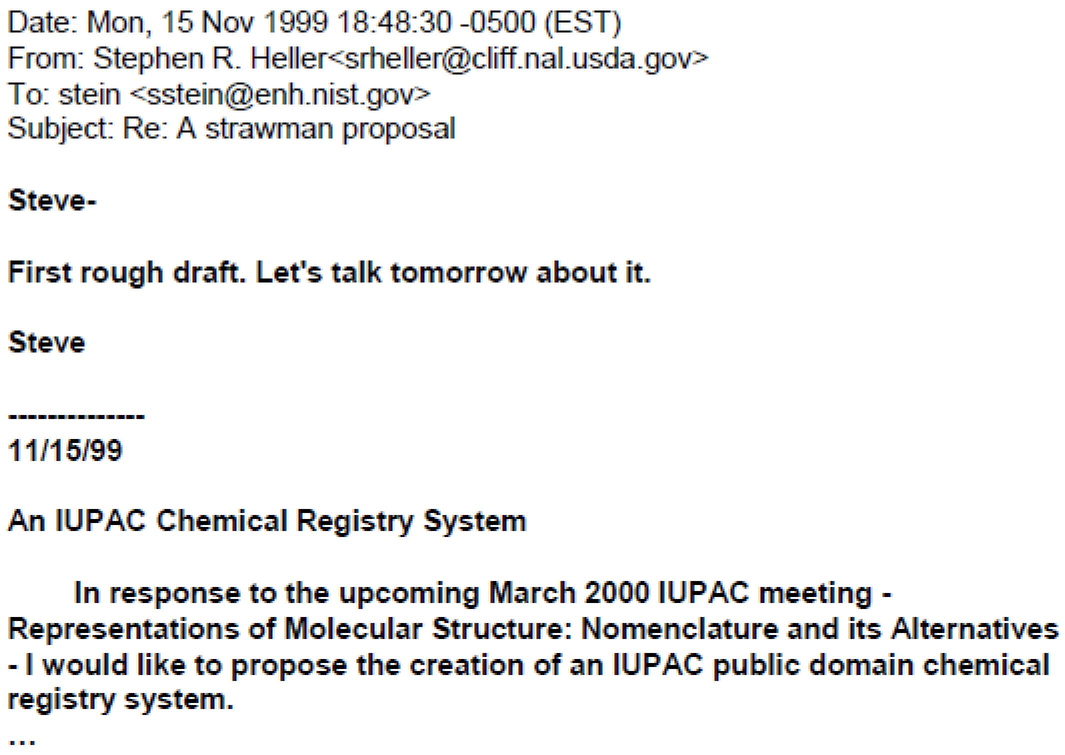
These slides were made from 100% recycled electrons

I will try to be politically correct.

I have even take a course in being PC.

But, I flunked it – twice.

My clone is giving this lecture. I am actually at Space Mountain ride in Disney World.



Date: Mon, 15 Nov 1999 18:48:30 -0500 (EST)
From: Stephen R. Heller<srheller@cliff.nal.usda.gov>
To: stein <sstein@enh.nist.gov>
Subject: Re: A strawman proposal

Steve-

First rough draft. Let's talk tomorrow about it.

Steve

11/15/99

An IUPAC Chemical Registry System

In response to the upcoming March 2000 IUPAC meeting - Representations of Molecular Structure: Nomenclature and its Alternatives - I would like to propose the creation of an IUPAC public domain chemical registry system.
…

Short History of InChI

1999: Steve’s Heller & Stein initiated a proposal at NIST for a public domain structure representation standard for the NIST databases
2000: Decided that InChI would be an IUPAC initiative
2001: The IUPAC Chemical Identifier project began
2005: version 1 was launched
2009: standard versions of InChI and the InChIKey were released, which took the original algorithm with its many variable parameters and fixed them so that interoperability between databases and resources with InChIs could be achieved
2009: the InChI Trust was formed
2011: version 1.04 released
2017: version 1.05 of the InChI, along with version 1.00 of Reaction InChI (RInChI)

How did we get here – really?

The Mass Spectral Search System, set up in the 70's at NIH, became part of the NIH/EPA Chemical Information Systems. I was involved in the earlier work and joined EPA. With EPA having a very practical need for mass spectral search and analysis, considerable funds were provided to increase the size and quality of the library. The initial library and software work was contracted and some quality control work was started, including adding CAS Registry Numbers

In the Beginning

By 1975 there were some 30,000 spectra in the library. While EPA was the driving force behind the rapid expansion and use of the library, it did not have the mission to provide data to the public. In 1978, the National Bureau of Standards – NBS (now the National Institute of Standards and Technology NIST) agreed to print a five volume collection of some 25,500 different mass spectra along with their chemical name, synonyms, and chemical structure. NBS/NIST agreed in 1980 to assume responsibility for the dissemination of the library.

In the End

Jumping forward a decade or two, in the late 1990’s CAS Registry Numbers could no longer be used, so the chemical structure of a compound became as the unique “key” for identifying the compound. For this purpose, chemical structure processing software was developed at NIST to enable compound “registration” (finding spectra for the same compound). This enabled the immediate inclusion of well over 10,000 compounds held in the archive for which the CAS Registry Numbers were unavailable. However this was not an ideal situation.

The IUPAC Awakening

Steve’s Heller & Stein came up with an outline of a plan in 1999 to develop a more rational chemical registration system for the database.

Meanwhile, with the ever increasing reliance on computer processing by chemists, it became evident to Ted Becker and Alan McNaught at IUPAC that this organization should explore new, computer-driven approaches to the problem of chemical identification

InChI Project Goal

To find & link everything about a chemical from many sources with the purpose of creating new information.

InChI Videos

1. What on Earth is InChI?

http://www.youtube.com/watch?v=rAnJ5toz26c

2. The Birth of the InChI

http://www.youtube.com/watch?v=X9c0PHXPfso

3. The Googlable InChIKey

http://www.youtube.com/watch?v=UxSNOtv8Rjw

4. InChI and the Islands

http://www.youtube.com/watch?v=qrCqJ0o4jGs

What is InChI?

The IUPAC International Chemical Identifier, or InChI, is a non-proprietary, machine-readable string of symbols which enables a computer to represent the compound in a completely unequivocal manner.

InChIs are produced by computer from structures drawn on-screen with existing structure drawing software, and the original structure can be regenerated from an InChI with existing structure drawing software.

InChI is really just a synonym.

http://en.wikipedia.org/wiki/International_Chemical_Identifier

Unique InChI Features

Only IUPAC International structure standard

Only Open Source structure standard

Only structure standard support by a wide majority of publishers, database producers, and chemistry software companies

Four Requirements for a Computer Representation Standard

Need

Definition/Specification

Timing/Infrastructure

Acceptance/Use

Why InChI? - Too Many Good and Excellent Identifiers ("Standards")

Structure diagrams
- various conventions
- contain 'too much' information

Connection Tables/Notations
- MolFiles, SDF, SMILES, SLN, ROSDAL, …

Pronounceable names (and mostly unpronounceable) and mostly complex names
- IUPAC, CAS 8th CI name, CAS 9th CI name, trivial, trade, WHO INN, ASK, ISO

(Dumb) Index Numbers
EINECS, ELINCS, FEMA, DOT, RTECS, CAS, Beilstein, USP, RTECS, EEC, RCRA, NCI, UN, USAN, EC, ChemSpider ID, REACH, PubChem CID, BAN, NSC, ASK, KEGG, BP, IND, MARTINDALE, MESH, IT IS, RX-CUI, NDF-RT, ATC, AHPA, USP/NF, UNII, MFCD#, and so on

“Standards are like toothbrushes – everyone has one but no one wants to use someone else's.”

Phil Bourne,
Former Associate Director for Data Science (Big Data), NIH

Definition/Specification

An arbitrary computer algorithm to ensure consistency and reproducibility and to be able to call it a real standard.

There really is no written standard
Software is the implementation

What “*is*” the InChI standard?

The InChI standard programmed into the algorithm is an arbitrary decision as to how structures are handled. In most cases there is total agreement (e.g., methane). In cases of more complex molecules where there is not agreement among chemists, one representation is chosen. As long as the arbitrarily chosen representation is properly programmed, one will always get the SAME result using it – which is what a standard is!

InChI Characteristics

1. Easy to generate

2. Expressive (it will contain structural information)

3. Unambiguous/Unique

4. Does not require a centralized operation (it can be generated anywhere – can use crowdsourcing/free labor)

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

InChI is for computers

An InChI string is not directly intelligible to the normal human reader. Like Bar Codes, and InChI QR codes - InChIs are not designed to be read by humans.

Or, put another way – never send a human to do a machine's job!

Technology is at its best when it is invisible.

How difficult is it to create an InChI?

Today, all the major structure drawing programs (ChemDraw, MDL/Symyx /Accelrys/BIOVIA Draw, ISIS Draw, ChemAxon Marvin Sketch, ACD Labs ChemSketch, CLiDE, Jmol, and so on) have incorporated the InChI algorithm in their products, with usually an “InChI” button for generating the InChI.

InChI is the worst computer readable structure representation except for all those other forms that have been tried from time to time.

With apologies to Sir Winston Churchill (House of Commons speech on November 11, 1947)

Plank’s Law

"New scientific truth does not triumph by convincing its opponents and making them see the light, but rather because its opponents eventually die, and a new generation grows up that is familiar with it.“

Max Planck,
"Scientific Autobiography and Other Papers",
Williams & Norgate,
London (1950), pages 33-34.

Timing &Infrastructure

InChI has become a standard only because of the world has changed in the last 20 years and the old order is dead or dying.

Without the Internet, without vast amounts of data and information becoming available in computer readable form for the first time, without Google (and other search engines), without structure drawing programs, and with most chemistry publishers now needing chemical structures in their products, InChI would be yet another interesting graph theory project that died like so many before it.

Without this perfect good storm that created a foundation for InChI, at best, I would be talking to a group a 5-7 people at an IUPAC meeting talk.

Three strategic pillars for success

Global adoption and use
Increasing engagement with the chemistry community for the benefit of science and business

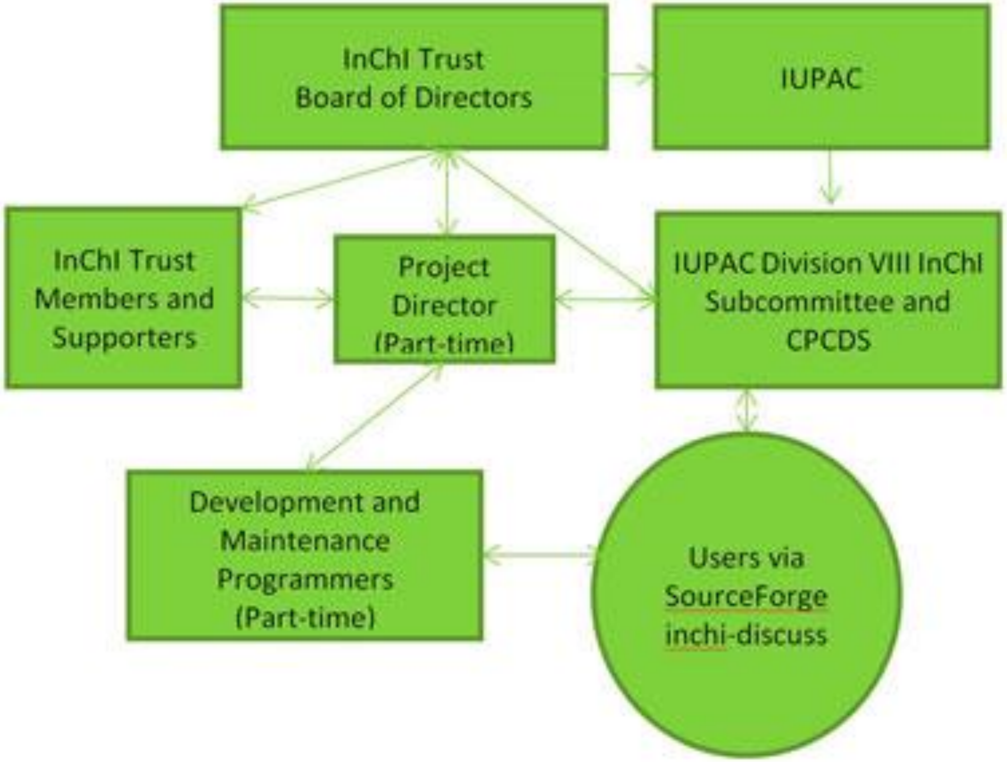
Maintenance & extension of the InChI and applications
To facilitate rapid and effective research discovery and business innovation

Governance
To provide an organizational framework that ensures the sustainability of the standard

The InChI Trust

To function and succeed, InChI had to become personality independent. InChI had to be “institutionalized”. If the work of this project was to be enduring it needed to turned over to an entity that would ensure its ongoing activities and be acceptable to the community. It was concluded that a not-for-profit organization would best fit the ongoing and future project needs. Thus the decision to create and incorporate the "InChI Trust“ as a UK charity.

Governance


InChI Trust
Board of Directors
IUPAC
InChI Trust
Members and
Supporters
Project
Director
(Part-time)
IUPAC Division VIII InChI
Subcommittee and
CPCDS
Development and
Maintenance
Programmers
(Part-time)
Users via
SourceForge
inchi-discuss

ACD/Labs

ALDRICH Chemistry

BIO-RAD

The Cambridge Crystallographic Data Centre

ChemAxon

CAS A DIVISION OF THE AMERICAN CHEMICAL SOCIETY

IUPAC

FDA

NIH National Institutes of Health

NIST National Institute of Standards and Technology U.S. Department of Commerce

OpenEye

RELX Group

ROYAL SOCIETY OF CHEMISTRY

SPRINGER NATURE

Taylor & Francis Group an informa business

WILEY Publishers Since 1807

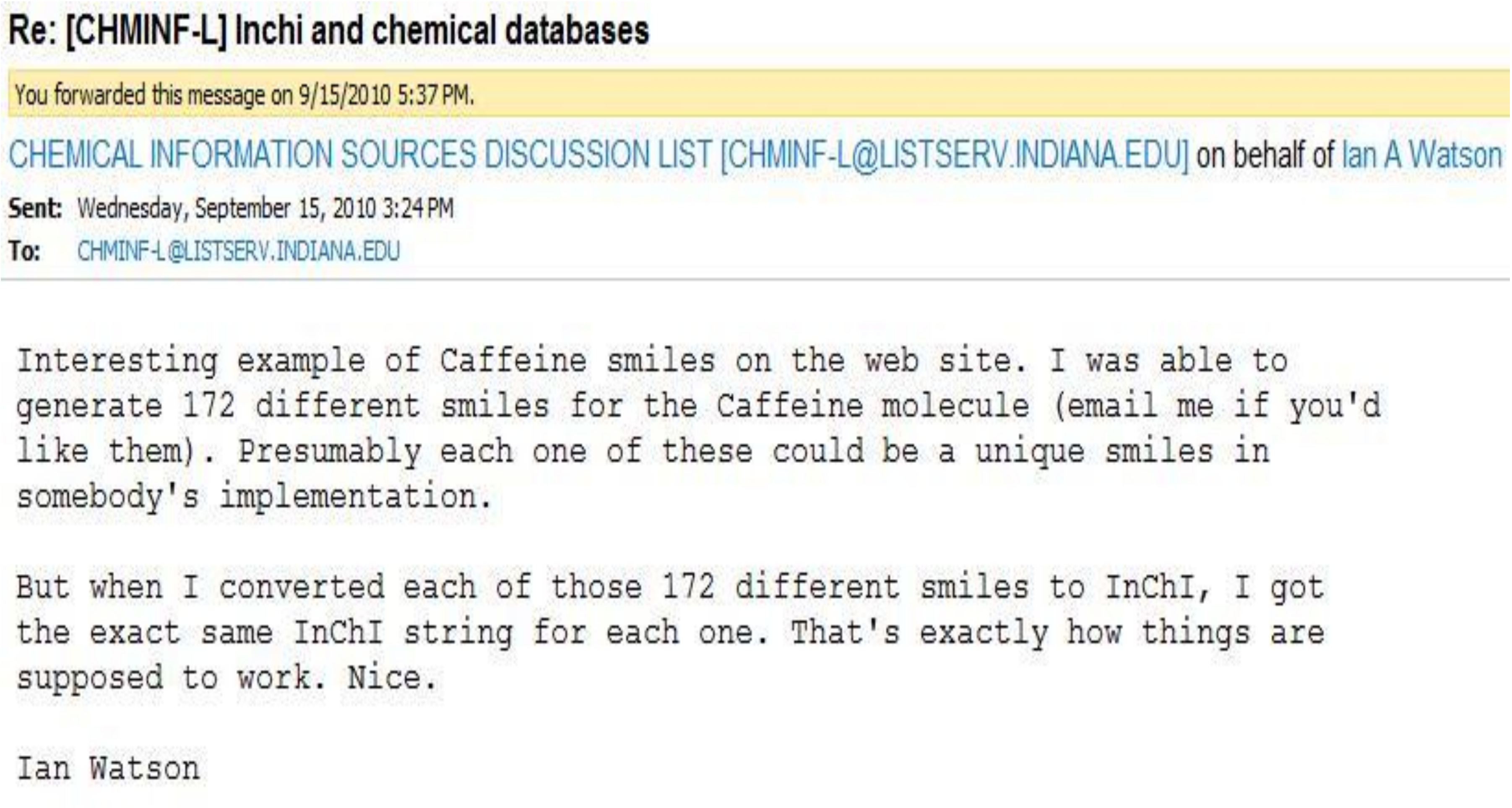
Current InChI Status

At present, practically speaking, InChI can handle simple organic molecules, which turns out to cover 99%+ of what people deal with every day. If it did not the every day needs of chemists and information specialists then the usage of InChI would not be as great as it is.

But InChI will never handle undefinable chemicals:

regular gas/summer or winter
balsamic vinegar
vegetable oil
partially hydrogenated oil
low sodium soy sauce

An example of Life before InChI – How to represent caffeine



Re: [CHMINF-L] Inchi and chemical databases

You forwarded this message on 9/15/2010 5:37 PM.

CHEMICAL INFORMATION SOURCES DISCUSSION LIST [CHMINF-L@LISTSERV.INDIANA.EDU] on behalf of Ian A Watson

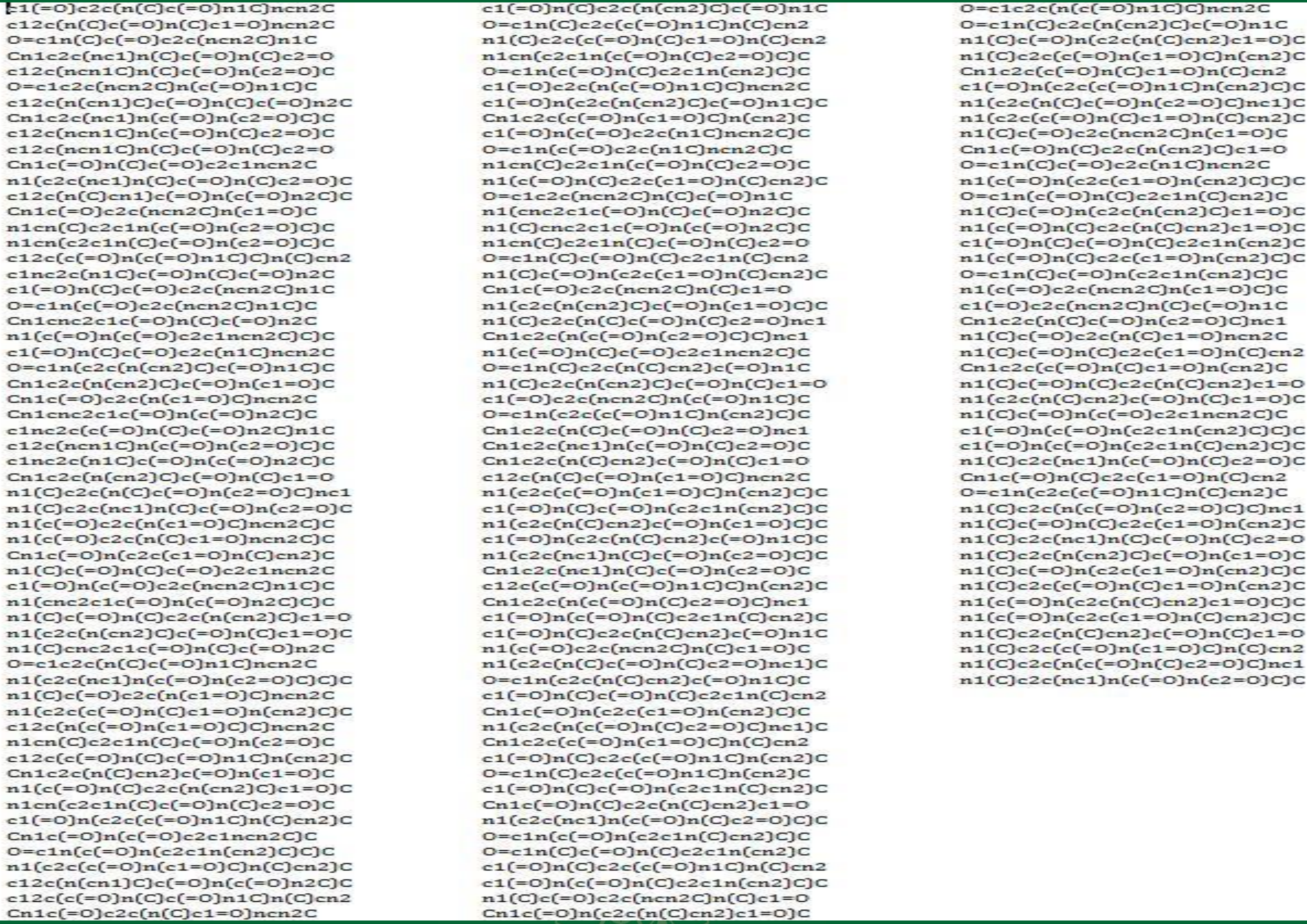
Sent: Wednesday, September 15, 2010 3:24 PM

To: CHMINF-L@LISTSERV.INDIANA.EDU

Interesting example of Caffeine smiles on the web site. I was able to generate 172 different smiles for the Caffeine molecule (email me if you'd like them). Presumably each one of these could be a unique smiles in somebody's implementation.

But when I converted each of those 172 different smiles to InChI, I got the exact same InChI string for each one. That's exactly how things are supposed to work. Nice.

Ian Watson



```
c1(=O)c2c(n(C)c(=O)n1C)ncn2C
c12c(n(C)c(=O)n(C)c1=O)ncn2C
O=c1n(C)c(=O)c2c(ncn2C)n1C
Cn1c2c(nc1)n(C)c(=O)n(C)c2=O
c12c(ncn1C)n(C)c(=O)n(c2=O)C
O=c1c2c(ncn2C)n(c(=O)n1C)C
c12c(n(cn1)C)c(=O)n(C)c(=O)n2C
Cn1c2c(nc1)n(c(=O)n(c2=O)C)C
c12c(ncn1C)n(c(=O)n(C)c2=O)C
c12c(ncn1C)n(C)c(=O)n(C)c2=O
Cn1c(=O)n(C)c(=O)c2c1ncn2C
n1(c2c(nc1)n(C)c(=O)n(C)c2=O)C
c12c(n(C)cn1)c(=O)n(c(=O)n2C)C
Cn1c(=O)c2c(ncn2C)n(c1=O)C
n1cn(C)c2c1n(c(=O)n(c2=O)C)C
n1cn(c2c1n(C)c(=O)n(c2=O)C)C
c12c(c(=O)n(c(=O)n1C)C)n(C)cn2
c1nc2c(n1C)c(=O)n(C)c(=O)n2C
c1(=O)n(C)c(=O)c2c(ncn2C)n1C
O=c1n(c(=O)c2c(ncn2C)n1C)C
Cn1cnc2c1c(=O)n(C)c(=O)n2C
n1(c(=O)n(c(=O)c2c1ncn2C)C)C
c1(=O)n(C)c(=O)c2c(n1C)ncn2C
O=c1n(c2c(n(cn2)C)c(=O)n1C)C
Cn1c2c(n(cn2)C)c(=O)n(c1=O)C
Cn1c(=O)c2c(n(c1=O)C)ncn2C
Cn1cnc2c1c(=O)n(c(=O)n2C)C
c1nc2c(c(=O)n(C)c(=O)n2C)n1C
c12c(ncn1C)n(c(=O)n(c2=O)C)C
c1nc2c(n1C)c(=O)n(c(=O)n2C)C
Cn1c2c(n(cn2)C)c(=O)n(C)c1=O
n1(C)c2c(n(C)c(=O)n(c2=O)C)nc1
n1(C)c2c(nc1)n(C)c(=O)n(c2=O)C
n1(c(=O)c2c(n(c1=O)C)ncn2C)C
n1(c(=O)c2c(n(C)c1=O)ncn2C)C
Cn1c(=O)n(c2c(c1=O)n(C)cn2)C
n1(C)c(=O)n(C)c(=O)c2c1ncn2C
c1(=O)n(c(=O)c2c(ncn2C)n1C)C
n1(cnc2c1c(=O)n(c(=O)n2C)C)C
n1(C)c(=O)n(C)c2c(n(cn2)C)c1=O
n1(c2c(n(cn2)C)c(=O)n(C)c1=O)C
n1(C)cnc2c1c(=O)n(C)c(=O)n2C
O=c1c2c(n(C)c(=O)n1C)ncn2C
n1(c2c(nc1)n(c(=O)n(c2=O)C)C)C
n1(C)c(=O)c2c(n(c1=O)C)ncn2C
n1(c2c(c(=O)n(C)c1=O)n(cn2)C)C
c12c(n(c(=O)n(c1=O)C)C)ncn2C
n1cn(C)c2c1n(C)c(=O)n(c2=O)C
c12c(c(=O)n(C)c(=O)n1C)n(cn2)C
Cn1c2c(n(C)cn2)c(=O)n(c1=O)C
n1(c(=O)n(C)c2c(n(cn2)C)c1=O)C
n1cn(c2c1n(C)c(=O)n(C)c2=O)C
c1(=O)n(c2c(c(=O)n1C)n(C)cn2)C
Cn1c(=O)n(c(=O)c2c1ncn2C)C
O=c1n(c(=O)n(c2c1n(cn2)C)C)C
n1(c2c(c(=O)n(c1=O)C)n(C)cn2)C
c12c(n(cn1)C)c(=O)n(c(=O)n2C)C
c12c(c(=O)n(C)c(=O)n1C)n(C)cn2
Cn1c(=O)c2c(n(C)c1=O)ncn2C
```

```
c1(=O)n(C)c2c(n(cn2)C)c(=O)n1C
O=c1n(C)c2c(c(=O)n1C)n(C)cn2
n1(C)c2c(c(=O)n(C)c1=O)n(C)cn2
n1cn(c2c1n(c(=O)n(C)c2=O)C)C
O=c1n(c(=O)n(C)c2c1n(cn2)C)C
c1(=O)c2c(n(c(=O)n1C)C)ncn2C
c1(=O)n(c2c(n(cn2)C)c(=O)n1C)C
Cn1c2c(c(=O)n(c1=O)C)n(cn2)C
c1(=O)n(c(=O)c2c(n1C)ncn2C)C
O=c1n(c(=O)c2c(n1C)ncn2C)C
n1cn(C)c2c1n(c(=O)n(C)c2=O)C
n1(c(=O)n(C)c2c(c1=O)n(C)cn2)C
O=c1c2c(ncn2C)n(C)c(=O)n1C
n1(cnc2c1c(=O)n(C)c(=O)n2C)C
n1(C)cnc2c1c(=O)n(c(=O)n2C)C
n1cn(C)c2c1n(C)c(=O)n(C)c2=O
O=c1n(C)c(=O)n(C)c2c1n(C)cn2
n1(C)c(=O)n(c2c(c1=O)n(C)cn2)C
Cn1c(=O)c2c(ncn2C)n(C)c1=O
n1(c2c(n(cn2)C)c(=O)n(c1=O)C)C
n1(C)c2c(n(C)c(=O)n(C)c2=O)nc1
Cn1c2c(n(c(=O)n(c2=O)C)C)nc1
n1(c(=O)n(C)c(=O)c2c1ncn2C)C
O=c1n(C)c2c(n(C)cn2)c(=O)n1C
n1(C)c2c(n(cn2)C)c(=O)n(C)c1=O
c1(=O)c2c(ncn2C)n(c(=O)n1C)C
O=c1n(c2c(c(=O)n1C)n(cn2)C)C
Cn1c2c(n(C)c(=O)n(C)c2=O)nc1
Cn1c2c(nc1)n(c(=O)n(C)c2=O)C
Cn1c2c(n(C)cn2)c(=O)n(C)c1=O
c12c(n(C)c(=O)n(c1=O)C)ncn2C
n1(c2c(c(=O)n(c1=O)C)n(cn2)C)C
c1(=O)n(C)c(=O)n(c2c1n(cn2)C)C
n1(c2c(n(C)cn2)c(=O)n(c1=O)C)C
c1(=O)n(c2c(n(C)cn2)c(=O)n1C)C
n1(c2c(nc1)n(C)c(=O)n(c2=O)C)C
Cn1c2c(nc1)n(C)c(=O)n(c2=O)C
c12c(c(=O)n(c(=O)n1C)C)n(cn2)C
Cn1c2c(n(c(=O)n(C)c2=O)C)nc1
c1(=O)n(c(=O)n(C)c2c1n(C)cn2)C
c1(=O)n(C)c2c(n(C)cn2)c(=O)n1C
n1(c(=O)c2c(ncn2C)n(C)c1=O)C
n1(c2c(n(C)c(=O)n(C)c2=O)nc1)C
O=c1n(c2c(n(C)cn2)c(=O)n1C)C
c1(=O)n(C)c(=O)n(C)c2c1n(C)cn2
Cn1c(=O)n(c2c(c1=O)n(cn2)C)C
n1(c2c(n(c(=O)n(C)c2=O)C)nc1)C
Cn1c2c(c(=O)n(c1=O)C)n(C)cn2
c1(=O)n(C)c2c(c(=O)n1C)n(cn2)C
O=c1n(C)c2c(c(=O)n1C)n(cn2)C
c1(=O)n(C)c(=O)n(c2c1n(C)cn2)C
Cn1c(=O)n(C)c2c(n(C)cn2)c1=O
n1(c2c(nc1)n(c(=O)n(C)c2=O)C)C
O=c1n(c(=O)n(c2c1n(C)cn2)C)C
O=c1n(C)c(=O)n(C)c2c1n(cn2)C
c1(=O)n(C)c2c(c(=O)n1C)n(C)cn2
c1(=O)n(c(=O)n(C)c2c1n(cn2)C)C
n1(C)c(=O)c2c(ncn2C)n(C)c1=O
Cn1c(=O)n(c2c(n(C)cn2)c1=O)C
```

```
O=c1c2c(n(c(=O)n1C)C)ncn2C
O=c1n(C)c2c(n(cn2)C)c(=O)n1C
n1(C)c(=O)n(c2c(n(C)cn2)c1=O)C
n1(C)c2c(c(=O)n(c1=O)C)n(cn2)C
Cn1c2c(c(=O)n(C)c1=O)n(C)cn2
c1(=O)n(c2c(c(=O)n1C)n(cn2)C)C
n1(c2c(n(C)c(=O)n(c2=O)C)nc1)C
n1(c2c(c(=O)n(C)c1=O)n(C)cn2)C
n1(C)c(=O)c2c(ncn2C)n(c1=O)C
Cn1c(=O)n(C)c2c(n(cn2)C)c1=O
O=c1n(C)c(=O)c2c(n1C)ncn2C
n1(c(=O)n(c2c(c1=O)n(cn2)C)C)C
O=c1n(c(=O)n(C)c2c1n(C)cn2)C
n1(C)c(=O)n(c2c(n(cn2)C)c1=O)C
n1(c(=O)n(C)c2c(n(C)cn2)c1=O)C
c1(=O)n(C)c(=O)n(C)c2c1n(cn2)C
n1(c(=O)n(C)c2c(c1=O)n(cn2)C)C
O=c1n(C)c(=O)n(c2c1n(cn2)C)C
n1(c(=O)c2c(ncn2C)n(c1=O)C)C
c1(=O)c2c(ncn2C)n(C)c(=O)n1C
Cn1c2c(n(C)c(=O)n(c2=O)C)nc1
n1(C)c(=O)c2c(n(C)c1=O)ncn2C
n1(C)c(=O)n(C)c2c(c1=O)n(C)cn2
Cn1c2c(c(=O)n(C)c1=O)n(cn2)C
n1(C)c(=O)n(C)c2c(n(C)cn2)c1=O
n1(c2c(n(C)cn2)c(=O)n(C)c1=O)C
n1(C)c(=O)n(c(=O)c2c1ncn2C)C
c1(=O)n(c(=O)n(c2c1n(cn2)C)C)C
c1(=O)n(c(=O)n(c2c1n(C)cn2)C)C
n1(C)c2c(nc1)n(c(=O)n(C)c2=O)C
Cn1c(=O)n(C)c2c(c1=O)n(C)cn2
O=c1n(c2c(c(=O)n1C)n(C)cn2)C
n1(C)c2c(n(c(=O)n(c2=O)C)C)nc1
n1(C)c(=O)n(C)c2c(c1=O)n(cn2)C
n1(C)c2c(nc1)n(C)c(=O)n(C)c2=O
n1(C)c2c(n(cn2)C)c(=O)n(c1=O)C
n1(C)c(=O)n(c2c(c1=O)n(cn2)C)C
n1(C)c2c(c(=O)n(C)c1=O)n(cn2)C
n1(c(=O)n(c2c(n(C)cn2)c1=O)C)C
n1(c(=O)n(c2c(c1=O)n(C)cn2)C)C
n1(C)c2c(n(C)cn2)c(=O)n(C)c1=O
n1(C)c2c(c(=O)n(c1=O)C)n(C)cn2
n1(C)c2c(n(c(=O)n(C)c2=O)C)nc1
n1(C)c2c(nc1)n(c(=O)n(c2=O)C)C
```

Large Databases with InChIs/InChIKeys

EBI UniChem –157 million
NIH/NCI – 110 million
NIH/PubChem - 97 million
RSC/ChemSpider – 67 million
Elsevier/Reaxys – 31 million
IUPAC – 0 million

Why is InChI a Success

InChI is able to put things together in a new way. We took IUPAC, the Internet, Open Source software, crowdsourcing (SourceForge), Graph theory, existing representation algorithms, digitized data available on the web, and search engines, combines them, and created a very valuable tool.

InChI only works because of new technology. Without these factors above, for all practical purposes, no one would even know InChI existed.

Success is uncoerced adoption

InChI is not a replacement for any existing internal structure representations. InChI is in ADDITION to what one uses internally. Its value to chemists is in FINDING and LINKING information

InChI Staff and Collaborators

The InChI project has had the unusual perfect “good storm” of cooperation and support. It is a truly international project with programming in Moscow, computers in the cloud, incorporated in the UK, and a project director in the USA. Collaborators from over a dozen countries, from academia, Pharma, publishers, and the chemical information industry, have all offered, and continue to offer, senior scientific staff to develop the InChI standard.

Project Director

The project Director oversees all aspects of the project. The volunteer IUPAC InChI subcommittee working groups defining the standards, the programming of these standards, lecturing on InChI, organizing meetings and workings on InChI.

But being the Project Director for InChI is like running a cemetery; You have a lot a people under you but nobody listens to you .

As to how many people are participating on these working groups, I would say about 1/3

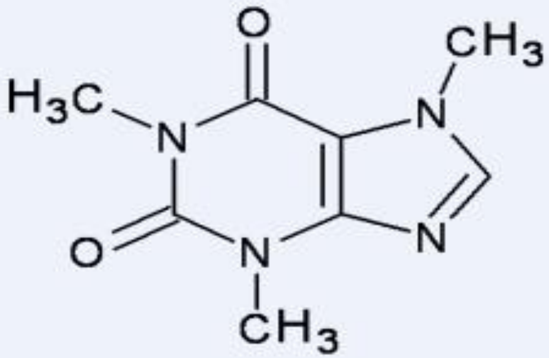
InChI layered structure design

The current InChI layers are:

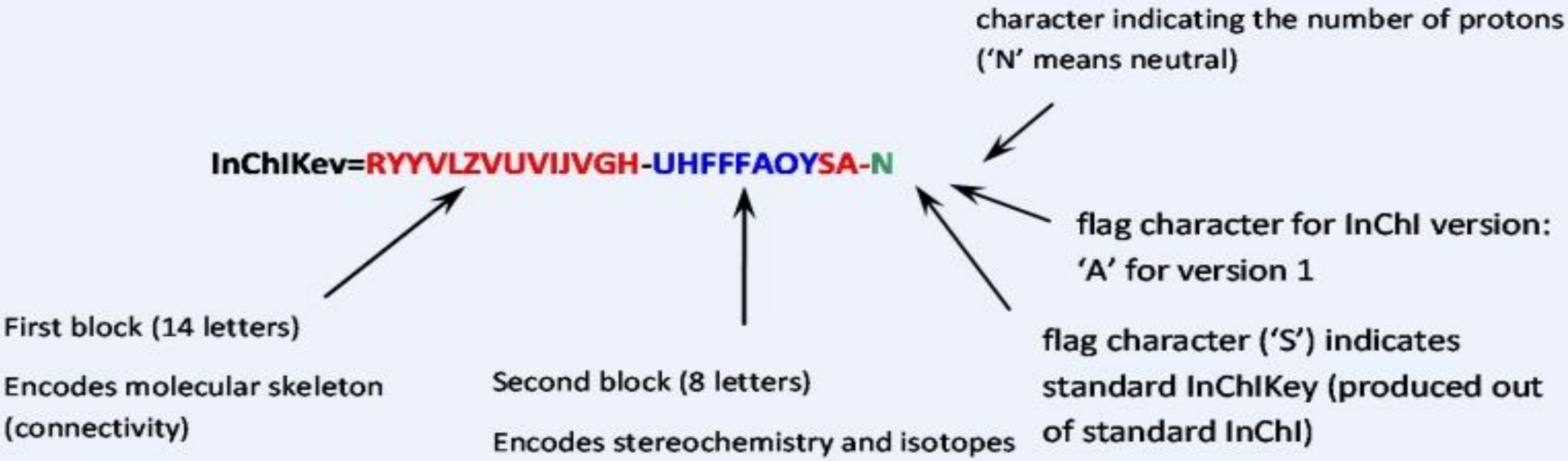
1. Formula
2. Connectivity (no formal bond orders)
   a. disconnected metals
   b. connected metals
3. Isotopes
4. Stereochemistry
   a. double bond (*Z/E*)
   b. tetrahedral (sp3)
5. Tautomers (on or off)

Charges are added to end of the string

The InChI Algorithm normalizes chemical representation and includes a “standardized” InChI, and the ‘hashed’ form called the InChIKey



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)

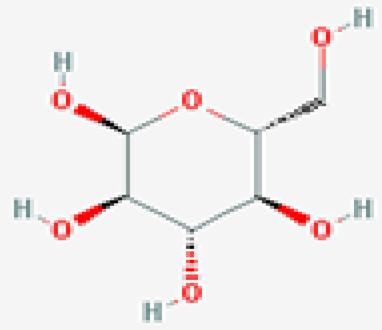
Second block (8 letters)

Encodes stereochemistry and isotopes

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

InChI is a string

InChI=1S/C6H12O6/c7-1-2-3(8)4(9)5(10)6(11)12-2/h2-11H,1H2/t2-,3-,4+,5-,6+/m1/s1



alpha-D-Glucose

Version/Type
Chemical formula
Connectivity
Charge/Proton
Stereochemical
Other (e.g., Isotopic)

“layered” line notation

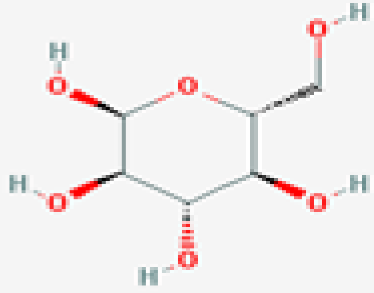
InChIKey is a “hashed” InChI

- Search engine friendly InChI
- May allow for ‘secure’ lookup of a chemical

WQZGKKKJIJFFOK-DVKNGEFBSA-N

Chemical formula
Connectivity
Stereochemical
Other (e.g., Isotopic)
Type
Version
Charge/Proton

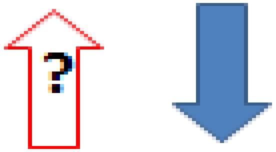
“layered” line notation



alpha-D-Glucose

InChIKey can be a ‘secret’

InChI=1S/C6H12O6/c7-1-2-3(8)4(9)5(10)6(11)12-2/h2-11H,1H2/t2-,3-,4+,5-,6+/m1/s1



WQZGKKKJIJFFOK-DVKNGEFBSA-N

There is no chemical information in an InChIKey … if you do not know the InChI, you cannot convert the InChIKey back into a chemical structure

Slide from Evan Bolton/NIH/PubChem

InChI/InChIKey Use and Utility

- InChI
  - Enabler of data exchange
  - Provides chemical structure normalization
- InChIKey
  - Compact form for structure lookup
  - Allows “secret” chemical information exchange

InChI characteristics

Consensus
Technical competence
Political and technical cooperation
Precompetitive collaboration – publishers, databases, software
No competition with commercial products
No mission creep
IUPAC blessing/endorsement & rapid IUPAC acceptance
Excellent understanding of what the Internet and how it can be effectively used in Chemical Information

Vision of the future

Current and Future Activities of InChI

Semi-Annual InChI Workshops

Mar 2017 – EBI Hinxton
Aug 2017 – NIH Bethesda

Aug 2018 – Boston MA

Feb 2019 – Cambridge UK
Aug 2019 – San Diego CA

Future Extensions

Organometallics (RFP just released - 3/19)
More Complete Tautomerism
Extended Stereochemistry
Inorganics
Positional isomers
Large molecules
Markush

Current and Future Applications

Reactions

Mixtures

InChI Resolver

QR codes for InChI

InChI open educational resources

InChI and the FAIR Project

Another use or application of InChI is with FAIR (Findable, Accessible, Interoperable, and Reusable) data principles project. InChI helps find information and data. InChI helps make the information and data accessible. And lastly InChI helps make the information and data interoperable.

Keep Calm
and Use InChI

Summary

If you are not part of the solution; you are part of the precipitate

Acknowledgements

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

Steve Bachrach, Colin Batchelor, John Barnard, Bob Belford, Evan Bolton, Ray Boucher, Steve Boyer, Ian Bruno, Steve Bryant, Alex Clark, Szabolcs Csepregi, Rene Deplanque, Josef Eiblmaier, Vincent Scalfani, Jeremy Frey, Nicko Goncharoff, Jonathan Goodman, Guenter Grethe, Richard Hartshorn, Jaroslav Kahovec , Richard Kidd, Hans Kraut, Alexander Lawson , Peter Linstrom, Gary Mallard, Leah McEwen, Bill Milne, Hunter Moseley, Moss, Peter Murray-Rust, Heike Nau, Marc Nicklaus, Carmen Nitsche, Matthias Nolte, Steffen Pauly, Igor Pletnev, Josep Prous, Peter Murray-Rust, 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.