# Standardizing metabolomic MS databases

*an open-source tool to cross-reference chemicals by structure codes,compound names and database identifiers*

**Gert Wohlgemuth, Pradeep Haldiya, Oliver Fiehn**

## Objective

Metabolomics needs standardized databases to compare data across studies, laboratories and different mass spectrometry platforms. The optimal way to annotate small molecules (0-1500 u) is by the corresponding structure in machine-readable formats. Open-source structure codes have been introduced by the IUPAC five years ago (InChI codes) but have not yet been widely adopted in literature. One problem is that there are no easy translator tools to be used by analytical chemists who lack time or expertise in computational chemistry to transform lists of identified metabolites into structure codes or to batch convert metabolite database identifiers into unambiguous and public InChI keys.

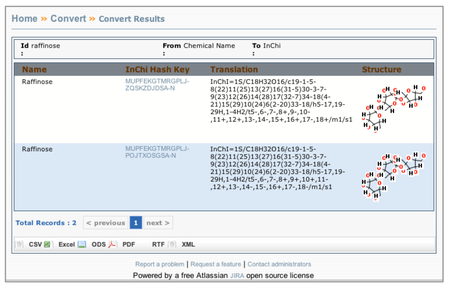
## References

The Chemical Translation Service (CTS) – a web-based tool
to improve standardization of metabolomic reports - Bioinformatics - submitted

http://cts.fiehnlab.ucdavis.edu

## Chemical names are poor identifiers

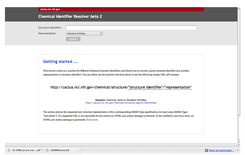
Is 'raffinose' a clear identifier? A synonym query for raffinose returns 10 hits in PubChem, 3 of which have 'raffinose' as major MeSH key identifier which are distinct in stereochemistry or structural modifications. This ambiguity is unavoidable when using names or synonyms. Accordingly, our Chemical Translation Service yields two structures.




## Similar tools are less powerful

**Chemical Identity Resolver**
http://cactus.nci.nih.gov/chemical/structure

easy to use
web based
one at a time conversion
batch convert not available
accessible for programmers




**The MetMask Project**
http://metmask.sourceforge.net/

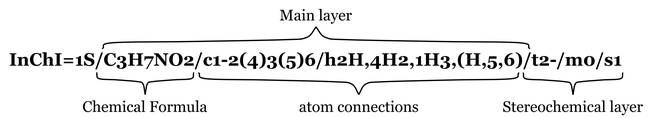
web based interface
R based interface
command line interface
hard to install
web interface does not allow batch converts


## What is a Std InChI Code

The Std InChI (IUPAC International Chemical Identifier) is a string of characters which are uniquely representing a chemical substance. It is derived from a structural representation of that substance in a specialist way to ensure that the a compound will always have the same Std InChI Code independent of the used tool to generate the structure. The InChI replaces outdated SMILES or SMART codes.

The Std InChI Code itself is separated into different layers, of which each presents a different part of the structure. This allows chemists to provide additional informations in a very easy and flexible way and so enhances the details.

The graphic below describes the layers in a simplified way




## What is a InChI Hash Key

The standard InChI Hash Key was developed, because the full standard InChI code can be over 4,000 letters long. Such lengths pose problems for DB indexing, but are even more problematic for chemists to publish structures e.g. in reports, or for use in Google queries.

An example for this would be Raffinose

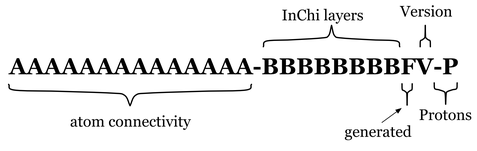
(InChI=1S/C18H32O16/c19-1-5-8(22)11(25)13(27)16(31-5)30-3-7-9(23)12(26)14(28)17(32-7)34-18(4-21)15(29)10(24)6(2-20)33-18/h5-17,19-29H,1-4H2/t5-,6-,7-,8+,9-,10-,11+,12+,13-,14-,15+,16+,17-,18+/m1/s1)

This long InChI code is translated to the Std InChI Key as

MUPFEKGTMRGPLJ-ZQSKZDJDSA-N

Hence, InChI Keys enable efficient queries for structures. The Key was defined as a hashcode of 27 characters, split into 5 blocks.

The first block encodes the achiral structure, i.e. the atom connectivity. All stereoisomers e.g. of raffinose therefore an identical first block and can be sorted and searched accordingly. The second block encodes the stereochemical information. The third block defines how it was generated, the 4th defines the version and the fifth indicates the charge state.

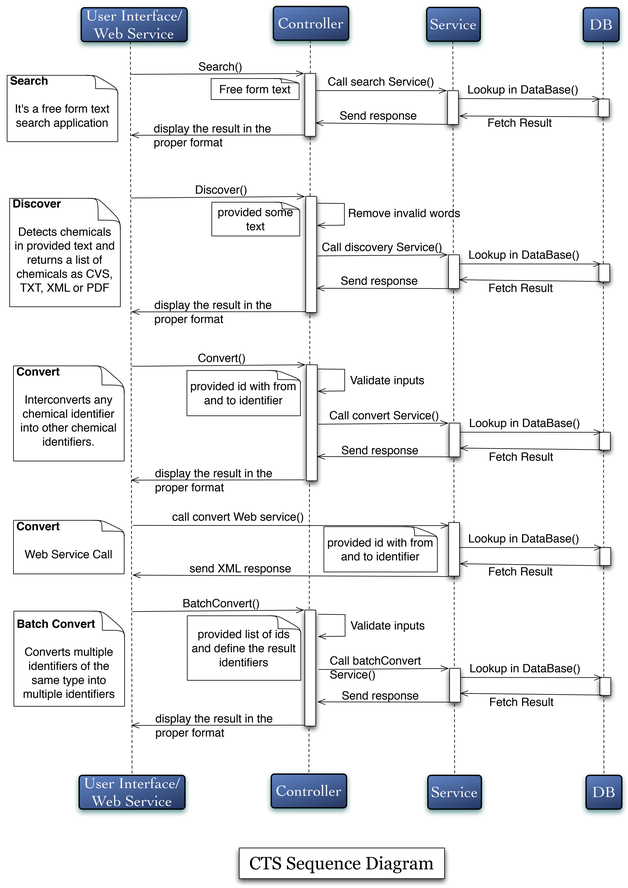



## Applications

Any list of identifiers can be converted into a range of other identifiers. Below, we exemplify transforming 1,400 PubChem compound identifiers (CID) of the chemical reference compounds available in the Fiehn laboratory into a single file that can be downloaded as standard office document (here: xls).

| | PubChem | InChI Hash Key | KEGG | CAS | CHEBI | LIPID MAP | HMDB | InChi | Formula | Exact Mass |
|---|---|---|---|---|---|---|---|---|---|---|
| 2-Undecanone | 8163 | KYWIYKKSMDLRDC-UHFFFAOYSA-N | C01875 | 112-12-9 | 17700 | LMFA12000002 | | InChI=1S/C | C11H22O | 170.167 |
| Apigenin | 5280443 | KZNIFHPLKGYRTM-UHFFFAOYSA-N | C01477 | 520-36-5 | 18388 | LMPK12110005 | HMDB0021 | InChI=1S/C | C15H10O5 | 270.053 |
| L-valine | 6287 | KZSNJWFQEVHDMF-BYPYZUCNSA-N | C00183 | 72-18-4, 5 | 16414 | | HMDB0087 | InChI=1S/C | C5H11NO2 | 117.079 |
| valine | 1182 | KZSNJWFQEVHDMF-UHFFFAOYSA-N | C16436 | 516-06-3 | | LMFA01100046 | | InChI=1S/C | C5H11NO2 | 117.079 |
| Igepal CA (630) | 24775 | LBCZOTMMGHGTPH-UHFFFAOYSA-N | | | | | | InChI=1S/C | C18H30O3 | 294.219 |
| N-carbamyl-L-glutamic acid | 3679006 | LCQLHJZYVOQKHU-UHFFFAOYSA-N | | | | | | InChI=1S/C | C6H10N2O5 | 190.059 |
| pyruvic acid | 1060 | LCTONWCANYUPML-UHFFFAOYSA-N | C00022 | 127-17-3 | 32816 | LMFA01060077 | HMDB002 | InChI=1S/C | C3H4O3 | 88.016 |

Several key features become apparent. First, all structures are now uniquely identified, and isomers (such as valine and L-valine) can be sorted next to each other, to identify structural overlaps. Secondly, not all chemicals are represented in biochemical databases. While pyruvate is found in all databases, LipidMAPS does not list the (biochemically more likely) L-valine but only the stereochemically undefined version valine.



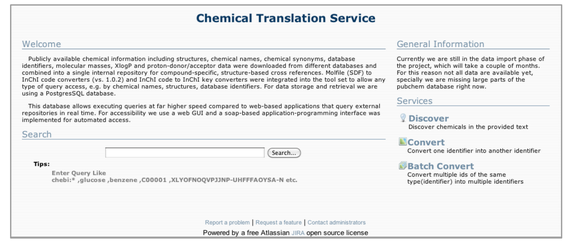

CTS Sequence Diagram

## Chemical Translation Service

The Chemical translation service (CTS) is online for public use (see screenshot below), available at **cts.fiehnlab.ucdavis.edu**
The program is completely based on open source software. Most importantly the Groovy on Rails Framework (Grails) in the Version 1.2.2.

The main reason for using Grails was the simplicity and high speed of development compared to classic approaches like Struts or Java Server Faces. For the rendering of molecules we utilized the Chemical Development Toolkit version 1.3.1. The reading and parsing of Molecular files was done by using a modified version of the JNI-InChi Library, which we enhanced with the support for the InChi 1.02b specification.

For the actual data storage we are using a PostgreSQL 8.2 Database running on Rocks Linux 5.3.




## Conclusions

The Chemical Translation Service is online and user friendly for all chemists. No password or programming experience is required. In order to provide high query speed, we have opted to import all major biochemical databases instead of delegating the queries to a multitude of external databases or websites.
At current, the 40 million entries in PubChem are not completely imported. Additional databases such as MetaCyc will be added as input/ output query options.