Chemical Translation Service

A web-based tool
to improve standardization of chemical identifiers.

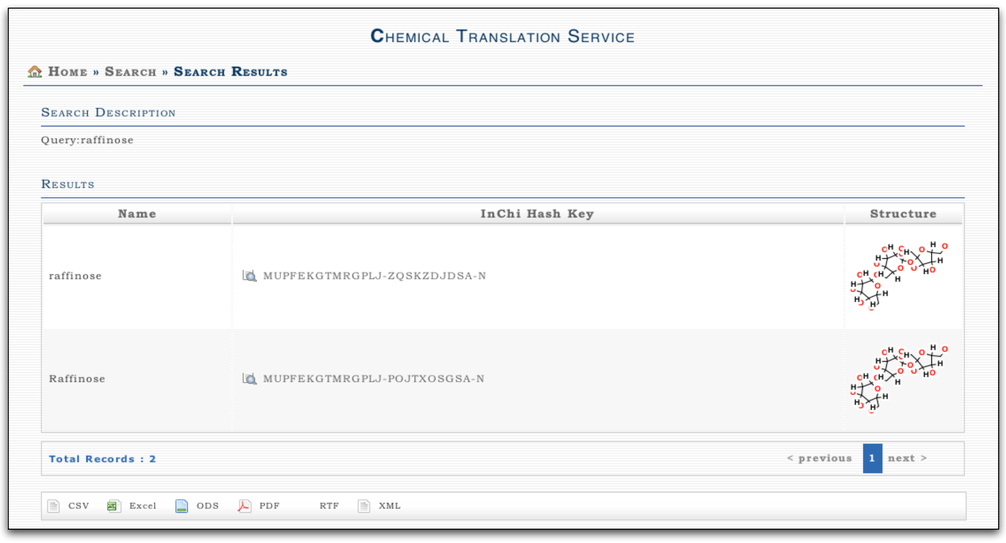
Pradeep Kumar Haldiya, Gert Wohlgemuth, Oliver Fiehn

Objective

Metabolomic publications and databases use different database identifiers or even trivial names, which disable queries across databases or between studies. The optimal way to annotate small molecules (0-1500 u) is by the corresponding structure in a machine-readable format. Open-source structure codes were 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 available for analytical chemists who lack time or expertise in computational chemistry to transform lists of identified metabolites into structure codes or to batch converts metabolite database identifiers (like CAS, KEGG, HMDB or PubChem) into unambiguous and public InChI keys.

Why are chemical names 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.



Chemical Translation Service

The Chemical translation service (CTS) is online for public use (see screenshot below). Chemical Translation Service that performs batch conversions of the most common compound identifiers, including CAS, CHEBI, compound formulas, Human Metabolome Database HMDB, InChI, InChIKey, IUPAC name, KEGG, LipidMaps, PubChem CID+SID, SMILES and chemical synonym names.

It also provides REST-based web service. This web service is required primarily for automatic access from other programs and databases.

The Chemical translation service (CTS) 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.

Why are InChI Hash Keys universal chemical identifier?

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

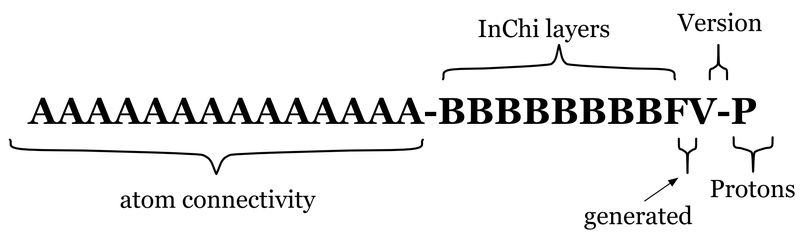
(InChI=1S/C18H32O16/c19-1-5-8(22)11(25)13(27)16(31-5)30-3-7-9(23)12(26)14(28)17(32-7)34-18(4-21)15(29)10(24)6(2-20)33-18/h5-17,19-29H,1-4H2/t5-,6-,7-,8+,9-,10-,11+,12+,13-,14-,15+,16+,17-,18+/m1/s1)

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

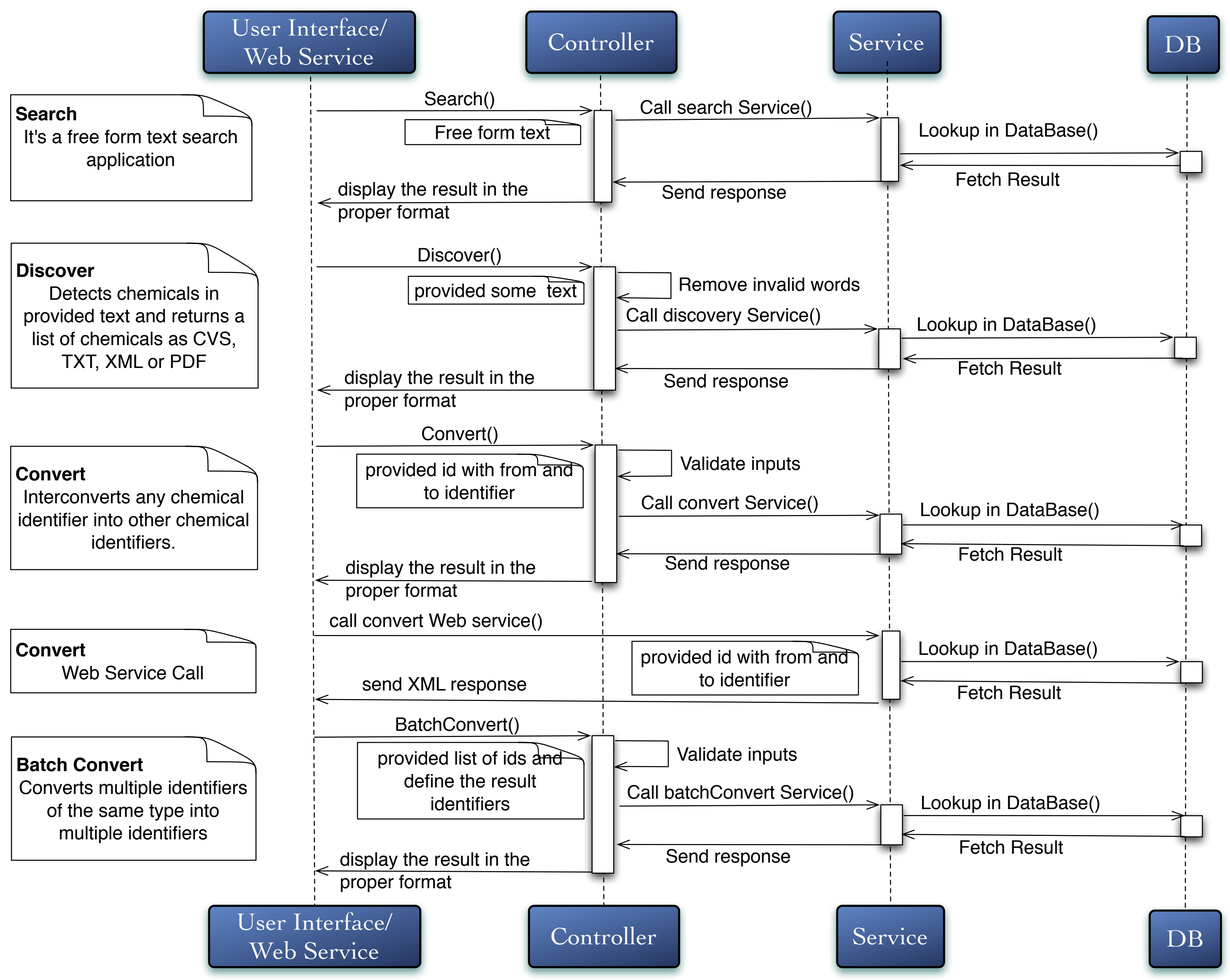
MUPFEKGTMRGPLJ-ZQSKZDJDSA-N

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.



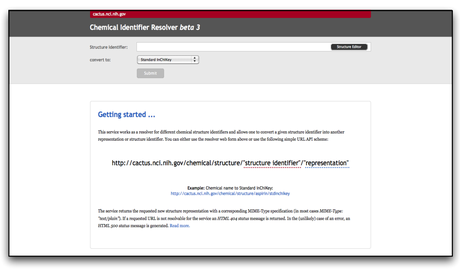
Sequence Diagram



Similar tools are less powerful

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

1. Easy to use
2. Web based
3. One at a time conversion
4. Batch convert not available
5. Accessible for programmers



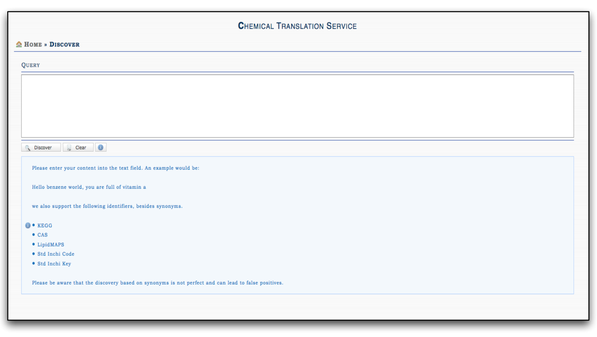
The MetMask Project
http://metmask.sourceforge.net/

1. Web based interface
2. R based interface
3. Command line interface
4. Hard to install
5. Web interface does not allow batch converts

Applications

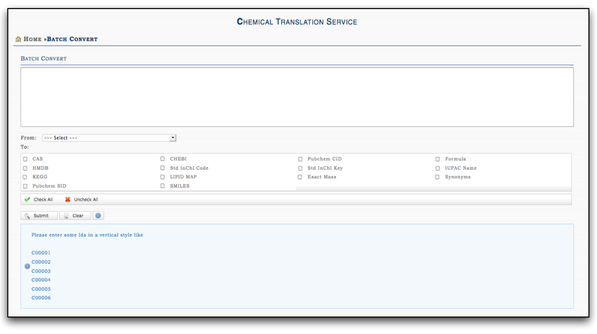
Search

Discover



Convert

Batch Convert



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.

References

Application Link
http://cts.fiehnlab.ucdavis.edu

Source Link
svn checkout http://chemical-compound-repository.googlecode.com/svn/trunk/ chemical-compound-repository-read-only

Contact us

Metabolomics Fiehn Lab
University of California, Davis
Davis, CA 95616 USA
http://fiehnlab.ucdavis.edu and
Mail to: wohlgemuth@ucdavis.edu, phaldiya@ucdavis.edu