BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 20 2010, pages 2647–2648
doi:10.1093/bioinformatics/btq476

*Databases and ontologies*

Advance Access publication September 9, 2010

# The Chemical Translation Service—a web-based tool to improve standardization of metabolomic reports

Gert Wohlgemuth[1], Pradeep Kumar Haldiya[1], Egon Willighagen[2], Tobias Kind[1] and Oliver Fiehn[1,*]

[1]University of California Davis, CA, Genome Center, USA and [2]Department of Pharmaceutical Science, Uppsala University, Sweden

Associate Editor: Jonathan Wren

**ABSTRACT**

**Summary:** Metabolomic publications and databases use different database identifiers or even trivial names which disable queries across databases or between studies. The best way to annotate metabolites is by chemical structures, encoded by the International Chemical Identifier code (InChI) or InChIKey. We have implemented a web-based 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. Batch conversion downloads of 1410 CIDs are performed in 2.5 min. Structures are automatically displayed.

**Implementation:** The software was implemented in Groovy and JAVA, the web frontend was implemented in GRAILS and the database used was PostgreSQL.

**Availability:** The source code and an online web interface are freely available. Chemical Translation Service (CTS): http://cts.fiehnlab.ucdavis.edu

**Contact:** ofiehn@ucdavis.edu

Received on May 11, 2010; revised on July 22, 2010; accepted on August 14, 2010

## 1 INTRODUCTION

The Metabolomics Standards Initiative (MSI) proposed the use of database identifiers for publishing reports in peer-reviewed journals or in data repositories (Sumner *et al.*, 2007), but the MSI did not specify best practice standards which identifier to use. Consequently, metabolomic data are presented by a wide variety of identifiers, mostly using publicly available databases such as KEGG, HMDB or PubChem. In other cases, authors merely use compound names without referencing to databases. Compound names are very poor descriptors (Kind *et al.*, 2009) as names often cannot be unambiguously mapped to authentic chemical structures, either because of missing chiral information (D, L) or because each chemical structure is associated with many synonym names, some of which may also be used for other structures. In addition, no database contains all identifiers of all other repositories. For example, KEGG LIGAND is a popular biochemical pathways database, but it is incomplete for many compounds found in human organs as given in the Human Metabolome database HMDB. Although each database lists outlinks to other databases, no single database provides comprehensive mapping options to other databases, and rarely there are batch query options offered.

Analytical chemists and biochemists may not be used to standard structure codes or lack expertise for downloading databases or installing software. We here present a publicly available tool that enables researchers to quickly convert lists of compound database identifiers, including the important InChI Keys.

**Software:** The Chemical Translation Service (CTS) was implemented using the programming languages Groovy (v1.7) and Java (v6.20). The open source web application framework Grails 1.2.2 and freely available plugins were used for the development of web services. For data storage, the PostgreSQL database, an open-source object-relational database management system was used. Easy access from other languages and platforms is provided via SOAP (Simple Object Access Protocol) web services.

**Hardware:** The used hardware was a dual quad-core Intel X5450 Xeon based server with 16 GB of RAM and an SSD (Solid State Disk) storage array with three disks in a Raid-0 configuration to store the Lucene index files. The average disk throughput was 600 MB/s. An additional SAS (Serial Attached SCSI) storage array with 16 disks in Raid-6 configuration was used to store the database content.

**Source Code:** The source code is hosted as a Google Code Project at http://code.google.com/p/chemical-compound-repository/

**Web Front End:** The database is freely available under: http://cts.fiehnlab.ucdavis.edu

## 2 RESULTS

The CTS consists of three major services.

(1) **The Discovery Service** detects chemicals in provided text and returns a list of chemicals as CSV, TXT, XML or PDF.

(2) **The Convert Service** interconverts any chemical identifier into other chemical identifiers.

(3) **The Batch Convert Service** converts multiple identifiers of the same type into multiple identifiers.

We recommend that users, especially chemists and biologists, use CTS for standardizing metabolomic reports into MSI-compliant formats before publishing data. Single identifiers can be converted such as KEGG identifier to PubChem ID, or SMILES to KEGG ID. Importantly, batch convert services are supported (Table 1),

*To whom correspondence should be addressed.

**Table 1.** Example result converting seven PubChem CIDs

| NAME | PubChem CID | InChI Hash Key | KEGG | CAS | ChEBI | LIPID MAPS | HMDB | Formula | Exact Mass |
|---|---|---|---|---|---|---|---|---|---|
| 2-Undecanone | 8163 | KYWIYKKSMDLRDC-UHFFFAOYSA-N | C01875 | 112-12-9 | 17 700 | LMFA12000002 | | C11H22O | 170.167 |
| Apigenin | 5 280 443 | KZNIFHPLKGYRTM-UHFFFAOYSA-N | C01477 | 520-36-5 | 18 388 | LMPK12110005 | HMDB02124 | C15H10O5 | 270.053 |
| L-valine | 6287 | KZSNJWFQEVHDMF-BYPYZUCNSA-N | C00183 | 72-18-4 | 16 414 | | HMDB00883 | C5H11NO2 | 117.079 |
| Valine | 1182 | KZSNJWFQEVHDMF-UHFFFAOYSA-N | C16436 | 516-06-3 | | LMFA01100046 | | C5H11NO2 | 117.079 |
| Igepal CA (630) | 24 775 | LBCZOTMMGHGTPH-UHFFFAOYSA-N | | | | | | C18H30O3 | 294.219 |
| *N*-carbamoyl-glutamic acid | 3 679 006 | LCQLHJZYVOQKHU-UHFFFAOYSA-N | | | | | | C6H10N2O5 | 190.059 |
| Pyruvic acid | 1060 | LCTONWCANYUPML-UHFFFAOYSA-N | C00022 | 127-17-3 | 32 816 | LMFA01060077 | HMDB00243 | C3H4O3 | 88.016 |

submitting multiple chemical identifiers of the same type and yielding a table of user-selected other identifiers. We currently use the CTS service for standardizing reports from different platforms in collaborative projects.

We provide different download formats including CSV, XLS, PDF and XML, in order to empower researchers to sort, compare and arrange properties of their lists in standard office software. For example, L-valine and valine differ only in the description of stereoconfiguration. Sorting via the atom connectivity part of the InChI hash keys (labeled red in Table 1) facilitates recognition of isomers or detection of doublets. It is apparent that databases deal differently with stereo configurations, as LipidMaps does not provide valine in its regular L-configuration whereas HMDB does and KEGG provides both valine enantiomers. Other metabolites such as pyruvate, the end product of glycolysis, is given by all databases, whereas the achiral form of *N*-carbamoylglutamate is not represented in any other database.

The Chemical Discovery Service annotates chemicals from any given text submitted in ASCII format by text mining. A hybrid approach using word stop-lists and fuzzy regular expression matching was used for detection of chemicals and a subsequent database matching was used for the output of all possible chemical identifiers. For example, 269 chemical names were given in the supplementary material of a published non-MSI compliant report (Sreekumar *et al*., 2009). These names could be read into 195 structures via synonym queries, e.g. for overlap analysis with other platforms or query of related studies. Misspelled compounds ('nonate') did not retrieve hits.

In comparison to the Chemical Resolver Identifier beta2 (Sitzmann, 2009), the CTS software comprises more commonly used database compound identifier inputs and allows any combination of output identifiers. Importantly, CTS also enables batch queries, a feature that is not given in other resources including the recently published MetMask tool (Redestig *et al*., 2010). MetMask requires installation of in-house databases and provides only limited services via web based queries. CTS maintains high-query speeds because all required data were downloaded from publicly available databases (except CAS) into a new in-house DB. Due to the proprietary nature of the Chemical Abstract Services CAS, these database identifiers are incomplete as only publicly available CAS numbers were downloaded. At present, the CTS database hosts 3.8 million entries which are indexed by InChI hash keys but continues to grow in size limited by PubChem data import speed. The CTS input databases are renewed monthly for new DB updates (e.g. HMDB, LipidMaps, KEGG).

The following advantages distinguish the CTS from other services:

- Generic compound searches over millions of compounds; can be limited by source DB query language and Lucene index.
- Regular-expression-based discovery of existing chemical names in any given text.
- Batch- or single-conversion of many publicly available chemical identifiers. Further databases (MetaCyc) are being added.
- Export of results into different formats including CSV, XLS, PDF and XML.
- Access via simple and intuitive web GUIs, optimized for the most common web browsers including Mozilla Firefox, Microsoft Internet Explorer and Apple Safari.
- Access via SOAP (Simple Object Access Protocol) web services, enabling other programming languages and platforms.
- InChI hash keys as main link between identifiers, providing structure-based indexing. InChI codes can be used in chemical databases for extensive substructure comparisons.
- Inclusion of further compound information, from monoisotope exact masses to proton donor potential, aiding identifications in mass spectrometry based metabolomic profiles.

## ACKNOWLEDGEMENTS

We thank Dinesh K. Barupal (Fiehnlab), P. Karp (BioCyc), Rima Kaddurah-Daouk (RKD, Duke, PI of the Pharmacometabolomics Network) and D. Wishart (HMDB) for stimulating discussions.

*Funding*: National Institutes of Health (R24 GM078233) and (RC2 GM092729) to RKD, Duke.

*Conflict of Interest*: none declared.

## REFERENCES

Kind,T. *et al*. (2009) How large is the metabolome? A critical analysis of data exchange practices in chemistry. *PLoS One*, **4**, Article No: e5440.

Redestig,H. *et al*. (2010) Consolidating metabolite identifiers to enable contextual and multi-platform metabolomics data analysis. *BMC Bioinformatics*, **11**, 214.

Sitzmann,M. (2009) NCI/CADD Chemical Identifier Resolver. Available at http://cactus.nci.nih.gov/chemical/structure (last accessed date May 9, 2010).

Sreekumar,A. *et al*. (2009) Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression, *Nature*, **457**, 910–914.

Sumner,L.W. *et al*. (2007) Proposed minimum reporting standards for chemical analysis. *Metabolomics*, **3**, 211–221.