

Semantic access to chemistry data with the ChEBI ontology and web services

Nico Adams, Paula de Matos, Adriano Dekker, Marcus Ennis, Janna Hastings, Kenneth Haug, Duncan Hull, Zara Josephs, Pablo Moreno, Steve Turner and Christoph Steinbeck

European Bioinformatics Institute, Hinxton, UK

Abstract. The Chemical Entities of Biological Interest (ChEBI) ontology is an ontology of chemical entities and their roles, being developed at the European Bioinformatics Institute (EBI). Recent developments include a submission tool for direct user submissions and enhancements to the search facilities available by web services.

Introduction

The recent broadening of the scope of efforts to understand whole-systems biology has created a need to understand biochemical small molecules and the roles they play in metabolism and other pathways and reactions. Small molecule data is typically grouped around small molecule structures which again can be marked up as substrates of enzymes, as participants in a particular metabolic pathway or as subjects of transport by a membrane protein. To enable semantic access, querying and visualisation of this small molecule data, there is a need for an ontology which organises the data according to chemical structure, reactivity and biological activity.

The Chemical Entities of Biological Interest (ChEBI) ontology is an ontology of chemical entities and their roles, being developed at the European Bioinformatics Institute (EBI) [1, 2]. ChEBI is molecule-centric with a number of annotations grouped around the 2D molecular graphs (connection tables) of small molecules. Each entry is manually annotated by expert annotators before being released. As a dictionary, the nomenclature provided includes an unambiguous ChEBI recommended name, IUPAC names, International Nonproprietary Names (INNs) and synonyms. Where feasible a molecular graph is provided accompanied by the chemical structural representations InChI, InChIKey, and SMILES. Additional chemical data such as formula, mass and charge are provided. Each entry is extensively cross-referenced. External databases link to ChEBI via the unique and stable ChEBI identifier. All data in ChEBI is publicly available, open and free for redistribution.

ChEBI is available via a public web facility (<http://www.ebi.ac.uk/chebi/>), FTP downloads (<http://www.ebi.ac.uk/chebi/downloadsForward.do>), and via web services. Additionally, ChEBI is redistributed through several life science-wide semantic resources, including the OBO Foundry [3] and BioGateway [4].

This poster describes recent developments in the ChEBI ontology and web services in 2009.

Submission Tool

To invite the community to participate more directly in the future growth and development of ChEBI, we have developed a web-based software utility to enable direct user submissions. User submissions are then publicly available (after the next release cycle) and cited to the submitter. Submitters have the option to remain anonymous if they wish. The ChEBI submission tool is available online at <https://www.ebi.ac.uk/chebi/submissions>.

The minimal information which is required for a ChEBI submission is a name, which must be unique within the database; either a text definition or a chemical structure; and a primary classification within the ontology. For example, the term 'insecticide' might be submitted to the ontology with definition 'A substance used to destroy pests of the class Insecta.' and primary classification 'is a pesticide (CHEBI:25944)'. Submission of the most complete dataset possible is encouraged, thus it is possible to add multiple synonyms and database cross-references, as well as to create multiple relationships within the ontology.

The captured submission is automatically validated for uniqueness, both of name and chemical structure (where applicable), and correctness. Final submission is not possible until all errors have been resolved. Once submitted, a submission receives a unique and stable ChEBI identifier.

ChEBI Web Services

With the October 2009 release, the ChEBI web services have been extended to enable new search functionality to be available via the web service. In particular, the facility to search by chemical structure has been added to the web service. The chemical structure search in ChEBI is backed by the OrChem Oracle chemistry cartridge [5]. The methods exposed by the ChEBI web service are:

```
getLiteEntity  
getStructureSearch  
getCompleteEntity  
getCompleteEntityByList  
getOntologyParents  
getOntologyChildren
```

Further details on the ChEBI web service implementation can be accessed at <http://www.ebi.ac.uk/chebi/webServices.do>.

Future Developments

Extensive refactoring of the ChEBI ontology is under way to increase interoperability with other resources in the OBO Foundry [3] and to align with the common upper-level ontology BFO. Furthermore, with the October 2009 release, ChEBI has incorporated the the ChEMBL drug discovery dataset [6], increasing its small molecules coverage to over 450,000 entities. However, the imported entities have not yet been classified in the ChEBI ontology. Future work will focus on algorithms for automatic structure-based classification of chemical entities into the ontology classes, increasing the size of the ChEBI ontology to cover all the chemicals of the ChEMBL drug discovery dataset.

Acknowledgements

ChEBI is funded by the European Commission under SLING, grant agreement number 226073 (Integrating Activity) within Research Infrastructures of the FP7 Capacities Specific Programme; and by the BBSRC, grant agreement number BB/G022747/1 within the “Bioinformatics and biological resources” fund.

References

1. *Chemical entities of biological interest: an update.* de Matos, P., Alcántara, R., Dekker, A., Ennis, M., Hastings, J., Haug, K., Spiteri, I., Turner, S., and Steinbeck, C. (2009). *Nucleic Acids Res.* doi:10.1093/nar/gkp886.
2. *ChEBI: a database and ontology for chemical entities of biological interest.* Degtyarenko, K., de Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., Alcántara, R., Darsow, M., Guedj, M. and Ashburner, M. (2008) *Nucleic Acids Res.* 36, D344D350
3. *The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration.* Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L.J., Eilbeck, K., Ireland, A., Mungall, C.J., OBI Consortium, Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.-A., Scheuermann, R.H., Shah, N., Whetzel, P.L. and Lewis, S. (2007) *Nature Biotech.* 25, 12511255.
4. *BioGateway*, <http://www.semantic-systems-biology.org/biogateway/>, last accessed November 2009.
5. *An open source chemistry search engine for Oracle.* Rijnbeek, M. and Steinbeck, S. *Journal of Cheminformatics*, in press.
6. *ChEMBL*. <http://www.ebi.ac.uk/chembl/>, last accessed November 2009.