EMBL-EBI

Towards **Automatic classification** within the **ChEBI ontology**

Janna Hastings, Paula de Matos, Marcus Ennis and Christoph Steinbeck

Introduction

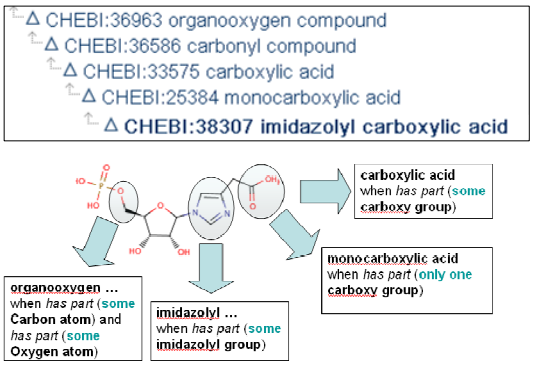
'Small molecules' are a core element of biomedical data. They serve as signalling molecules, influence the behaviour of proteins and enzymes and generally form the small molecule metabolism in living organisms. As such, chemical ontologies, which provide stable identifiers and a shared vocabulary for small molecules, are crucial to enabling the interoperation and semantic querying of such data. ChEBI[1] is a publicly available, manually annotated database of chemical entities and a chemical ontology, containing approximately 18000 annotated entities at the last release (June 2009). ChEBI is widely used for annotation of chemicals within biological databases, for text-mining, and as a chemistry component for intelligent and semantic web applications. Classification of entities into ontologies may be accomplished manually by careful annotation, as is the case for ChEBI. However, the amount of information on small molecules, like other areas of biomedical data, is growing exponentially and thus there is no hope to continue to deal with them *solely* by manual means.

To date, automatic classification of chemical entities within the ChEBI ontology has not been possible. However, with the growth in usage and the ever-increasing backlog of requests from users both within the UK and abroad, we feel that automatic classification is essential. The size of the domain of biologically interesting chemicals is essentially infinite, but even if we focus only on those chemical entities which are actively required for systems biological research and drug discovery, we find that the size of the domain vastly exceeds the ability of manual classification methods to provide a high percentage of coverage.

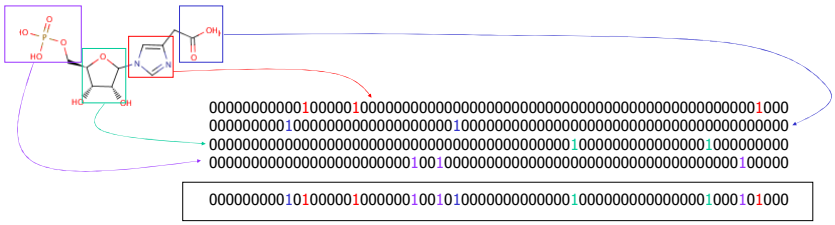
The domain of chemistry is more tractable for automatic classification into ontologies than the domains of many other bio-ontologies such as that of the Gene Ontology, as chemical structures are already computationally accessible in standard formats, and the structures are integrally related to the ontological classification.

Our work will synthesise logic-based approaches with relevant cheminformatics techniques.

1. Semantically define classes



2. Use cheminformatics tools to extract features from structures

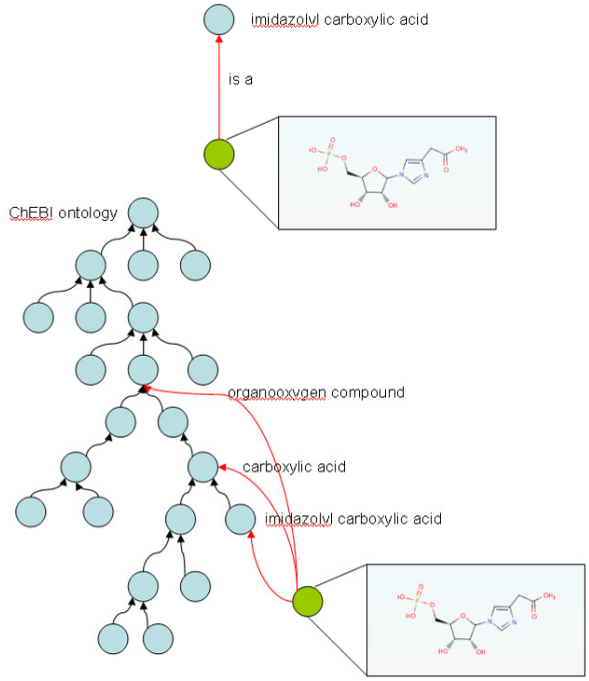


3. Apply automated classification

Description logic-based reasoning technology is a prime candidate for development of such an automatic classification system as it allows the rules of the classification system to be encoded within the knowledgebase. Already at 18000 entities, ChEBI is a fair size for a real-world application of description logic reasoning technology, and as the ontology is enhanced with a richer density of asserted relationships, the classification will become more complex and challenging. We have successfully tested a description logic-based classification of chemical entities based on specified structural properties using the hypertableaux-based HermiT[2] reasoner, and found it to be sufficiently efficient to be feasible for use in a production environment on a database of the size that ChEBI is now.

Future work will combine the cheminformatics algorithms for feature detection and graph isomorphism such as in the open source CDK[3] library, with logic-based and stochastic algorithms for automated classification to provide a general solution for automated structure-based classification of chemical entities.

4. Extract and assert most specific classification



References

1. Degtyarenko, K., de Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., Alcántara, R., Darsow, M., Guedj, M. and Ashburner, M. (2008) ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res.* **36**, D344–D350.
2. Shearer, R., Motik, B. and Horrocks, I. (2008) HermiT: A Highly-Efficient OWL Reasoner. In Dolbear, C., Ruttenberg, A. and Sattler, U. (Eds.), *Proceedings of the 5th Workshop on OWL: Experiences and Directions (OWLED 2008)*, Karlsruhe, Germany, October 26–27, 2008.
3. Steinbeck, C., Han, Y., Kuhn, S., Horlacher, O., Luttmann, E., and Willighagen, E. (2003) The Chemistry Development Kit (CDK): an open-source Java library for chemo- and bioinformatics. *J. Chem. Inf. Comput. Sci.* **43**, 493–500.

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.

Janna Hastings
Software Engineer
Cheminformatics and Metabolism team

janna.hastings@ebi.ac.uk
chebi-help@ebi.ac.uk

EMBL- EBI
Wellcome Trust Genome Campus
Hinxton
Cambridge
CB10 1SD
UK

T +44 (0) 1223 494 444
F +44 (0) 1223 494 468
http://www.ebi.ac.uk

http://www.ebi.ac.uk/chebi

Nature Precedings : doi:10.1038/npre.2009.3525.1 : Posted 31 Jul 2009