Beilstein-Institut
Experimental Standard Conditions of Enzyme Characterization
September 12th – 16th, 2011, Rüdesheim/Rhein, Germany

Ontology-based search in SABIO-RK

Ulrike Wittig*, Enkhjargal Algaa, Andreas Weidemann, Renate Kania, Maja Rey, Martin Golebiewski, Lei Shi, Lenneke Jong and Wolfgang Müller

Scientific Databases and Visualization Group,
Heidelberg Institute for Theoretical Studies (HITS),
Schloss-Wolfsbrunnenweg 35, 69118 Heidelberg, Germany

E-Mail: *Ulrike.Wittig@h-its.org

Received: 31st January 2012/Published: 15th February 2013

Abstract

The SABIO-RK database (http://sabio.h-its.org/) is established as a resource for biochemical reactions and their kinetic data. Data are manually extracted from scientific literature and stored in a structured and standardised format. Additionally SABIO-RK allows direct submission of data from lab experiments in an automated workflow, e. g. within project collaborations for storage and exchange of unpublished experimental results and later publishing the data. To access the kinetic data in the database, web interfaces and web services are available offering complex searches using various criteria. For specific search criteria different classification levels of organisms, tissues, and reactants, can be selected based on biological ontologies. Biological ontologies are developed for a hierarchical classification of biological objects and for modelling a domain using shared vocabularies. Ontological relations are implemented in SABIO-RK to extend the search functionalities.

Introduction

SABIO-RK [1] is a manually curated database containing kinetic data of biochemical reactions and their related information. The data are manually extracted from literature [2] or are automatically submitted from lab experiments [3]. SABIO-RK combines available kinetic parameters along with their corresponding rate equations, as well as kinetic law and parameter types and experimental and environmental conditions (pH, temperature, buffer) under which the kinetic data were determined. It stores detailed information about the biochemical reactions and pathways including their reaction participants (substrates, products), modifiers (inhibitors, activators, cofactors), cellular location, enzyme information (e. g. UniProt [4] accession number, EC number [5], isozymes, protein complex composition, wild type/mutant information, molecular weight), kinetic parameters, corresponding rate equation, and biological source (organism, tissue, cell location). As data within SABIO-RK are strongly interlinked, they are mainly extracted from literature. [6] Each entry within SABIO-RK is referenced to the original publication and data directly submitted from lab experiments are referenced to the original source of the raw data (e. g. the MeMo-RK database storing the experimental raw data of our collaboration partners in Manchester). Biology and biochemistry experts are doing the curation and annotation to controlled vocabularies, ontologies and external data sources.

Ontologies and Controlled Vocabularies

A defined and shared vocabulary is important to avoid misinterpretations and helps to exchange data between resources correctly. Controlled vocabularies provide predefined terms for a specific domain, which have been selected by domain experts. In many cases these vocabularies are structured like taxonomies or hierarchical classifications and typically include so-called is_a and part_of relationships. To illustrate these relationships an example: Vehicles are cars, bicycles, trains and so on. Therefore the entities *car*, *bicycle* and *train* are denoted by *is_a relationships* to *vehicle*, even though all three have different characteristics. A car consists of several parts, including wheels or an engine, thus the entities *wheel* and *engine* have *part_of* relationships to *car*.

Ontologies, including controlled vocabularies and all relationships between objects, represent the knowledge of a well-defined domain and are used to describe properties of the objects.

Biological ontologies and controlled vocabularies are also being developed to force the usage of a shared language for naming of biological objects. By agreeing on a particular ontological representation, for example the development of biological databases, a common vocabulary can be used for the analysis of biological data and their comparison between different resources. In biology, chemistry or medicine, there are many synonymous terms, abbreviations and acronyms that can refer to the same object. For example *alpha-ketoglutarate*

and *2-oxoglutarate* are synonymous names of the same chemical compound. For the identification of synonyms of chemical compounds our group developed ChemHits (http://sabio.h-its.org/chemHits/) as an analysis tool for chemical compound names.

With the increasing amount of data generated from high-throughput experiments, there is an increasing need for structuring the data. Ontologies assure the unambiguous identification of objects you are looking for and thus help to analyse the data correctly. Additionally ontologies also assist to build and maintain ontologies themselves. [7 – 9]

In biology the systematic definition of terms and hierarchical relations is very old. For the first time it was used to create a taxonomy for organisms. Organism names could change over time because new characteristics are described and the classification of the organism changes within the taxonomy: e. g. *Streptococcus faecalis* was used in the past and now it is changed to the genus *Enterococcus faecalis*. Because both organism names are used in publications a classification scheme and controlled vocabulary is needed including all synonyms. Another system for nomenclature and classification was established in the 1950's. At that time the number of newly discovered enzymes increased very rapidly. Now there was the need to unravel the different names for the same enzyme and to establish a system where enzymes with similar functions could be grouped in classes, leading to the EC classification system [5].

Most of the available biological ontologies and controlled vocabulary only contain hierarchical is_a relationships between objects. Based on these parent-child relations classifications and hierarchical structures of objects can be derived. Examples of such classifications are the enzyme classification system offered by IUBMB [5] or the organism taxonomy provided by NCBI [10]. There are also ontologies which have not only is_a, but also part_of relationships between objects. The BRENDA Tissue Ontology (BTO) [11] for example, contains information about tissues and cell types. The relationships between tissues and cell types are represented as part_of relationships, e. g. hepatocyte is part_of liver.

Beside these, the ChEBI ontology for small chemical compounds [12] is one of the few ontologies containing additional relationships, apart from is_a and part_of relations. In addition, the ChEBI ontology contains chemistry-specific relationships which can be used to convey additional information about the chemical compound. It includes acid-base relations, relations between different stereoisomeric forms of chemical compounds, or relations to define the functional role of a chemical compound.

To improve the functionality of the SABIO-RK database by offering extended queries hierarchies (is_a and part_of relations) based on the ontological relations from several ontologies are implemented in the SABIO-RK search options.

During the process of data insertion and curation controlled vocabularies are available for students extracting the data from literature and for database curators working on the data annotation and the quality control. For consistency and to avoid duplicate entries, for example lists of compounds, reactions, organisms, tissues, cellular locations, kinetic law types, parameter types, and units already existing in the SABIO-RK database are provided as selection lists for students and curators. Most of these lists are first generated by extracting terms from external sources and extended by terms extracted from literature. Enzyme names and EC numbers are extracted from IUBMB, organism names from NCBI taxonomy, tissues and cellular locations from BRENDA [13], types of kinetic laws and parameters from Systems Biology Ontology (SBO) [14], and units from the International System of Units [15]. Synonymic terms are referred to the same recommended term to enable the search for alternative names. New terms extracted from literature are added to the term lists and in many cases they were also submitted as new terms to the corresponding ontology. Therefore the SABIO-RK curators also help in the further development of external controlled vocabularies and ontologies (e. g. SBO or BTO).

Table 1. Overview of annotations and links in SABIO-RK to external databases, ontologies, and controlled vocabularies [4 – 5, 10 – 14, 18 – 23] and options how to search for them in SABIO-RK.

	Annotation to		**Link to**	**Search by**	
	Controlled vocabulary/ Ontology	Database		Name	ID
Reaction		KEGG	KEGG		X
Compound	ChEBI	ChEBI	ChEBI	X	X
		KEGG	KEGG	X	X
		PubChem	PubChem	X	
Enzyme	IUBMB	ExplorEnz	ExplorEnz		
		BRENDA	BRENDA		
		ExPASy	ExPASy		
		KEGG	KEGG		
		Reactome	Reactome		
		IntEnz	IntEnz		
			IUBMB	X	X
Protein		UniProt	UniProt		X
Organism	NCBI Taxonomy		NCBI Taxonomy	X	
Tissue	BTO		BTO	X	
Cell location			Gene Ontology	X	
Kinetic parameter type	SBO			X	
Kinetic law type	SBO			X	
Role in reaction	SBO			X	
Publication		PubMed	PubMed		X

Based on the known terms from selection lists data in SABIO-RK can be unambiguously identified and annotated to external data resources and ontologies. Biological ontologies and controlled vocabularies used for annotations in SABIO-RK are for example ChEBI, SBO, BTO, and NCBI taxonomy. In addition, comprehensive annotations to external databases enable the user to obtain further details, for example about reactions, compounds, enzymes, proteins, tissues, or organisms. A detailed listing of available annotations and links to databases and ontologies in SABIO-RK is represented in Table 1.

Annotations to external databases and ontologies together with the data from SABIO-RK can be also exported in SBML (Systems Biology Markup Language) [16], compliant with the MIRIAM standard [17], e. g. for setting up biochemical reaction network models. SABIO-RK reaction and kinetic law identifiers both are themselves listed as MIRIAM data types and are also exported with the data from allowing tracking of the data back to SABIO-RK as data source.

Ontology-based Data Search in SABIO-RK

Data in SABIO-RK can be accessed via web-based user interfaces (classical and new interface) or via web-services (our previous SOAP slowly phasing out and being replaced by a more modern RESTful interface). Queries are available using names of biochemical objects or internal SABIO-RK or external database identifiers. Searching reactions using identifiers is offered for example by reaction identifiers from KEGG or SABIO-RK, by compound identifiers from KEGG [18], ChEBI, PubChem [10], or SABIO-RK, by enzyme EC numbers, UniProt accession numbers, or PubMed identifiers. Most of the search criteria can be defined by name searches, e. g. compounds, enzymes, organisms, tissues, cell locations, kinetic parameters, or kinetic law definitions (Table 1).

Some of the search criteria (e. g. organism, tissue, and compound) are based on biological ontologies. Based on different classification levels SABIO-RK can be searched not only for specific terms but also for a group of terms which have similar characteristics defined by subparts in the hierarchical tree of an ontology. These hierarchical relationships between objects extracted from is_a relations are implemented in the SABIO-RK search options. Therefore the search for organisms can be either defined by using a specific term, e. g. *Rattus norvegicus* or can be extended by defining a group of organisms based on the NCBI taxonomy, e. g. the search for all rodents including for example mouse or hamster by using the search term *Rodentia (NCBI)*. The selection lists contain single terms and terms with *NCBI* in parenthesis representing terms from the NCBI taxonomy. The search for NCBI taxonomy terms always includes all ‘‘children’’ and ‘‘grandchildren’’ of this term extracted from the NCBI taxonomy tree.

The tissue search in SABIO-RK includes the possibility to use BRENDA Tissue Ontology (BTO) terms. Therefore the search for tissues can be also defined by simple terms or terms extracted from the BTO. For example the search for *kidney* gives fewer search results compared to *kidney (BTO)* because the last one not only searches for the tissue *kidney* but also for several kidney cell lines or subparts of the tissue which are represented as is_a or part_of relationships in the ontology. In Figure 1 a screenshot of the SABIO-RK web interface represents the ontology-based search results for organism *Rodentia (NCBI)* and tissue *kidney (BTO)*.

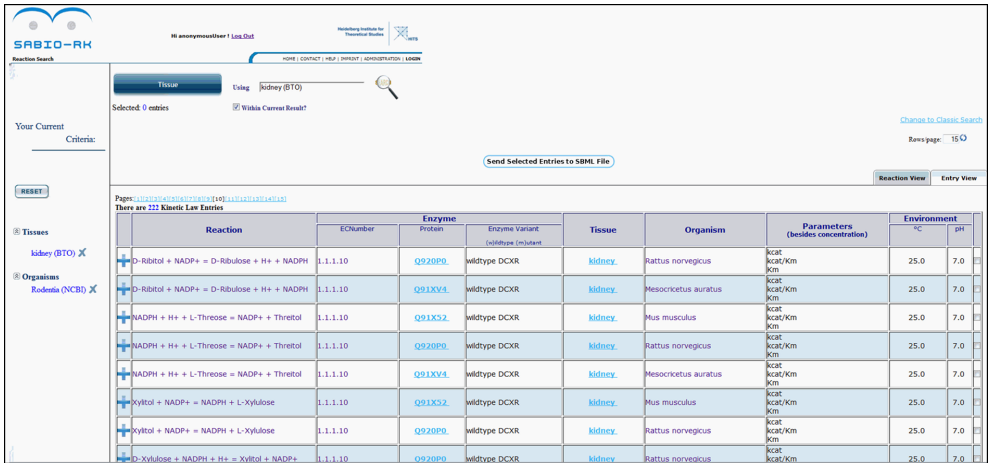

Figure 1. Screenshot of the SABIO-RK web interface for the search for *Rodentia (NCBI)* and *kidney (BTO)*.

At the moment only the classical SABIO-RK web-based interface offers for chemical compounds the search using ChEBI ontology terms. The is_a relationships extracted from the ChEBI sub-ontology ‘‘Molecular Structure’’ are implemented in the database search options for reaction participants to include the search for compound classes based on the hierarchical compound classification. For example queries can be defined to find all reactions containing an amino acid as reaction participant. Therefore for the query the reactant *amino acid* (CHEBI) has to be selected and the result would contain all reactions with any amino acid as substrate or product.

Based on the implementation of the hierarchical structures extracted from ontologies there is more functionality available so that the database user is able to decide which level of information is needed for the search. Queries including ontology-based terms result in more and comprehensive data. If a search result is dissatisfying an amplification of the search domain would help by using more general terms in the classification scheme. It also helps especially for tissue searches to get all related entries for one tissue including its cell lines because in the literature tissues and cell lines are described equivalent for similar

experiments. For example in some publications the results are specified for liver but the experiments are done on hepatocytes which are specific cells of the liver. Ontology-based searches for tissues using BTO includes both tissues and corresponding cell lines in one query to offer a combination of related terms with the regard to the contents.

On the other hand ontology-based searches offer the possibility to easier compare kinetic data within for example groups of organisms with same characteristics. Therefore SABIO-RK offers with these new organism search options the comparison of all kinetic data of a group of organisms like for example plants, mammals, or vertebrates.

Conclusion

SABIO-RK is a curated database containing biochemical reactions and their kinetics. Extracted from some selected biological ontologies and controlled vocabularies the hierarchical relations are implemented in the SABIO-RK search options for advanced functionality of the database. Annotations in SABIO-RK to controlled vocabularies, ontologies, and external databases allow comprehensive searches in the database, linking to external sources and the comparison of data. The search for organisms, tissues, and compounds can be extended by the search using ontological terms from NCBI taxonomy, BRENDA Tissue Ontology, and ChEBI Ontology, respectively. Future work will include the extension of the ontology-based searches for other data in SABIO-RK, especially for cell locations.

Acknowledgement

The SABIO-RK project is supported by the Klaus Tschira Foundation (http://www.klaus-tschira-stiftung.de/), the German Federal Ministry of Education and Research (http://www.bmbf.de/) through Virtual Liver and SysMO-LAB (Systems Biology of Microorganisms), and the DFG LIS (http://www.dfg.de/), under the short title ‘‘Integrated Immunoblot Environment’’.

References

[1] Wittig, U., Kania, R., Golebiewski, M., Rey, M., Shi, L., Jong, L., Algaa, E., Weidemann, A., Sauer-Danzwith, H., Mir, S., Krebs, O., Bittkowski, M., Wetsch, E., Rojas, I., Müller, W. (2012) SABIO-RK – database for biochemical reaction kinetics. *Nucleic Acids Res.* **40:**D 790 – 6.
doi: http://dx.doi.org/10.1093/nar/gkr1046.

[2] Wittig, U., Golebiewski, M., Kania, R., Krebs, O., Mir, S., Weidemann, A., Anstein, S., Saric, J., Rojas, I. (2006) SABIO-RK: integration and curation of reaction kinetics data. *Lecture Notes in Computer Science* **4075:**94 – 103.
doi: http://dx.doi.org/10.1007/11799511_9.

[3] Swainston, N., Golebiewski, M., Messiha, H.L., Malys, N., Kania, R., Kengne, S., Krebs, O., Mir, S., Sauer-Danzwith, H., Smallbone, K., Weidemann, A., Wittig, U., Kell, D.B., Mendes, P., Müller, W., Paton, N.W., Rojas, I. (2010) Enzyme kinetics informatics: from instrument to browser. *FEBS Journal* **277:**3769 – 79.
doi: http://dx.doi.org/10.1111/j.1742-4658.2010.07778.x.

[4] The UniProt Consortium (2011) Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res.* **39:**D214 – 9.
doi: http://dx.doi.org/10.1093/nar/gkq1020.

[5] IUBMB Enzyme Classification: http://www.chem.qmul.ac.uk/iubmb/enzyme/

[6] Wittig, U., Kania, R., Rojas, I., Müller, W. (2010) Herausforderungen bei der Extraktion von biochemischen Daten aus der Literatur. In *Lecture Notes in Informatics (LNI) – Proceedings, Series of the Gesellschaft für Informatik (GI).*

[7] Rojas, I., Ratsch, E., Saric, J., Wittig, U. (2004) Notes on the use of ontologies in the biochemical domain. *In Silico Biol.* **4**(1):89 – 96.

[8] Bodenreider, O., Stevens, R. (2006) Bio-ontologies: current trends and future directions. *Brief Bioinform.* **7**(3):256 – 74.
doi: http://dx.doi.org/10.1093/bib/bbl027.

[9] Rubin, D.L., Shah, N.H., Noy, N.F. (2008) Biomedical ontologies: a functional perspective. *Brief Bioinform.* **9**(1):75 – 90.
doi: http://dx.doi.org/10.1093/bib/bbm059.

[10] Sayers, E.W., Barrett, T., Benson, D.A., Bolton, E., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., DiCuccio, M., Federhen, S., Feolo, M., Fingerman, I.M., Geer, L.Y., Helmberg, W., Kapustin, Y., Landsman, D., Lipman, D.J., Lu, Z., Madden, T.L., Madej, T., Maglott, D.R., Marchler-Bauer, A., Miller, V., Mizrachi, I., Ostell, J., Panchenko, A., Phan, L., Pruitt, K.D., Schuler, G.D., Sequeira, E., Sherry, S.T., Shumway, M., Sirotkin, K., Slotta, D., Souvorov, A., Starchenko, G., Tatusova, T.A., Wagner, L., Wang, Y., Wilbur, W.J., Yaschenko, E., Ye, J. (2011) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **39:**D38 – 51.
doi: http://dx.doi.org/10.1093/nar/gkq1172.

[11] Gremse, M., Chang, A., Schomburg, I., Grote, A., Scheer, M., Ebeling, C., Schomburg, D. (2011) The BRENDA Tissue Ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources. *Nucleic Acids Res.* **39:**D 507 – 13. doi: http://dx.doi.org/10.1093/nar/gkq968.

[12] de Matos, P., Alcántara, R., Dekker, A., Ennis, M., Hastings, J., Haug, K., Spiteri, I., Turner, S., Steinbeck, C. (2010) Chemical Entities of Biological Interest: an update. *Nucleic Acids Res.* **38:**D 249 – 54. doi: http://dx.doi.org/10.1093/nar/gkp886.

[13] Scheer, M., Grote, A., Chang, A., Schomburg, I., Munaretto, C., Rother, M., Söhngen, C., Stelzer, M., Thiele, J., Schomburg, D. (2011) BRENDA, the enzyme information system in 2011. *Nucleic Acids Res.* **39:**D 670 – 6. doi: http://dx.doi.org/10.1093/nar/gkq1089.

[14] Courtot, M., Juty, N., Knüpfer, C., Waltemath, D., Zhukova, A., Dräger, A., Dumontier, M., Finney, A., Golebiewski, M., Hastings, J., Hoops, S., Keating, S., Kell, D.B., Kerrien, S., Lawson, J., Lister, A., Lu, J., Machne, R., Mendes, P., Pocock, M., Rodriguez, N., Villeger, A., Wilkinson, D.J., Wimalaratne, S., Laibe, C., Hucka, M., Le Novère, N. (2011) Controlled vocabularies and semantics in systems biology. *Mol. Syst. Biol.* **7:**543. doi: http://dx.doi.org/10.1038/msb.2011.77.

[15] International System of Units (SI): http://www.bipm.fr/en/si/

[16] Hucka, M., Finney, A., Sauro, H.M., Bolouri, H., Doyle, J.C., Kitano, H., Arkin, A.P., Bornstein, B.J., Bray, D., Cornish-Bowden, A. *et al.* (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* **19:**524 – 31. doi: http://dx.doi.org/10.1093/bioinformatics/btg015.

[17] Le Novère, N., Finney, A., Hucka, M., Bhalla, U.S., Campagne, F., Collado-Vides, J., Crampin, E.J., Halstead, M., Klipp, E., Mendes, P., Nielsen, P., Sauro, H., Shapiro, B., Snoep, J.L., Spence, H.D., Wanner, B.L. (2005) Minimum Information Required In the Annotation of Models (MIRIAM). *Nat. Biotechnol.* **23:**1509 – 15. doi: http://dx.doi.org/10.1038/nbt1156.

[18] Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M., Hirakawa, M. (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.* **38:**D 355 – 60. doi: http://dx.doi.org/10.1093/nar/gkp896.

[19] The Gene Ontology Consortium (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.* **25**(1):25 – 9. doi: http://dx.doi.org/10.1038/75556.

[20] McDonald, A.G., Boyce, S., Moss, G.P., Dixon, H.B., Tipton, K.F. (2007) ExplorEnz: a MySQL database of the IUBMB enzyme nomenclature. *BMC Biochem.* **8:**14.
doi: http://dx.doi.org/10.1186/1471-2091-8-14.

[21] Bairoch, A. (2000) The ENZYME database in 2000. *Nucleic Acids Res.* **28:**304 – 5.
doi: http://dx.doi.org/10.1093/nar/28.1.304.

[22] Matthews, L., Gopinath, G., Gillespie, M., Caudy, M., Croft, D., de Bono, B., Garapati, P., Hemish, J., Hermjakob, H., Jassal, B., Kanapin, A., Lewis, S., Mahajan, S., May, B., Schmidt, E., Vastrik, I., Wu, G., Birney, E., Stein, L., D'Eustachio, P. (2009) Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res.* **37:**D 619 – 22.
doi: http://dx.doi.org/10.1093/nar/gkn863.

[23] Fleischmann, A., Darsow, M., Degtyarenko, K., Fleischmann, W., Boyce, S., Axelsen, K.B., Bairoch, A., Schomburg, D., Tipton, K.F., Apweiler, R. (2004) IntEnz, the integrated relational enzyme database. *Nucleic Acids Res.* **32:**D 434 – 7.
doi: http://dx.doi.org/10.1093/nar/gkh119.