SABIO-RK: Integration and Curation of Reaction Kinetics Data

Ulrike Wittig, Martin Golebiewski, Renate Kania, Olga Krebs, Saqib Mir, Andreas Weidemann, Stefanie Anstein, Jasmin Saric, and Isabel Rojas

Scientific Databases and Visualization Group, EML Research gGmbH, Schloss-Wolfsbrunnenweg 33, 69118 Heidelberg, Germany
Ulrike.Wittig@eml-r.villa-bosch.de
http://sabiork.villa-bosch.de/

Abstract. Simulating networks of biochemical reactions require reliable kinetic data. In order to facilitate the access to such kinetic data we have developed SABIO-RK, a curated database with information about biochemical reactions and their kinetic properties. The data are manually extracted from literature and verified by curators, concerning standards, formats and controlled vocabularies. This process is supported by tools in a semi-automatic manner. SABIO-RK contains and merges information about reactions such as reactants and modifiers, organism, tissue and cellular location, as well as the kinetic properties of the reactions. The type of the kinetic mechanism, modes of inhibition or activation, and corresponding rate equations are presented together with their parameters and measured values, specifying the experimental conditions under which these were determined. Links to other databases enable the user to gather further information and to refer to the original publication. Information about reactions and their kinetic data can be exported to an SBML file, allowing users to employ the information as the basis for their simulation models.

1 Introduction

The biosciences have undergone some dramatic changes in the last few years. Novel lab approaches like high-throughput methods enable scientists to rapidly produce an enormous amount of data. For researchers this poses problems connected with retaining an overview of these data and accessing them. Thus one of the biggest challenges in biological science at present is to achieve data comparability and ease of access for the scientific community. To attain this goal, experimental data from different sources need to be standardized and integrated into databases.

At the moment only a small number of databases exist which contain information about biochemical reaction kinetics. The BRENDA enzyme database [1] offers a comprehensive list of kinetic parameters based on literature information. UniProt [2] started to include kinetic parameters as comments related to biophysicochemical properties, also manually extracted from publications. The BioModels database [3] stores published mathematical models of biological interest annotated and linked to relevant data resources (e.g. publications or databases). The models include kinetic

U. Leser, F. Naumann, and B. Eckman (Eds.): DILS 2006, LNBI 4075, pp. 94 – 103, 2006.

laws and their parameters represented in SBML (Systems Biology Mark-up Language) format [4] and can be used for simulations of biochemical reactions or networks.

In order to compare kinetic data and develop biochemical network models, kinetic parameters need to be consistently described and related to kinetic mechanisms, equations representing the kinetic laws and environmental conditions. The known mechanisms of biochemical reactions should be reflected in mathematical formulas, which have to be linked to the corresponding parameters, such as kinetic constants and concentrations of each reaction participant. As kinetic constants highly depend on environmental conditions, they only can be specified completely by describing these conditions used for determination. Data sets based on an experiment assayed under similar experimental conditions should be associated to each other to facilitate the comparison. Therefore, users interested in information about reaction kinetics require databases that merge and structure all these data.

The SABIO-RK (**S**ystem for the **A**nalysis of **Bio**chemical Pathways - **R**eaction **K**inetics) database is designed to meet these requirements and to support researchers interested in information about biochemical reactions and their kinetics. This report will mainly focus on the database content, integration and curation processes. Modelling of the database and the retrieval of data by database searching will be briefly discussed.

2 Data Integration

SABIO-RK represents an extension of the SABIO biochemical pathway database also developed at EML Research [5]. Figure 1 represents a simplified schema of the main database objects and their relations in SABIO and SABIO-RK. SABIO contains information about biochemical pathways, reactions and their participants (enzymes, reactants etc.). These data are connected with specific protein information, organisms or cellular locations. SABIO-RK combines the general data about biochemical reactions stored in SABIO with information about their kinetic properties. The type of kinetic law and its representation in a formula is given if provided in the literature. This also includes effectors (e.g. cofactors, activators or inhibitors) of the reactions and their type of interaction (e.g. competitive or non-competitive inhibition). The kinetic laws are represented with their parameters, including their measured values. Since many of the publications only contain kinetic constants (e.g. Km, kcat or Vmax) but have no description of the kinetic law type, these parameter values are also inserted independent from a kinetic law type.

Additionally the database contains descriptions of the experimental conditions (e.g. pH, temperature, and buffer) for the measured parameter values. In the buffer description all components of the assay are represented including coupled enzyme assays.

In order to establish a broad information basis, data from different sources are integrated into SABIO-RK (Figure 1). Most of the reactions, their associations with biochemical pathways and their enzymatic classifications (EC classifications [6]) are downloaded from the KEGG (Kyoto Encyclopedia of Genes and Genomes) database [7] and stored in SABIO. In contrast, the kinetic data contained in SABIO-RK are

manually extracted from scientific articles and verified by curators. At the moment it is very difficult to extract this information automatically, such as by the use of text mining technologies, given that most of the data are highly scattered through various publications and are frequently found in tables, formulas or graphs. However, we are working on the development of support tools, one of them for the identification of synonyms of chemical compound names, as we will describe in section 3.

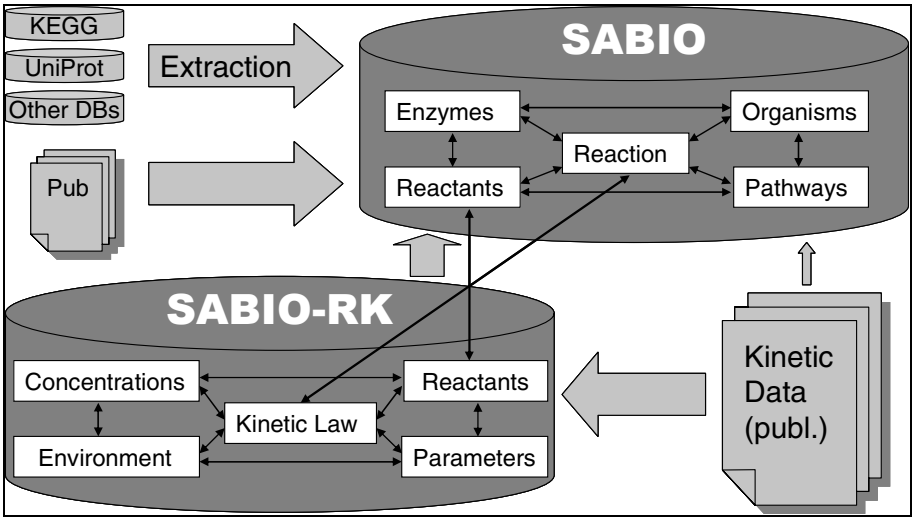


Fig. 1. Population, content and schematic relation of SABIO and SABIO-RK. SABIO contains general information about biochemical pathways and reactions in different organisms, including details about corresponding enzymes and reactants. Most of these data are collected from other databases like KEGG and UniProt. SABIO-RK extends SABIO by storing information about the reactions' kinetic properties, such as the kinetic laws with their corresponding parameters and environmental conditions under which they were determined.

As standards for publishing data of biochemical reactions and reaction kinetics are lacking, the curators are faced with problems like synonymic or aberrant notations of compounds and enzymes, multiplicity of parameter units and missing information about assay procedures and experimental conditions. During the curation process, we unify and structure the data consistently in order to facilitate the comparison of the kinetic data obtained under different experimental conditions or from different organisms, tissues etc. Furthermore structured data enable the user to understand the behaviour of a biochemical reaction under environmental changes like for example increase of temperature or pH variations.

The information source of each database entry is clearly shown and linked to the PubMed [9] database in order to allow the user to refer to the original paper to obtain additional information about the experiment described.

Systematic names of organisms named in the publication are identified according to the NCBI taxonomy [8] and additional information about specific strains of organisms is stored in the database. If the enzyme of the original organism was

expressed in another organism, the host organism is represented in the general comment line of an entry.

The extraction work is done by students using a web-based interface to enter the extracted information into a temporary SQL database. A list of publications, expected to contain kinetic data was obtained by keyword searches in the PubMed database and is used as the basis for data extraction. Before transferring the data to the final database SABIO-RK, they are checked, complemented and verified by a team of biological experts to eliminate possible errors and inconsistencies.

As of May 2006, data from more than 1400 publications were evaluated from which 820 were found to contain useful kinetic data that were extracted and inserted into the intermediate database. About 65% are already curated and inserted into the SABIO-RK database. From one publication more than one database entry can arise if different reactions, enzymes, environmental conditions etc. are connected to measured parameter values. Currently SABIO-RK contains about 5100 curated single database entries referring to about 190 organisms, 1100 different reactions, and 320 enzymes catalysing these reactions. Each database entry comprises at least one kinetic parameter. Since there are many publications processed where no information about the kinetic law is addressed, currently in SABIO-RK only 30% of all entries are related to a kinetic law formula. Of the about 20.000 chemical compound names included in the database, 13.470 refer to different compounds, i.e. the average number of alternative names per compound is 1.5.

SABIO-RK only refers to the original source of kinetic data compared parameter values extracted from a referenced paper are not linked to this publication. To avoid redundancies, copying of errors and linking to disparate experimental conditions, the original source of the referenced values is included in the database as separate entries. However, a comparison of the parameter values is possible since entries from different sources are linked by the same reaction or enzyme, assuming that the experimental conditions are the same.

Links to the UniProt database enables the user to gather further information about proteins corresponding to the enzymes.

3 Curation Process

The curation of extracted data is used to achieve correctness and consistency within the database. Already existing standards for data formats are applied as well as new standards are defined if necessary. For example, the unification of parameter units or chemical compound names involves existing standards as the SI system for unit notation or the nomenclature recommendations for chemical compounds of the International Union of Pure and Applied Chemistry (IUPAC). In contrast to, for enzyme specifications (mutants, isoforms, etc.) database-internal norms are assigned additionally to the enzyme classification (EC) system of the International Union of Biochemistry and Molecular Biology (IUBMB). Already existing controlled vocabularies are used for the representation of organisms [8], tissues [1], cell locations [1] etc.

During the curation process, most of the data are unified and structured, with the exception of some information which is stored as comment lines or descriptions, such

as in the case of the buffers' compositions. The description of a buffer can be very complex containing for example information about coupled enzyme reactions and synthetic or labelled derivatives of physiological compounds. Therefore, currently information about the buffer composition is stored as a free text. Additional comment lines also contain information about host organisms in which proteins are expressed (e.g. recombinant enzymes expressed in Escherichia coli), or information about the enzyme proteins, especially the protein name if no EC classification is known.

The fact that chemical compounds often have multiple alternative names complicates the work of the database curators. They need to find out whether a compound described in the publication is already contained in the database, possibly with synonymic names, or it is necessary to include this new compound in the database. To address this problem we have developed a tool for the linguistic analysis of chemical terminology, more precisely the names of organic compounds, named CHEMorph [10]. CHEMorph analyses systematic and semi-systematic names, class terms, and also otherwise underspecified names, by using a morpho-syntactic grammar developed in accordance with IUPAC nomenclature [11]. It yields an intermediate semantic representation of a compound which describes the information encoded in a name. The tool provides SMILES strings [12] for the mapping of names to their molecular structure and also classifies the terms analysed. The general process together with an example analysis is shown in Figure 2. The systematic compound name *7-hydroxyheptan-2-one* is transformed into a semantic representation which describes the following: The compound *compd* with its three *parameters* in parentheses, which are the formal descriptions of (i) the skeleton structure, (ii) the name's prefix, and (iii) the name's suffix. From this semantic expression, the corresponding SMILES string [CC(=O)CCCCCO] and the class list is calculated. Currently by matching the yielded SMILES strings, CHEMorph can be used to identify synonymous compound names in order to check if a compound is already contained in SABIO-RK. Future developments will include the matching of chemical structures generated from SMILES strings. With the help of additional Natural Language Processing (NLP) methods, the existing compounds in SABIO-RK can be analyzed for wrong synonyms and multiple entries. By this, the completeness and consistency of the compound data can still be improved.

Also organisms often have synonymic names. They can be described in the literature by their common or systematic name. The SABIO-RK database refers to the NCBI taxonomy and uses systematic names to be able to compare data. Since some authors only give the common name of an organism the curator has to deduce the systematic name. For example the organism described as "rat" can be transferred to Rattus sp. or Rattus norvegicus where the latter is mainly used in laboratories. The curators have to decide if the general organism name is used or not.

Units of kinetic parameters and concentrations can be written in different ways and often have multiple scales. Different systems of standardisation exist in parallel, for example enzyme activities can be noted in *katal* (*mol/s*), *international units* (*µmol/min*), *mg/min* or similar units. This makes the comparison of the data quiet difficult. Therefore a list of scaled and standardized units was established within SABIO-RK based on the recommendations of the International System of Units (SI) [13]. All parameter units and concentrations stored refer to a list that relates synonymic notations with the correct SI standards.

For consistency and to avoid duplicate entries, lists of compounds, organisms, tissues, compartments, kinetic law types and parameter units already existing in the SABIO-RK database are provided for selection at the input interface. These lists also contain synonyms referring to the same content to enable the search for alternative names of compounds, tissues etc. These may mean that the information presented to the database user is not exactly that included in the paper because the entries are presented with recommended names, however the user can always obtain the synonyms of the entries with multiple names. Already existing reactions can be searched in the database by defining one or two reaction participants.

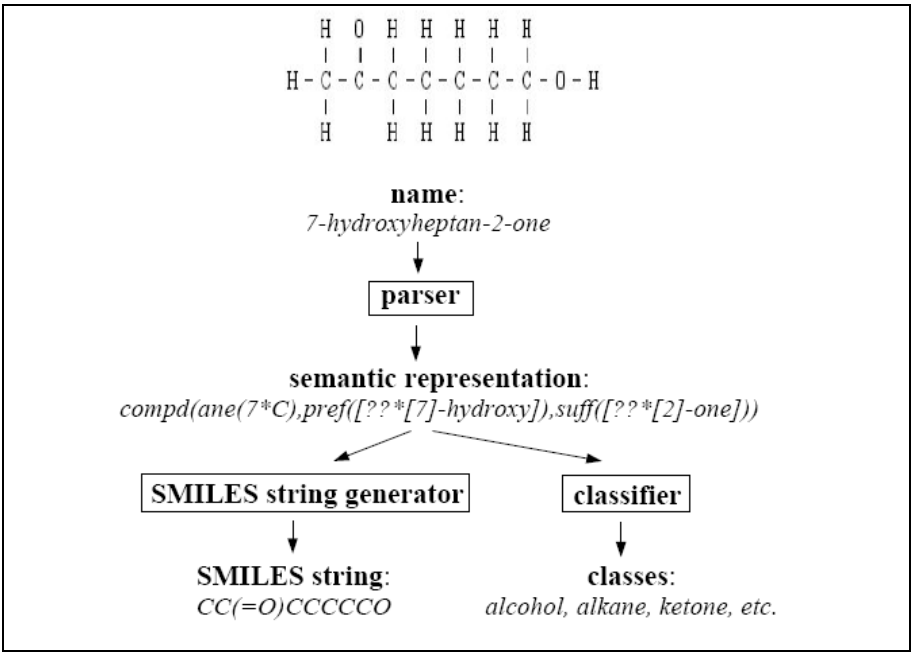


Fig. 2. Overview of the CHEMorph system to support manual database curation. A chemical compound name is parsed and gets a semantic representation assigned, which is taken as a basis to calculate a SMILES string and the classes the compound belongs to.

Enzymes variants catalysing the reactions are distinguished by a description of their subform, like wildtype or mutant protein species. They are named as *wildtype* or *mutant* followed by their name. Different isoforms of an enzyme are also named as *wildtype* followed by the name or abbreviation of the isoform. Furthermore, by including the specification of one or more accession number(s) of the UniProt database (if available), a direct link is provided to the properties of the enzyme.

Additionally, the curators are confronted with missing or only partial information in the literature. For example a reaction definition can be incomplete supposing that only substrates of reactions are named without a definition of the reaction products. Knowing the chemical mechanism of the enzymatic reaction an equation could be

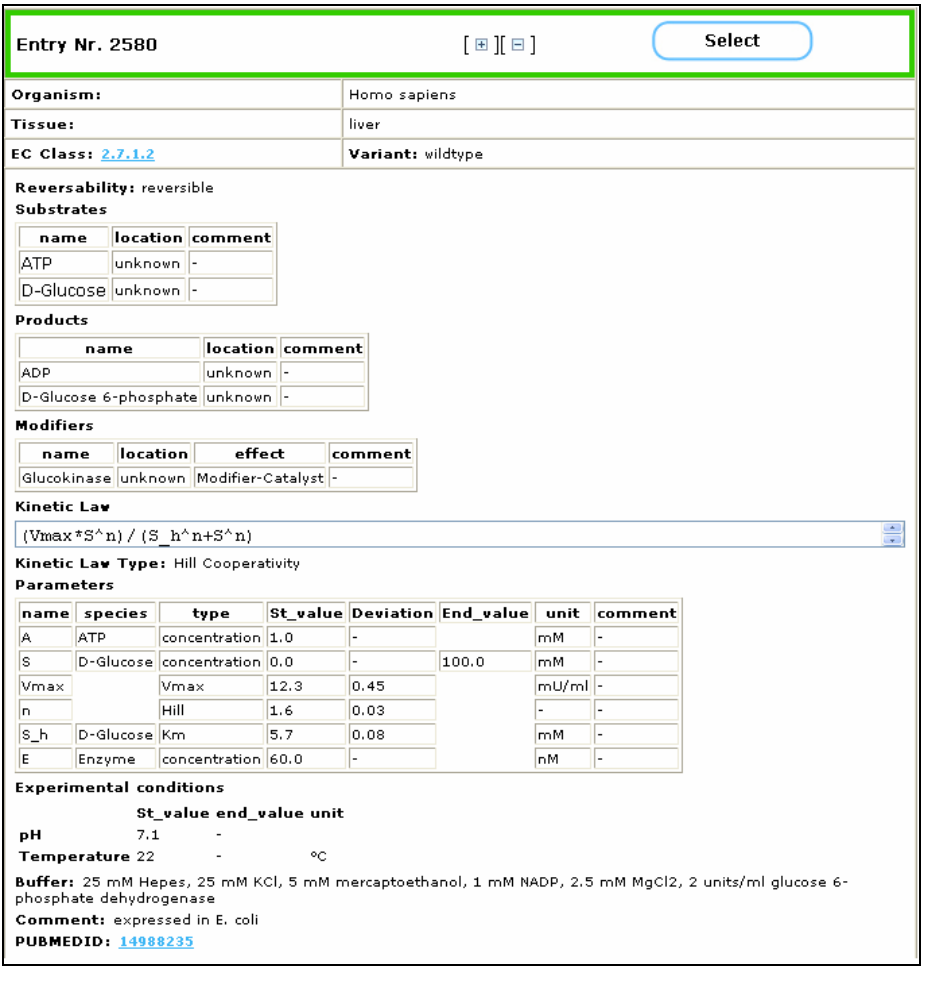
completed manually by biochemical experts, but this work could be very time-consuming and furthermore the result could be imprecise. Therefore general compound classes representing specific chemical properties are used as reaction participants. A tool developed for the SABIO biochemical pathway database allows for the classification of chemical compounds based on their functional groups using SMILES strings [14]. Different levels of compound classes based on this compound classification system can deduce more information about the unknown products of a reaction. For compounds for which a SMILES string cannot be assigned, e.g. underspecified or class names such as *deoxysugar* CHEMorph can be used for a classification based on the name.

Sometimes, publications contain incomplete datasets, i.e. not all parameters are measured or initial concentrations of reaction participants (reactants, effectors or enzymes) are missing. For these cases, SABIO-RK contains all parameters or concentrations required for a complete kinetic formula, independent of the existence of the corresponding values. Missing values are left blank or represented by null values. In this way, when exporting the information in SBML the user maintains a reference to all parameters, with or without values.

One major point is that the database only contains information that is mentioned in the corresponding paper. There is neither any interpretation of data by the biological experts, nor the addition of further information. For example, if the authors describe the kinetic mechanism of the reaction as competitive inhibition and no explicit formula is given in the publication, the SABIO-RK database will not show a kinetic formula but the kinetic law type named *Competitive inhibition*.

4 Search and Retrieval

The web-based user interface of SABIO-RK (Figure 3) enables the user to search for reactions and their kinetics by specifying characteristics of the reaction. These characteristics may include biochemical pathways in which the reaction participates, reactants of the reaction (substrates and products), classification of the enzymes catalysing the reaction and organisms, tissues or cellular locations in which the reaction takes place. The search for kinetic data can be further specified by the experimental conditions used for their determination (currently only pH value and temperature), which are considered solely for the retrieval of kinetic data. The system retrieves all entries satisfying the given criteria and indicates whether kinetic information for the associated reaction under the search criteria is specified (organism, tissue, cellular location and experimental conditions). Apart from this, the system also indicates whether there are kinetic data available for the enzymes catalysing each reaction. This approach has been selected to support the variations in the definition of the reactions composing a pathway, e.g. where a reaction can be substituted for by a very similar reaction with a slight change in the reactants. The next version of the interface will also enable the user to search for networks or paths of reactions between two compounds or enzymes. The kinetic data can then be viewed and selected for export in SBML format. Reactions with no kinetic data can also be included in the SBML file.



Entry Nr. 2580 [⊞][⊟] Select

Organism:	Homo sapiens
Tissue:	liver
EC Class: 2.7.1.2	**Variant:** wildtype

Reversability: reversible

Substrates

name	location	comment
ATP	unknown	-
D-Glucose	unknown	-

Products

name	location	comment
ADP	unknown	-
D-Glucose 6-phosphate	unknown	-

Modifiers

name	location	effect	comment
Glucokinase	unknown	Modifier-Catalyst	-

Kinetic Law

(Vmax*S^n)/(S_h^n+S^n)

Kinetic Law Type: Hill Cooperativity

Parameters

name	species	type	St_value	Deviation	End_value	unit	comment
A	ATP	concentration	1.0	-		mM	-
S	D-Glucose	concentration	0.0	-	100.0	mM	-
Vmax		Vmax	12.3	0.45		mU/ml	-
n		Hill	1.6	0.03		-	-
S_h	D-Glucose	Km	5.7	0.08		mM	-
E	Enzyme	concentration	60.0	-		nM	-

Experimental conditions

	St_value	end_value	unit
pH	7.1	-	
Temperature	22	-	°C

Buffer: 25 mM Hepes, 25 mM KCl, 5 mM mercaptoethanol, 1 mM NADP, 2.5 mM MgCl2, 2 units/ml glucose 6-phosphate dehydrogenase

Comment: expressed in E. coli

PUBMEDID: 14988235

Fig. 3. SABIO-RK database entry. An example data set represents a specified reaction including kinetic data, experimental conditions and additional information extracted from a publication.

5 Summary

The SABIO-RK database has been designed to meet the Systems Biology community requirements. It aims to support modelers with high quality data in setting-up in-silico models describing biochemical reaction networks. The database enables complex searches for reactions, parameters, etc. and uses existing or defines new standards for data formats. Selected kinetic data can be exported in SBML format to build models for the simulation of complex biochemical processes. SABIO-RK also bundles information for researchers interested in comparing reaction kinetic data originating from different sources.

6 Future Directions

In future, not only kinetic data from published literature will be inserted into the database but also data directly entered by scientists doing the lab experiments. Thus, all the needed information can be given by the experimenters and no information is lost. In doing so, users would be able to directly compare their own experimental results in SABIO-RK with kinetic data extracted from literature. Furthermore detailed descriptions of the kinetic reaction mechanism will be included in the near future to give the opportunity to represent kinetic properties of sub-reactions or binding mechanisms of enzymes in the database. Finally, data export functions of the user interface will be expanded, since a lot of the information stored in SABIO-RK can not yet be formally described in SBML.

Currently SABIO-RK contains mainly metabolic reactions we aim in the near future to incorporate more signalling reactions. Here the difficulty lies in the representation of the signalling reactions, i.e. multiplicity of states of a compound and general descriptions of compounds or compound families.

In order to improve the support to modeller, we are working on a visual display of the networks been set-up by the users, using this platform for present different types of information, such as the existence or not of kinetic information under certain experimental conditions.

Acknowledgement

The project is funded by the Klaus Tschira Foundation and partially by the German Research Council (BMBF). We would also like to thank the members of the Bioinformatics and Computational Biochemistry and the Molecular and Cellular Modelling Groups of EML Research for their helpful discussions and comments. Last but not least, we thank all the student helpers, who have contributed to the population of the database.

References

1. Schomburg I, Chang A, Ebeling C, Gremse M, Heldt C, Huhn G, Schomburg D (2004) BRENDA, the enzyme database: updates and major new developments. *Nucleic Acids Res*, 32, D431-3
2. Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res*, 33, D154-D159
3. Le Novere N, Bornstein B, Broicher A, Courtot M, Donizelli M, Dharuri H, Li L, Sauro H, Schilstra M, Shapiro B, Snoep JL, Hucka M (2006) BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. *Nucleic Acids Res*, 34, D689-91

4. Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, Kitano H, Arkin AP, Bornstein BJ, Bray D, Cornish-Bowden A, Cuellar AA, Dronov S, Gilles ED, Ginkel M, Gor V, Goryanin II, Hedley WJ, Hodgman TC, Hofmeyr JH, Hunter PJ, Juty NS, Kasberger JL, Kremling A, Kummer U, Le Novere N, Loew LM, Lucio D, Mendes P, Minch E, Mjolsness ED, Nakayama Y, Nelson MR, Nielsen PF, Sakurada T, Schaff JC, Shapiro BE, Shimizu TS, Spence HD, Stelling J, Takahashi K, Tomita M, Wagner J, Wang J (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19, 524-31
5. Rojas I, Bernardi L, Ratsch E, Kania R, Wittig U, Saric J (2002) A database system for the analysis of biochemical pathways. *In Silico Biol* 2,0007
6. IUBMB: http://www.chem.qmul.ac.uk/iubmb/enzyme/
7. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res*, 34, D354-7
8. NCBI Taxonomy: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Taxonomy
9. PubMed: http://www.pubmed.gov
10. Anstein S, Kremer G, Reyle U (2006) Identifying and Classifying Terms in the Life Sciences: The Case of Chemical Terminology. *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*. To appear
11. IUPAC: http://www.chem.qmul.ac.uk/iupac/
12. Weininger D (1988) SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J Chem Inf Comput Sci*, 28, 31-36
13. International System of Units (SI): http://www.bipm.fr/en/si/
14. Wittig U, Weidemann A, Kania R, Peiss C, Rojas I (2004) Classification of chemical compounds to support complex queries in a pathway database. *Comp Funct Genom,* 5, 156-62