SABIO-RK: A data warehouse for biochemical reactions and their kinetics.

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

Scientific Databases and Visualisation Group, EML Research gGmbH,
Schloss-Wolfsbrunnenweg 33, 69118 Heidelberg, Germany

Abstract

Systems biology is an emerging field that aims at obtaining a system-level understanding of biological processes. The modelling and simulation of networks of biochemical reactions have great and promising application potential but require reliable kinetic data. In order to support the systems biology community with such data we have developed SABIO-RK (System for the Analysis of Biochemical Pathways - Reaction Kinetics), a curated database with information about biochemical reactions and their kinetic properties, which allows researchers to obtain and compare kinetic data and to integrate them into models of biochemical networks. SABIO-RK is freely available for academic use at http://sabio.villa-bosch.de/SABIORK/.

1 Introduction

Systems biology deals with the analysis and prediction of the dynamic behaviour of biological networks through mathematical modelling based on experimental data [1].

It focuses on the connections and interactions of the components in the cell and in general in the organism, all as part of one system. The modelling and simulation of biochemical reaction networks requires consistent information about the kinetics of the reactions involved, such as the kinetic laws mathematically describing the dynamics of the reactions together with their respective parameters. Since kinetic parameters highly depend on the environmental conditions under which they where determined, descriptions of the kinetic parameters should be given together with the experimental conditions used for their determination as well as the biological sources (for example organism, tissue, cellular location) of the enzymes and other reaction components.

In the process of setting up models of biochemical networks, researchers need to compare and integrate information from many different sources, such as various experiments, additional information presented in publications and data provided by different databases. Not only they are confronted with the problem of collecting this information scattered through various resources, but also they need to overcome the problem of integrating data described in many different formats. Furthermore, experiments are performed using different technologies and terminologies, plus each special field uses its own vocabulary and concepts. One key challenge in systems biology is therefore to provide mechanisms to collect and integrate the necessary data to be able to compare them and use these data for the setting-up of biochemical reaction network models.

* Corresponding author, Olga.Krebs@eml-r.villa-bosch.de

To date there are a small number of databases available containing kinetic data of biochemical reactions; here we will mention a few of them. BRENDA [2] is a comprehensive database providing information about enzymes. Enzyme entries contain information about the reactions catalysed by the given enzyme including data describing their kinetics. Kinetikon provides detailed knowledge about biochemical reaction kinetics. However, its content is currently very limited, concentrating on reactions in yeast. The KDBI database (Kinetic Data of Biomolecular Interactions) [3] is a collection of experimentally determined kinetic data of binding or interaction events described in the literature, like protein-protein, protein-RNA, protein-DNA, protein-ligand, RNA-ligand, and DNA-ligand binding.

Although there are several databases like the above mentioned offering experimentally obtained kinetic data, they hardly offer a detailed description of the kinetics for single reactions comprising the kinetic law for the reaction rate with its parameters, as well as the corresponding environmental and experimental conditions.

The supported ways to query the databases for kinetic information are also rather limited. Apart from this, data export mechanisms and remote access facilities are also lacking. For example only Kinetikon offers export facilities of the data in SBML (Systems Biology Mark-up Language) [4], a commonly used data exchange format. Moreover, to the best of our knowledge none of the above mentioned databases offers programming interfaces, such as web-services that allow direct querying to the database from external applications. This last point is very important in order to be able to include resources containing kinetic data about biochemical reactions within workflows for the generation of biochemical network models. For instance, using web-services simulation tools for biochemical network models can automatically obtain kinetic data for certain reactions by querying the corresponding databases containing these data.

Apart from the databases containing information about the kinetics of single reactions based on experiments there is another group of databases containing published simulation models of biochemical reaction networks. This group includes BioModels [5], JWS [6], and DOQCS [7]. The models are mostly annotated (mainly in BioModels) and linked to the original sources and other databases with related data. The models presented in these databases contain detailed information about the reaction rate laws and parameters, however they frequently lack documentation on the environmental conditions under which the models hold. In many cases this is due to the lack of this particular information in publications. Ideally researchers should be able to compare and combine data from databases containing experimentally obtained data with data from model database. This can be accomplished by annotating the data to common ontologies, controlled vocabularies, or external databases.

SABIO-RK is an information system created to store, manage and distribute data about biochemical reactions and their kinetic properties. It allows researchers to obtain and compare kinetic data and to integrate them into models of biochemical networks. The main goal of SABIO-RK is the effective integration of curated (by biological experts), annotated (by the use of controlled vocabularies), consistent (to avoid redundant information) and comprehensive data from multiple data sources. In this way we provide users with high quality information about reaction kinetics, and allow easy access and efficient analysis of these data.

In this paper we will describe the main characteristics of SABIO-RK and its architecture. The conceptual data model and the database content will be discussed. The system architecture will be introduced including the design and implementation of the major system components.

2 Data in SABIO-RK

The knowledge base of SABIO-RK is composed of two tightly integrated concept-based components: **SABIO**, representing data relevant to biochemical reactions and pathways, and **Reaction Kinetics** representing the kinetic data obtained from experimental assays (see Figure1).

The **SABIO** component (published in [8]) contains detailed description of reaction equations, with their participating molecules (including information like systematic names and synonyms, chemical structure and provider identifiers for small molecules and proteins), an organism taxonomy, tissue descriptions and cellular locations. The database schema also supports the storage of information about proteins and genes (this information is however not yet used as search criteria in the SABIO-RK interface, see section 3). Most of the data are obtained by integrating data from several electronically available data sources. References to other databases enable the user to gather further information and to refer to the origin of data.

Most of the reactions, their associations with pathways, organisms and enzymes were downloaded from the KEGG database [9], and protein and gene annotation from Swiss-Prot [10] and from the organism-specific databases like GDB for human or SGD for yeast.

The data from external sources have not been altered but they have been enriched by information from manual annotation, consistently structured, and curated to eliminate redundancy. The curation process also includes the unification and standardisation of the data. Already existing standards for data formats are applied as well as new standards are defined if necessary. For example, most chemical compounds are known by a variety of names and authors frequently use trivial names.

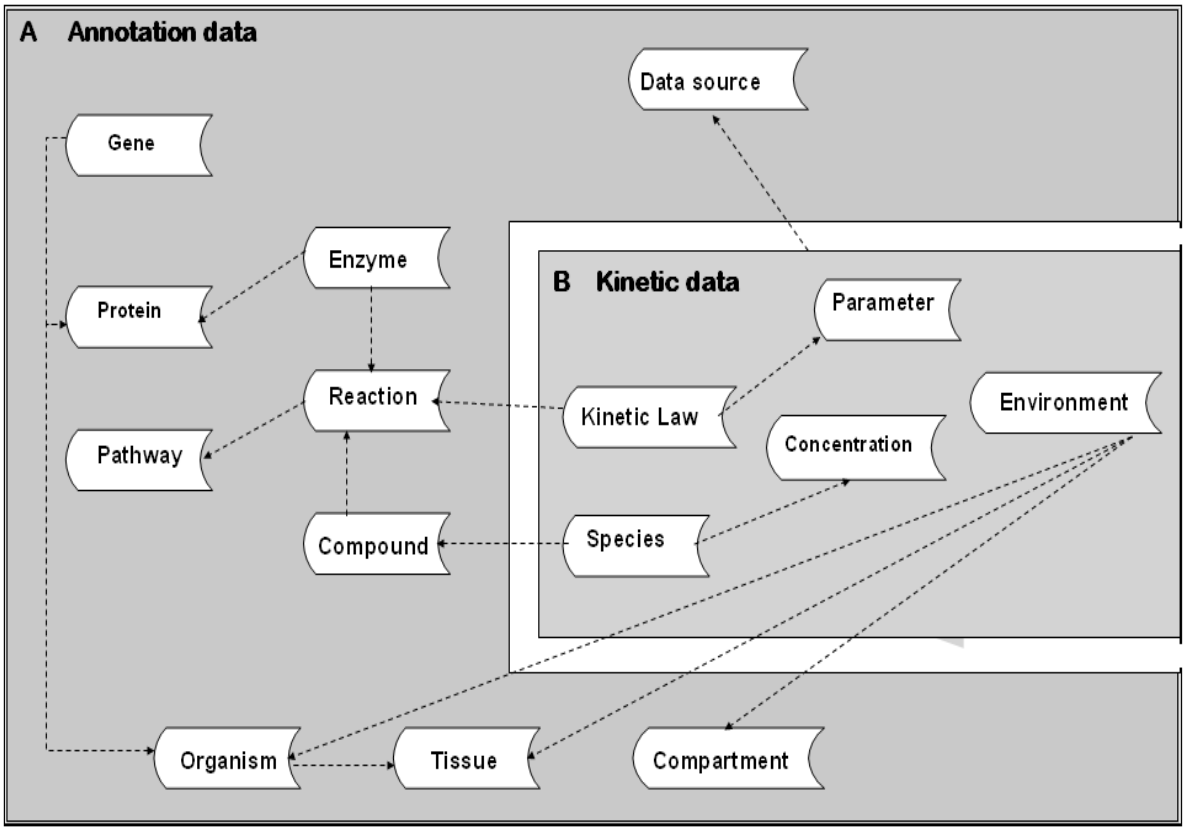

Figure 1: Conceptual model of SABIO-RK (simplified schema).

The curator has to check whether the 'new' compound, which was identified in a paper, is already stored within the SABIO-RK database. To support the curation process chemical compound names are automatically analyzed and identified. These comprise orthographic variance, syntactic and morphological variance (slightly diverging from standard IUPAC nomenclature) and combinations thereof.

To overcome this terminological hurdle we have developed two methods for chemical compound identification (a publication of this work is in preparation), i.e. normalisation of chemical compound names and calculation of the structure from chemical compound names. The normalisation component is able to produce the most common orthographic, morphological and syntactic variances. We use all variants for identifying possible candidates for the compound identification in the database. For the calculation of structure from chemical compound names we have developed a tool called CHEMorph [11]. 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. 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.

Apart from having alternative names for compounds we also maintain alternative, synonymic and recommended names for all other named entities, like pathways, organisms, tissues, etc. For enzyme specifications, in addition to the enzyme classification system of the International Union of Biochemistry and Molecular Biology (IUBMB) database, internal norms for protein specification are assigned. This comprises the differentiation between "wildtype" and "mutant" proteins followed by specification like for example "wildtype phosphorylated", "wildtype isozyme II" or "mutant K212Q". Existing controlled vocabularies are used for the representation of organisms, tissues, cellular locations, etc.

The Reaction Kinetics component of the database extends and supplements the **SABIO** content by storing highly interrelated information about biochemical reactions and their kinetics, mainly experimentally obtained. It includes reactants and modifying compounds (i.e. inhibitors or activators) of reactions, information about the catalyzing enzymes, and the kinetic laws governing the reactions.

The database provides information about the mathematical formulas describing the rates of reactions together with their corresponding parameters, such as kinetic constants and concentrations of each reaction participant. The experimental conditions under which the parameters were determined, such as pH, temperature and buffers are also included, as well as organism, tissue and cellular location where the reactions take place.

The database allows the storage of multiple kinetic descriptions for the same reaction, according to factors such as the organism in which they were measured (or for which they were estimated), or the environmental conditions under which the experiments were done.

In the process of population of the database with kinetic data we are confronted with many problems related to the integration and verification of the data, such as the unambiguous identification of reaction components (as mentioned above), missing specifications of experimental conditions and multiplicity of parameter units and mathematical equations. To address some of these problems we apply specific controlled vocabularies and link the data to biological ontologies. Controlled vocabularies are used (i) for the roles of reaction participants (e.g. 'substrate', 'catalyst, inhibitor', 'product', etc), (ii) for parameter types ('Vmax', 'Km', 'kcat', etc), (iii) for a taxonomy of kinetic rate equations ('Michaelis-Menten', 'Ordered Bi Bi', 'Competitive inhibition' etc). This procedure, in combination with the usage of merged synonymic notations and the annotation of compounds, enzymes and other components, facilitates the comparison and integration of the data.

The curators' work is also supported by some automatic routines to check the consistency of the entered data. For example, when a rate equation is entered, it is verified for the correct mathematical format. Moreover, each parameter in the kinetic law equation has to be defined in the list of parameters; for consistency this has to be done even if no value for this parameter is given in the paper. All units of parameter values are specified as scalable SI (International System of Units) units compatible with the unit requirements of the SBML specification. SABIO-RK currently (Nov: 206) contains 158 different biochemical pathways, 19450 chemical compounds, 7300 reactions, and 4225 different enzymes with corresponding proteins information. The reaction kinetics' data stored in SABIO-RK were manually extracted from 820 articles and are related to 1460 different biochemical reaction, 416 distinct EC numbers in 272 organisms. The stored parameters mainly describe steady-state kinetics for reactions related to metabolic pathways. Almost 8200 database entries (42% having a rate equation) describe about 6300 enzyme activities represented as rate constants, kcat or Vmax values and compound properties represented in 7500 Km (Michaelis constant) and 1550 Ki (inhibitor constant) values.

3 System architecture

The system architecture and data flow are summarized in Figure 2. Data acquired from external databases in different formats (e.g. flat files, database dumps, Excel tables, etc) first enters the **data integration layer**, where the data source wrappers and parsers extract the required information from source files and integrate them by mapping into the database schema. Data from structured flat files are mainly integrated into the database using an Oracle tool called SQL Loader. Regular updates of external data from relational databases (like ChEBI [13]) are performed using Oracle import utilities.

The introduction of the data manually extracted from literature into the database is supported by a customised web-based input interface. This interface implements methods to semi-automatically curate the introduced data, and supports curators in the verification, standarisation and corrections of the data.

The core of the system, the **data storage layer**, is a relational database implemented in Oracle 10g. This layer offers the storage, indexing and management of the data related to all aspects relevant to a complete representation of the data. This layer is also responsible for the data access control, query processing and for metadata management.

Data transformation, cleaning and consistency checking is performed using common database technologies like triggers, PL SQL procedures and functions.

A MySQL database is used by the input interface for the temporary storage of kinetic data extracted from literature sources for the curation by experts.

The **application layer** is the middle tier which bridges the user interfaces and the underlying database, hiding technical details from the end user. Apache Tomcat is used to deploy the SABIO-RK web application. The web pages are created mainly using JSP (JavaServer Pages) and JasvaScript technology. The business logic is defined by Java Beans at two levels, one actually describing the concepts and operations over the concepts represented in the database and a top-one directly supporting the search facilities offered by the user interface. Communication to the database is carried out by the concept classes accessing the Oracle database via the JDBC protocol.

Complementary to the SABIO-RK web application Oracle Application Express is used to create forms and reports to support the process of data curation, revision and management.

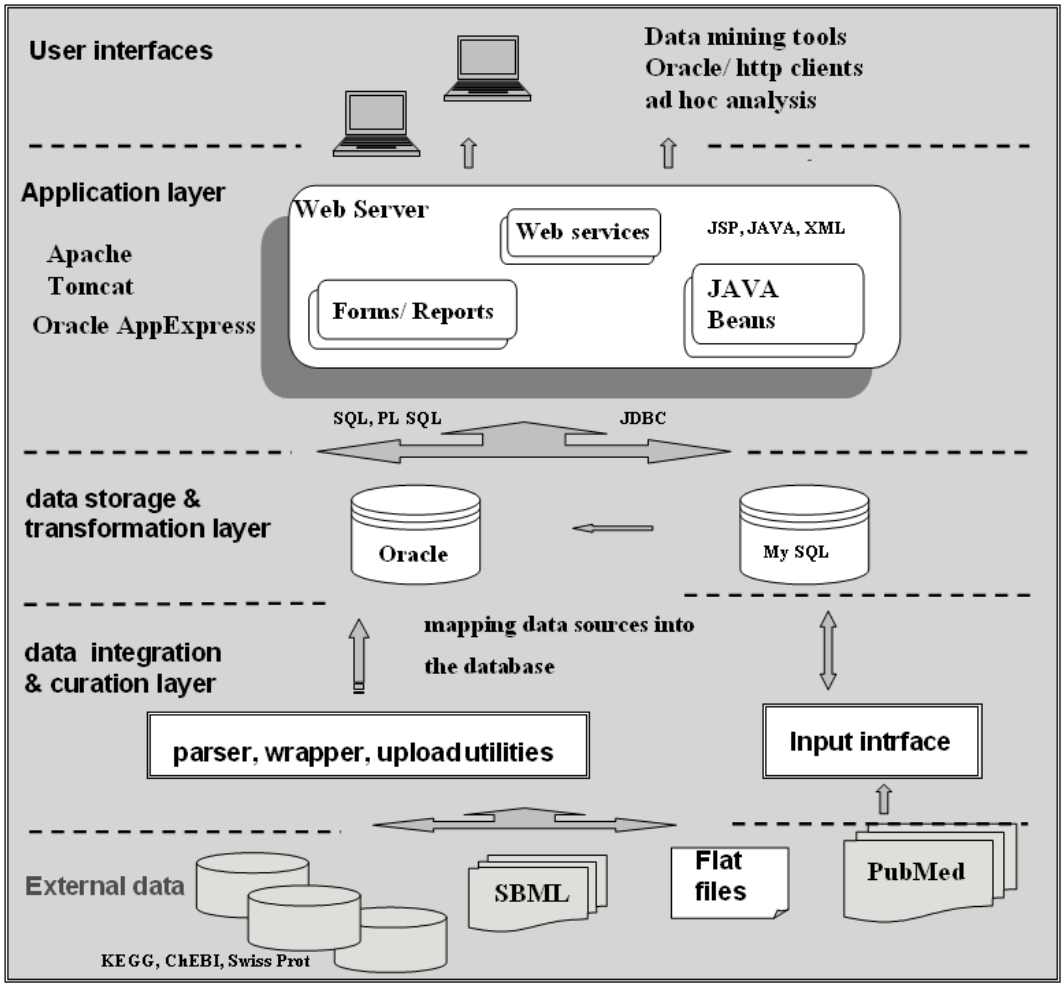

Figure 2. Detailed architecture of the system components and data workflow schema.

Additionally, the application layer includes web services to allow user's application to directly query the database without having to select the data manually through the web-interface. This aspect permits, for example, simulation tools to directly access the data stored in SABIO-RK for the creation of simulation models.

Users can access SABIO-RK by a web-based **interface** (Figure 3) that allows the search for biochemical reactions and their kinetics by specifying characteristics of the reactions of interest (such as reactants, enzymes or pathways) as well as the kinetic data searched (e.g. from a particular tissue, determined under certain experimental conditions or only certain parameter types). All reactions matching the search criteria are presented, allowing the view of further details about reactions, catalyzing enzymes and the kinetic data upon selection. Data about biochemical reactions and their kinetic parameters with their respective rate equations can be exported in SBML file format, thus allowing the import into simulation and modeling programs supporting SBML.

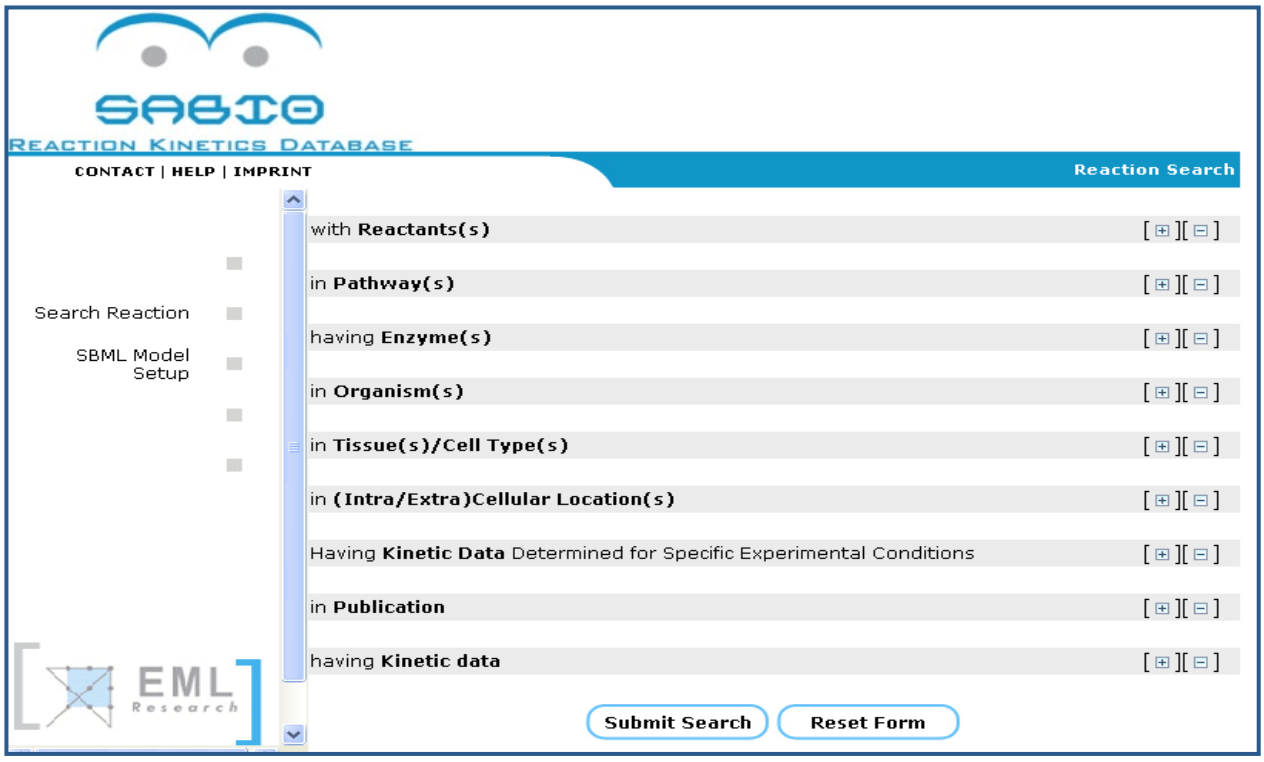

Figure 3. The SABIO-RK user interface allows the search for reactions and their kinetics by combining multiple criteria such as reactants, catalyzing enzymes, organisms, etc.

4 Summary and future directions

SABIO-RK has been mainly conceived as a system to support the analysis of biochemical reaction networks and their kinetics. The database stores data covering all aspects relevant to a complete representation of the biochemical pathways (currently mainly metabolic), and of biochemical reactions and their kinetics within the context of cellular locations, tissues and organisms. It contains information about reactants and effectors, catalyzing enzymes, and the kinetic laws for determining the rates of reactions. The type of the kinetics, modes of inhibition or activation, and corresponding equations are also shown together with their parameters and experimental conditions.

The system is accessible via a web-based interface which allows the querying, navigation and visualisation of the data. Although the main motivation for the creation of SABIO-RK was to provide a resource for modelers of biochemical networks to extract information about reactions and their kinetics, the database is also aimed at experimentalists needing information about assays for the determination of reactions kinetics and their results. The kinetics data is mainly extracted from literature sources and then revised and supplemented by a group of curators. Links to several external databases enable the user to gather further information about chemical compounds, enzymes and reactions, or to refer to the original publications. These links together with the use of synonymic notations for compounds and enzymes also facilitate the comparison of the data.

One of our main goals is to offer to researchers the possibility to directly introduce their experimental results into the database using a web-based input interface.

In the next database release an extension of the data model to store information about reaction mechanisms in terms of elementary steps of biochemical reactions will be included. Apart

from this we are working on the representation of reaction kinetics of signaling and regulatory reactions in the database.

5 Acknowledgements

SABIO-RK 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] Schilling M, T. Maiwald, S. Bohl, M. Kollmann, C. Kreutz, J. Timmer, and U. Klingmüller. Quantitative data generation for systems biology: the impact of randomisation, calibrators and normalisers. IEE Proceedings, Systems Biology, Vol. 152, Issue 4, p. 193-200, 2005

[2] Schomburg I, Chang A, Ebeling C, Gremse M, Heldt C, Huhn G, Schomburg D. BRENDA, the enzyme database: updates and major new developments. Nucleic Acids Res, 32, D 431-433, 2004

[3] Ji ZL, Chen X, Zhen CJ, Yao LX, Han LY, Yeo WK, Chung PC, Puy HS, Tay YT, Muhammad A, Chen YZ. KDBI: Kinetic Data of Bio-molecular Interactions database., Nucleic Acids Res. Jan 1;31(1):255-7, 2003

[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. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics,* 19, 524–531, 2003

[5] 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. BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. Nucleic Acids Res. Jan 1;34(Database issue):D689-91, 2006

[6] Olivier BG, Snoep JL. Web-based kinetic modelling using JWS Online. Bioinformatics. Sep 1;20(13):2143-4. Epub 2004 Apr 8, 2004

[7] Sudhir Sivakumaran , Sridhar Hariharaputran , Jyoti Mishra and Upinder S. Bhalla The Database of Quantitative Cellular Signaling: management and analysis of chemical kinetic models of signaling networks.Bioinformatics Vol. 19 no. 3, Pages 408-415, 2003

[8] Rojas I, Bernardi L, Ratsch E, Kania R, Wittig U, Saric J.A database system for the analysis of biochemical pathways. In Silico Biol 2,0007, 2002

[9] Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. Jan 1;28(1):27-30, 2000

[10] Bairoch A., Apweiler R. The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999.Nucleic Acids Res. January 1; 27(1): 49–54., 1999

[11] Kremer, Gerhard; Anstein, Stefanie; Reyle, Uwe Analysing and Classifying Names of Chemical Compounds with CHEMorph in Sophia. Ananiadou and Juliane Fluck, editors, Proceedings of the Second International Symposium on Semantic Mining in Biomedicine (SMBM 2006) pp. 37-43 JULIE Lab, Friedrich-Schiller-Universitat Jena, Germany, 2006

[12] Weininger D SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Comput Sci, 28, 31-36, 1988

[13] Brooksbank, C., Cameron, G., Thornton, J. The European Bioinformatics Institute's data resources: towards systems biology Nucleic Acids Res, 33, D46–D53, 2005.