the FEBS Journal

REVIEW ARTICLE

Challenges for an enzymatic reaction kinetics database

Ulrike Wittig, Maja Rey, Renate Kania, Meik Bittkowski, Lei Shi, Martin Golebiewski, Andreas Weidemann, Wolfgang Müller and Isabel Rojas

Scientific Databases and Visualization Group, Heidelberg Institute for Theoretical Studies (HITS), Germany

Keywords
biocuration; database; ontology; reaction kinetics; standardization

Correspondence
U. Wittig, Scientific Databases and Visualization Group, Heidelberg Institute for Theoretical Studies (HITS), Schloss-Wolfsbrunnenweg 35, 69118 Heidelberg, Germany
Fax: +49 6221 533298
Tel: +49 6221 533217
E-mail: ulrike.wittig@h-its.org

Note
This paper is dedicated to the memory of Isabel Rojas (1968–2013) who was our group leader and the founder of the SABIO-RK database.

(Received 31 May 2013, revised 27 September 2013, accepted 2 October 2013)

doi:10.1111/febs.12562

The scientific literature contains a tremendous amount of kinetic data describing the dynamic behaviour of biochemical reactions over time. These data are needed for computational modelling to create models of biochemical reaction networks and to obtain a better understanding of the processes in living cells. To extract the knowledge from the literature, biocurators are required to understand a paper and interpret the data. For modellers, as well as experimentalists, this process is very time consuming because the information is distributed across the publication and, in most cases, is insufficiently structured and often described without standard terminology. In recent years, biological databases for different data types have been developed. The advantages of these databases lie in their unified structure, searchability and the potential for augmented analysis by software, which supports the modelling process. We have developed the SABIO-RK database for biochemical reaction kinetics. In the present review, we describe the challenges for database developers and curators, beginning with an analysis of relevant publications up to the export of database information in a standardized format. The aim of the present review is to draw the experimentalist's attention to the problem (from a data integration point of view) of incompletely and imprecisely written publications. We describe how to lower the barrier to curators and improve this situation. At the same time, we are aware that curating experimental data takes time. There is a community concerned with making the task of publishing data with the proper structure and annotation to ontologies much easier. In this respect, we highlight some useful initiatives and tools.

Introduction

In the present review, we aim to draw attention to the view of database curators with respect to data contained in their publications focussing on enzymatic reaction kinetics. Experimentalists working in the laboratory are keen to publish their results in a scientific journal and to share their findings with the scientific community. However, do they consider whether and how these published data are reused or extracted from the publication for further research? Are they aware of the problems that other scientists have to deal with when reading the paper and extracting information? Of course, there are other experimentalists who are trying to repeat experiments or compare the results with their own data, and there are modellers who are trying to integrate published data into simulatable computer models. Both search for kinetic parameters and additional information in databases for enzymatic reaction kinetics and appreciate the structured,

Abbreviations
SBML, Systems Biology Markup Language; SBO, Systems Biology Ontology; SBPAX, Systems Biology Pathway Exchange; STRENDA, Standards for Reporting Enzymology Data.

coherent and searchable format of already published information, which is typically widely distributed in the literature. Experimentalists use such databases to obtain an overview of published data in their area of interest and for reference. Moreover, the kinetic parameters, together with their assay conditions, help in the design of new experimental set-ups for related studies. Modellers use such databases to extract bundled information in standardized formats and integrate data from different sources into their models.

For dynamic modelling and simulation of biochemical reactions and complex networks, computational methods are used that either describe the reaction dynamics as an approximate estimation applying convenience kinetics [1] or require detailed information on the reactions and their kinetics. This required information includes kinetic parameters with their rate equations that describe the dynamic behaviour of the reactions over time, as well as detailed descriptions of how these were determined.

The kinetic data needed to create a model can have different sources: experimental results from personal experiments or collaboration partners, data obtained from other models or fitting procedures, as well as data from scientific publications. Up to now, the main sources are scientific publications. Although standardization efforts already exist, as well as standard identifiers, a wide range of information is nonstandardized or implicit in publications and has to be reconstructed. In particular, information about which equations were used for the determination of kinetic parameters is rarely given. Additionally, detection of the relevant information and extracting this from publications is very cumbersome and time consuming, especially because a structured and standardized format for the description of the data and its context is missing and corresponding information is scattered over the whole publication. Also, there is almost no use of controlled vocabularies, annotations to ontology terms or unique database identifiers for the described data and the referenced entities (U. Wittig, unpublished data).

In recent years, the number of available biological databases has grown, offering the advantage of providing unified structures. Included amongst them are databases that contain enzyme and reaction kinetics data to support the modelling and simulation processes: enzyme databases such as BRENDA [2] or protein databases such as UniProtKB [3], which both also store kinetic parameters, as well as databases storing complete models with their parameters, including BioModels [4], JWS Online [5] and DOQCS [6]. When we started to develop SABIO-RK [7] in 2005, there was no database that stored kinetic parameters for single reactions together with their corresponding kinetic rate equation, as well as the experimental conditions under which the kinetic parameters were estimated. Model databases actually store equations with their parameters, although only in the context of the corresponding model, usually obtained from quite generic fitting procedures that fit the whole model to experimental results. This makes it difficult to reuse these data and integrate them into other models. The modellers' experience was that they could use the kinetic parameters from databases but, additionally, they always had to read the details in the publication to extract the rate laws (if available at all) and check the constraints. It is also important to know the experimental conditions under which the kinetic parameters were estimated to evaluate the portability of the kinetic parameters and to understand the role of the enzyme within the reaction process. However, to relate the assay conditions to the kinetic parameters and rate laws, modellers also had to extract the relevant information directly from the publications because no database stored all this required information in one place.

The SABIO-RK database has been developed to meet these requirements and to support scientists in modelling and understanding of complex biochemical networks by structuring kinetic data and related information from the literature. Compared to most of the other databases with a focus on proteins and enzymes, SABIO-RK uses a reaction-oriented approach to represent the available kinetic parameters from publications or directly from *in vitro* wet-laboratory experiments together with kinetic laws, corresponding rate equations and the environmental conditions (e.g. pH, temperature, buffer), including the initial concentrations of reaction participants and enzymes under which the kinetic data were measured. From the beginning, SABIO-RK was implemented to represent both metabolic and signalling reactions but, because most publications of the past 50 years focused on metabolic reactions, this type of reaction represents the vast majority in SABIO-RK. In the future, there will be more signalling reactions included in SABIO-RK because, in recent years, knowledge of kinetic data for signalling reactions is growing.

Enzymatic reaction kinetics

The kinetics of enzyme catalyzed reactions is described by the reaction rates that depend on the mechanisms regarding how chemical compounds (e.g. substrates, products, cofactors, activators, and inhibitors) interact with the catalyzing enzyme. The first mathematical descriptions of kinetic rate equations for such reactions were described in detail by Victor Henri [8] and Leonor Michaelis and Maud Leonora Menten [9]. Besides the intrinsic thermodynamic and sterical properties of these interactions encoded in the kinetic parameters, the most essential external factors affecting the enzymatic reaction kinetics are the concentrations of the enzyme and the interacting chemical compounds, as well as pH, ionic strength and temperature. The influence of these environmental conditions on enzymatic reactions was one of the most important findings of Michaelis and Menten in 1913 and the subsequent use of buffered systems was a major prerequisite for obtaining reproducible kinetic data [10,11]. The concentrations of the reaction participants have an impact on the frequency of collisions between them because higher concentrations increase the probability of such encounters between these molecules, and therefore increase the probability for a conversion. The factors pH, ionic strength and temperature are very important because they not only influence the frequency of collisions between the molecules [12], but also can influence the active site of the enzyme by changing the charges of amino acids residues, etc. [13].

Thus, knowledge of the experimental conditions during the determination of kinetic parameters and rate equations is of crucial relevance for assessing the boundary conditions of the described kinetics. This information is important for the appropriateness of the chosen parameter values in the context of biochemical reaction networks that are modelled in systems biology. In classical enzymology, the parameter values of a reaction are determined by measuring the initial rate of product formation at increasing substrate concentrations up to a maximum concentration. From a secondary plot that describes reaction velocity as a function of substrate concentration (e.g. by transformation according to Lineweaver–Burk [14] or by plotting the reaction velocity against the logarithm of the substrate concentration [9]), the parameters of this function can be inferred (e.g. the Michaelis constant K_m and the turnover number k_{cat}) that describe the kinetic properties of the reaction and its dependency on concentrations of the participating molecules. Under nonsaturating substrate concentrations, K_m can be determined and, through extrapolation to infinite substrate concentrations, an apparent unimolecular rate constant k_{cat} can be estimated if the enzyme concentration $[E]$ in the assay is known using the definition of $V = [E] \times k_{cat}$ [15]. The International Union of Biochemistry and Molecular Biology recommends to use V instead of V_{max} because the rate does not define a maximum in the mathematical sense but a limit [16,17]. However, in publications analyzed for SABIO-RK, V_{max} is mainly used but, often, it is not clear whether the substrate concentration is saturating, nor whether the authors are aware of the difference between the maximum and limiting rate. Because of the often ambiguous and sometimes even incorrect descriptions in the original articles, and also because of the aim of distinguishing between V and v for rates, SABIO-RK uses the term V_{max}.

In publications, the majority of kinetic parameters specifying biochemical reactions are V_{max} or rather V, k_{cat} and K_m values. These kinetic constants are critical to understanding how enzymes work. Kinetic parameters are important for computational modelling (e.g. to reconstruct and understand biochemical networks, their regulation and the interaction of the network components). These parameters are also used by wet-laboratory scientists to obtain insights into the mechanisms regarding how chemical compounds interact with selected enzymes under specific environmental conditions. Therefore, the collection of information about enzyme kinetics in a specialized database and its presentation in a structured and standardized format will support diverse users with different backgrounds and requirements.

In summary, knowledge of kinetic parameters without being aware of further details about the kinetic mechanism and environmental conditions (pH, temperature and buffer) is not sufficient for the modelling and simulation of experimental results. Ideally, the correct mathematical description of the corresponding rate equation should exactly characterize the complete information about how all the reactants and modifying ligands of a reaction interact to affect the reaction velocity. The completely known kinetic mechanism of a reaction may indicate the way in which the activity of the enzyme is regulated.

Over past decades, the description of enzymatic reaction kinetics in the scientific literature remains far from complete. During the analysis of papers for data extraction for the SABIO-RK database, we could not perceive a difference in the completeness of information by comparing articles from different publication years over the past five decades. Sometimes, concentration ranges have to be estimated from graphs or the original data of the graphs are not provided at all. Apart from kinetic parameters, the kinetic mechanisms and rate equations are rarely given. More than 90% of publications analyzed contain no rate equation. Authors who applied the Michaelis–Menten equation to determine kinetic parameters only used the term 'Michaelis–Menten' in approximately 40% of the papers to describe the experimental set-up and the parameter determination. In many publications, authors have used graphical representations to estimate affinity constants for substrates (K_m) or inhibitors (K_i) but a detailed characterization of the kinetic mechanism and their description in mathematical equations is missing. There are different ways of applying the graphical representation for the determination of the kinetic parameters (e.g. Lineweaver–Burk, Hanes–Woolf or Eadie–Hofstee plots). Many of the publications describe the use of the double-reciprocal Lineweaver–Burk plot ($1/v$ against $1/S$), remaining unaware of the fact that results could be misleading with regard to the experimental errors. Parameters determined by Hanes–Woolf (S/v against S) or Eadie–Hofstee (v against v/S) plots are more accurate because these plots are more robust against error-prone data [16]. For publications where such graphical representations were used without describing the rate equation, SABIO-RK provides the equation supplementary with the corresponding kinetic law type 'Michaelis–Menten'.

When Michaelis and Menten published their paper in 1913 about the kinetic analysis of invertase 'all of the original data for each of the figures was provided in tables, a useful feature lacking in today's publications' [18]. Nowadays,+ it is 'increasingly common that papers not only report no primary data but also report no secondary data either' [19] by giving only the final results of the data processing, which makes it hard to reproduce and compare the data with other experiments. The format in which kinetic parameters are represented within articles comprises tables, figures or free text, and the information is highly scattered. For most journals, there is no specification provided to authors with respect to how kinetic parameters or kinetic rate equations should be represented in a publication. Additionally, the publication of the data in more detail as supplementary information (e.g. online) would greatly increase its value and reusability.

Standard formats and controlled vocabularies

Especially for computational modelling and computer-assisted exchange of knowledge, a definition of standard data exchange formats and a common language are essential [20]. Therefore, the use of existing ontologies and controlled vocabularies for all reaction participants (e.g. small chemical compounds and proteins), as well as for kinetic rate laws, parameters, units, etc., is becoming more and more essential. This requires that wet-laboratory scientists, who prefer focusing on research rather than the importance of automatic data processing, should also use standard terminologies and formats when describing their data in publications, with the aim of facilitating the exact identification of objects and enabling their peers to compare different experimental results. Ever since the foundation of protein and gene related databases (e.g. UniProtKB [3], PDB [21], DDBJ/EMBL/GenBank [22]) in the 1980s, protein and gene identifiers have been provided in publications to some extent. However, according to our observations, there is no tendency for their elevated usage over the last 20 years (U. Wittig, unpublished data).

Also, EC numbers are only used in less than half of the publications, although they are the most unique identifiers for the enzymatic activity of a protein in a biochemical reaction given by the International Union of Biochemistry and Molecular Biology. In addition, unique identifiers are mostly missing for small chemical compounds acting as reaction participants (e.g. substrates, products, inhibitors, activators).

For kinetics data related information in articles, the situation is even more inexplicit. Apart from lacking identifiers, there is also no exact terminology for kinetic rate laws, parameter types or unit definitions. Already in the original paper of Michaelis and Menten, a 'loose usage of concepts' was criticized [18]. However, at present, ontologies are developed such as the Systems Biology Ontology (SBO) [20]. This is composed of hierarchically arranged sets of controlled vocabularies that are commonly used in mathematical modelling in the field of systems biology. Besides others, SBO includes the description of kinetic rate laws, reaction participant's roles and kinetic parameter types. The use of SBO terms would be beneficial because it is also employed for the unambiguous description in databases and within the standard data exchange formats: Systems Biology Markup Language (SBML) [23] and Systems Biology Pathway Exchange (SBPAX) [24]. However, in the literature concerning biochemical reaction kinetics data, no usage of a controlled terminology or unique identifiers can be found. For example, the maximal velocity of an enzymatic reaction could not only be given in a publication as V or V_m or V_{max}, but also as the 'maximum rate' or 'maximum velocity'. Independent of the naming, this kinetic parameter should be referred to the SBO entry SBO:0000186 for a unique identification or, more specifically, to the forward maximal velocity SBO:0000324 or the reverse maximal velocity SBO:0000325, which are both ontological children of the superior term. Besides term definition and ontological relations to other objects, the SBO also provides the equation for the calculation of the parameter. If this would be kept in mind by authors, no misrepresentation would occur, such as the use of the unit $1/s$ for a maximal velocity, as is often used in articles. The maximal velocity is defined by the equation $V_{max} = [E] \times k_{cat}$. This formula implies that the unit for the enzyme concentration $[E]$ has to be included in the unit for V_{max}. Because k_{cat} is the maximum number of substrate molecules that an enzyme converts per time per catalytic site, the above mentioned unit $1/s$ refers to k_{cat} instead of V_{max}. Obviously, both terms are often mixed up in publications. Another misrepresentation of kinetic parameters in the literature is the often confusing and sometimes wrong assignment of 'enzyme activity', 'specific activity', V_{max} and 'specific V_{max}'. 'Enzyme activity' describes the catalytic effect (conversion of substrate per time) exerted by the active enzyme for given concentrations of all reaction participants, whereas 'specific activity' gives this information for the active enzyme per total amount of protein used in the assay. V_{max} (or more precisely the limiting rate V) represents the activity of the enzyme under infinite substrate concentrations, a constant at a given temperature and a given enzyme concentration. Finally, the 'specific V_{max}' defines the limiting rate per total amount of protein used in the assay. Sometimes, all four terms can be found mixed up within one publication. Especially if the substrate and/or enzyme concentration is not correctly defined, it is difficult to estimate the correct parameter type.

The reference to the term 'enzyme concentration', which, for example, is important for the calculation of V_{max}, can be different dependent on the purity of the enzyme. There are publications that describe the concentration per number of cells or based on the fresh weight of the cells or tissue. Other authors describe an experimental assay that may contain crude extract, or purified or partially purified protein containing the enzyme. Thus, without the molecular weight of the enzyme protein, the enzyme composition of catalytic sites and detailed information about the purity of the enzyme, it is very hard, if not impossible, to calculate parameters such as V_{max} [25].

As already noted above, the experimental conditions are essential for the interpretation of the kinetic behaviour of an enzymatic reaction. As described for enzyme concentrations, there is also no standard for the representation of assay buffers and substrate concentrations. First, information about the experimental conditions can be distributed over the whole publication. Of course, assay information is expected in the Materials and methods section of a paper, although most legends of any tables and figures also contain information about pH, temperature, buffer, and/or concentrations of reaction participants. Often, participant concentrations are only given within figures and have to be estimated from the axis of a diagram. Dependent on the quality of the diagram, the extracted values might be inexact. Sometimes conflicts can be even observed between the information written in the Materials and methods sections and in the legends or figures.

Additionally, sometimes experimental conditions are not adequately described. In approximately 10% of papers, the information on the assay temperature is missing and, in 3% of the papers, it is just specified as 'room temperature'. Furthermore, in 20% of publications, the concentrations for compounds and buffer components are not provided in standard units (SI units, according to the International System of Units) but, for example, as the absolute amount of compound used in the assay volume, and therefore these have to be converted manually.

Standardization initiatives and tools

As a result of the increasing need to standardize data and its description (so-called metadata), there are more and more different initiatives that define exchange formats, controlled vocabularies and reporting guidelines representing the knowledge of a specific field. At the same time, there is insight that scientists need help to create data conforming to standards in a time-saving manner. The goal here is to make adherence to standards less of a chore and rather a direct time saver in the self-management of local data.

Most of the initiatives in the biomedical field are registered at BioSharing (http://www.biosharing.org), which is a catalogue of reporting standards. The Minimum Information for Biological and Biomedical Investigations [26] project provides a framework for coherent minimum reporting guidelines for different data types in the form of minimum information checklists. Especially for enzyme kinetics, the Standards for Reporting Enzymology Data (STRENDA) initiative [27] defined minimal information for reporting enzyme data (http://www.strenda.org). Many of the existing databases already use standards defined within these different standardization projects such as STRENDA, and journals should now also implement them and enforce authors to publish complete, consistent and structured data in standard formats. It should be an integral part of the peer reviewing process and be assimilated in the instructions to authors. The main goal of the STRENDA initiative is to establish a database that could be used by authors when submitting a manuscript to a journal to store data about the experimental set-up and the measured results in a standardized format in parallel to the written paper. After the publication of these data, other databases such as SABIO-RK would be able to download these structured data. From a database point of view, we prefer the direct submission of published data in a structured and standardized format instead of the need to manually extract information from the text. Authors would be guided through an electronic submission form and standards and identifiers could be automatically included. SABIO-RK together with collaboration partners [28] have already implemented a data submission workflow directly from laboratory experiments into a database. Here, methods for comprehensive data annotation in spreadsheets using, for example, RIGHTFIELD [29] or in SBML files using SEMANTICSBML [30] could support the integration process. The tool EXCEMPLIFY [31] is built to create RIGHTFIELD-compatible templates at the same time as helping experimentalists with the handling of experimental data. Thus, even preliminary data that are likely to be discarded later can be enriched using metadata of the experiments compliant with definitions in ontologies and controlled vocabularies.

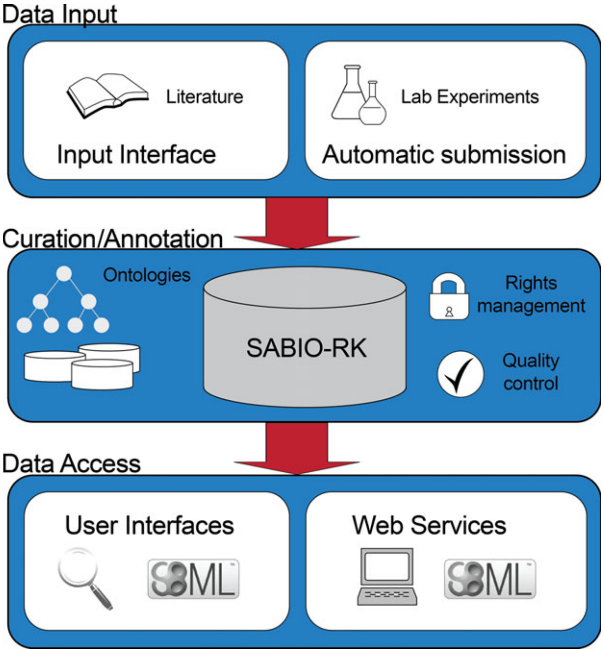


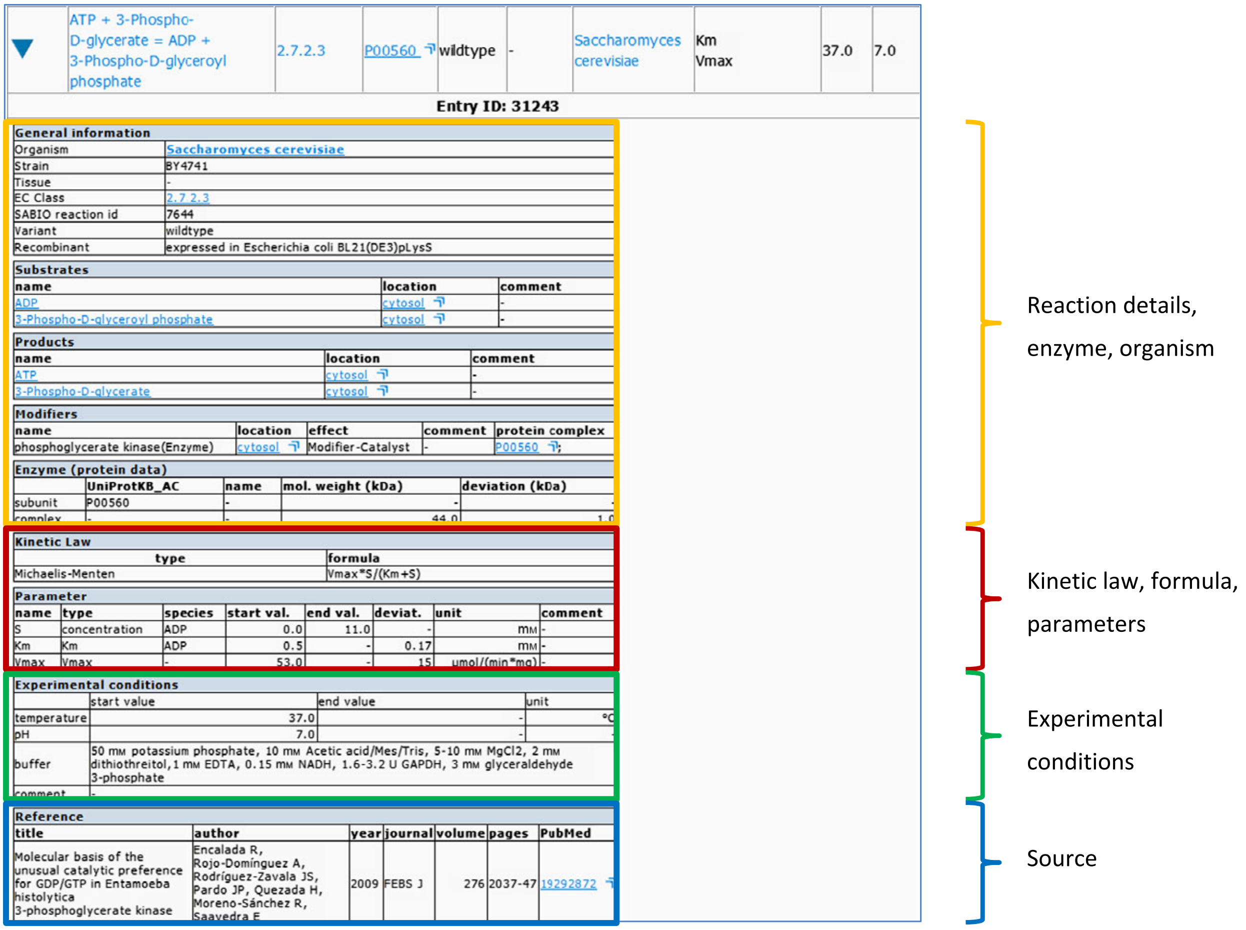
Fig. 1. Representation of the SABIO-RK data workflow.

SABIO-RK database

SABIO-RK is a manually curated database for enzymatic reaction kinetics. Data are either extracted from scientific articles [32] or directly submitted by wet-laboratory experimentalists [30]. During a typical workflow (Fig. 1), published data are manually inserted by the students or biological experts who first read the publications using a web-based input interface. Subsequently, the same input interface is used by database curators who read the paper a second time to validate the data and to adjust them to SABIO-RK data standards. This double check is needed to avoid errors and inconsistencies. Finally, the data are transferred to the public online database.

In the SABIO-RK database, biochemical reactions are defined by their reaction participants (substrates, products), modifiers (inhibitors, activators, cofactors), catalyst details (e.g. EC enzyme classification, UniProtKB accession numbers, protein complex composition of the active enzyme, isozymes, wild-type/mutant information, molecular weight) and their biological source (organism, tissue/cell type, cell location). This is not restricted to any organism classes. SABIO-RK data can be simply accessed through web-based user interfaces and web services. Various search criteria are selectable to search for biochemical reactions and their kinetics. Beside a free text search, complex and detailed queries can be executed in the advanced search. This may include the combination of several search criteria [e.g. reaction participants (substrates, products, inhibitors, activators etc.), pathways, enzymes, organisms, tissues or cellular locations, kinetic parameters, environmental conditions or literature sources]. When entering the search terms, the number of kinetic data entries is displayed that is available in the database matching the search criteria. Further sorting and grouping features are implemented in three views with a different focus, which also offer alternatives for further modification of the query. The search criteria also comprise SABIO-RK internal identifiers and identifiers from external databases (e.g. UniProtKB [3], KEGG [33], ChEBI [34]) based on supplementary added annotations. Selected complete database entries or grouped datasets can be exported in different file formats: SBML, BioPAX/SBPAX and a simple table format. With the exception of the latter, annotations to external databases and ontologies are always included [7].

SABIO-RK stores all of the kinetic information for one specific reaction under specific experimental conditions from a defined biological source in one dataset called the database entry. This information can be viewed and exported as a single dataset. As shown in Fig. 2, the general information about the organism and the tissue is described, as well as the enzyme and reaction participants (yellow), followed by kinetic information including rate laws and formulas and the



▼	ATP + 3-Phospho-D-glycerate = ADP + 3-Phospho-D-glyceroyl phosphate	2.7.2.3	P00560	wildtype	-	Saccharomyces cerevisiae	Km Vmax	37.0	7.0

Entry ID: 31243

General information	
Organism	Saccharomyces cerevisiae
Strain	BY4741
Tissue	-
EC Class	2.7.2.3
SABIO reaction id	7644
Variant	wildtype
Recombinant	expressed in Escherichia coli BL21(DE3)pLysS

Substrates

name	location	comment
ADP	cytosol	-
3-Phospho-D-glyceroyl phosphate	cytosol	-

Products

name	location	comment
ATP	cytosol	-
3-Phospho-D-glycerate	cytosol	-

Modifiers

name	location	effect	comment	protein complex
phosphoglycerate kinase(Enzyme)	cytosol	Modifier-Catalyst	-	P00560 ;

Enzyme (protein data)

	UniProtKB_AC	name	mol. weight (kDa)	deviation (kDa)
subunit	P00560	-	-	-
complex	-	-	44.0	1.0

Kinetic Law

type	formula
Michaelis-Menten	Vmax*S/(Km+S)

Parameter

name	type	species	start val.	end val.	deviat.	unit	comment
S	concentration	ADP	0.0	11.0	-	mM	-
Km	Km	ADP	0.5	-	0.17	mM	-
Vmax	Vmax	-	53.0	-	15	μmol/(min*mg)	-

Experimental conditions

	start value	end value	unit
temperature	37.0	-	°C
pH	7.0	-	-
buffer	50 mM potassium phosphate, 10 mM Acetic acid/Mes/Tris, 5-10 mM MgCl2, 2 mM dithiothreitol,1 mM EDTA, 0.15 mM NADH, 1.6-3.2 U GAPDH, 3 mM glyceraldehyde 3-phosphate		
comment	-		

Reference

title	author	year	journal	volume	pages	PubMed
Molecular basis of the unusual catalytic preference for GDP/GTP in Entamoeba histolytica 3-phosphoglycerate kinase	Encalada R, Rojo-Domínguez A, Rodríguez-Zavala JS, Pardo JP, Quezada H, Moreno-Sánchez R, Saavedra E	2009	FEBS J	276	2037-47	19292872

Reaction details, enzyme, organism
Kinetic law, formula, parameters
Experimental conditions
Source

Fig. 2. SABIO-RK database entry representing the data structure.

corresponding parameters (red). Then, the experimental conditions pH, temperature and buffer are represented (green) and, finally, the original source of the data is cited (blue).

One of the main goals of the SABIO-RK database is to facilitate and support the process of computational modelling. Accordingly, SABIO-RK is integrated in systems biology applications [35] and a number of modelling platforms, including CELLDESIGNER [36], VIRTUAL CELL [37] or SYCAMORE [38], which either make use of SABIO-RK's web services or the web interface.

The SABIO-RK database is mainly populated with data manually extracted from the literature, which requires biological expert knowledge for an understanding of the publication, the extraction and standardization of relevant information, and the guarantee of high-quality data in the database. As a result of the missing controlled vocabularies and annotations to standard identifiers, the manual data extraction comprises extra work for the biological expert to interpret and assign the information. To reduce errors and inconsistencies during data insertion, database internal selection lists with controlled vocabularies are used and constraints are included to check and structure the data. For example, a consistency check is implemented within the input interface to control the parameters given in the rate equation with the list of available parameters. If not all of the parameters are given in the paper, for consistency reasons, 'dummy' parameters are created with values of 'null' to offer complete datasets for modellers during data export, especially in SBML format. Therefore, parameters are sometimes defined in the database, although no values were provided in the original literature. Only 78.8% of the database entries in SABIO-RK containing a rate equation include a substrate affinity constant (K_m or S_half) together with a reaction velocity constant (V_{max} or k_{cat}). Therefore, for more than 20% of the entries, either a substrate affinity constant or a velocity constant, or both, are missing, with the last representing, for example, rate equations for inhibitions where only a inhibition constant is given without further substrate or reaction-related parameters.

The manual data extraction and curation process also includes the annotation of data to ontologies, controlled vocabularies and external databases. SABIO-RK uses the following biological ontologies and controlled vocabularies for the various attributes: ChEBI [34], SBO [20], BTO (BRENDA Tissue Ontology) [39], NCBI (National Center for Biotechnology Information) organism taxonomy [40] and Gene Ontology [41]. Based on these annotations, the correct interpretation, exchange, comparison and cross-referencing of data is possible. On the other hand, external databases such as KEGG [33], UniProtKB [3] or ChEBI use these annotations to cross-reference to SABIO-RK database entries [7].

As of May 2013, the SABIO-RK database stores kinetic parameters for 5737 different biochemical reactions in approximately 44 000 database entries. On average, ten database entries are extracted from one publication because any possible variation of the experimental conditions or tissues or organisms results in the creation of a new entry in the database. Based on the information available in publications, rate equations are available for approximately 52% of all entries in the database. The majority of these database entries are defined as being of Michaelis–Menten kinetic law type and represent 42% of all SABIO-RK entries (Fig. 3). Michaelis–Menten kinetics is provided for 3659 different biochemical reactions. Because only one-third of these reactions are single-substrate reactions (water is ignored as a substrate for that calculation), two-thirds of the biochemical reactions with Michaelis–Menten kinetics in SABIO-RK are multiple-substrate reactions and therefore do not represent real Michaelis–Menten laws *in vivo*. For multiple-substrate reactions, there are different types of kinetic mechanisms (e.g. ordered or random sequential ternary-complex, Theorell-Chance or Ping-Pong mechanisms for Bi-Bi reactions). These kinetic law types are usually only tagged by the type of reaction, although the corresponding rate equation is not provided by the authors. This missing information cannot be supplied by SABIO-RK database curators without making further assumptions, and therefore it cannot be considered by modellers for their computational model set-up [42]. If needed, modellers are able to handle this by using convenience kinetics [1], in contrast to enzymologists who typically need to know the correct and detailed enzyme kinetic mechanism.

For the determination of kinetic parameter values for multiple-substrate reactions, Michaelis–Menten kinetics could be applied if one substrate is varied and the other substrate(s) is kept constant on a saturating

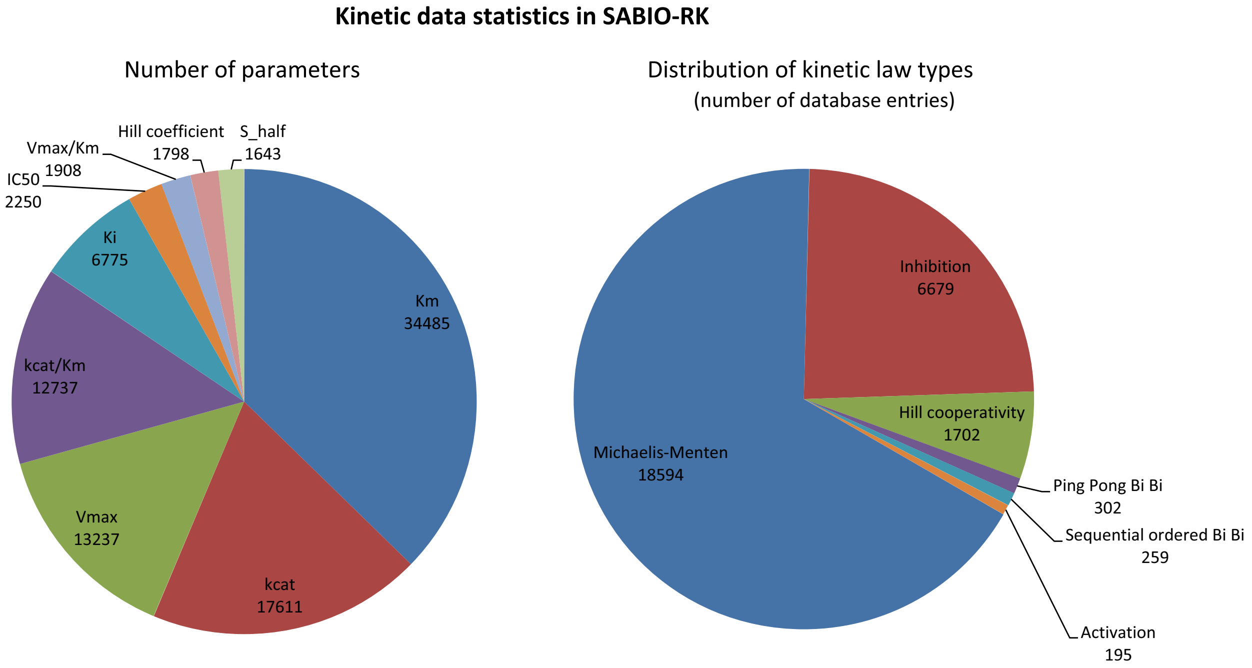


Fig. 3. Statistics on the most frequent kinetic parameters and kinetic law types stored in SABIO-RK.

level. Therefore, kinetic parameters such as K_m values are measured under pseudo-single-substrate conditions for one substrate at varied concentrations to determine its K_m value, whereas all other substrates are kept constant under saturating concentrations. Many publications describe the use of Michaelis–Menten kinetics for this single substrate without explaining detailed kinetics for the whole reaction. Under these conditions, the reaction could be seen as a single-substrate reaction and the parameter values should be given as 'apparent' values. In the literature, the naming of such parameter values as 'apparent' is not consistent. These *in vitro* analyses of multiple-substrate reactions cannot be extrapolated to *in vivo* conditions within living cells where the concentrations of the reaction participants are different from the *in vitro* saturating conditions [15]. Because of numerous problems with respect to measuring *in vivo* data, it should be realized that all current data in SABIO-RK are *in vitro* data and therefore any models built from these data should be critically regarded and extrapolated to the situation in living cells. 'Models are not descriptions of reality; they are descriptions of our assumptions about reality' [10].

Summary

In the present review, we describe the challenges experienced with respect to the development and maintenance of a database for biochemical reaction kinetics using SABIO-RK as a paradigm. The typical workflow not only includes data extraction from the literature, but also extensive manual work to complete and annotate the information for data storage and export in a standardized way. From computational modelling and database development points of view, it appears that authors are usually unaware of the importance of the reusability of their published data. Accordingly, we highly recommend that authors use standard terminologies and unique identifiers referring to databases and ontologies. This should be ideally supported by publishers and journal editors. Furthermore, all assay details (e.g. temperature, pH, etc.) should be described in the publication and should not only refer to other publications. When writing papers, authors should keep in mind the reusability of their data. Minimal information for enzyme kinetics are recommended within the STRENDA initiative, which defined guidelines for the publication of enzyme data that are already accepted by some biological journals and have been inserted in the author's guidelines of these journals. The International Society for Biocuration (http://www.biocurator.org) representing the biocuration community also initiates collaborations between database curators and publishers. Ideally, authors should take advantage of the possibility of storing supplementary data related to the publication and additionally submit their data directly to databases. This would support both computational modelling and database population, although changes for a better way of writing papers would only affect future publications. The extraction of data from already existing publications will still comprise time-consuming manual curation work.

Acknowledgements

Special thanks are extended to our database users for their feedback during the database development process and for many discussions about their requirements. The SABIO-RK project is financed by the Klaus Tschira Foundation (http://www.klaus-tschira-stiftung.de/), the German Federal Ministry of Education and Research (http://www.bmbf.de/) via the projects Virtual Liver Network (grant number 0315749) and SysMO-LAB (Systems Biology of Microorganisms), and the DFG LIS (http://www.dfg.de/) as part of the project 'Integrierte Immunoblot Umgebung'.

References

1 Borger S, Uhlendorf J, Helbig A & Liebermeister W (2007) Integration of enzyme kinetic data from various sources. *In Silico Biol* **7**, S73–S79.

2 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**, D670–D676.

3 The UniProt Consortium (2011) Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res* **39**, D214–D219.

4 Le Novère N, Bornstein B, Broicher A, Courtot M, Donizelli M, Dharuri H, Li L, Sauro H, Schilstra M, Shapiro B *et al.* (2006) BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. *Nucleic Acids Res* **34**, D689–D691.

5 Olivier BG & Snoep JL (2004) Web-based kinetic modelling using JWS Online. *Bioinformatics* **20**, 2143–2144.

6 Sivakumaran S, Hariharaputran S, Mishra J & Bhalla US (2003) The Database of Quantitative Cellular Signaling: management and analysis of chemical kinetic models of signaling networks. *Bioinformatics* **19**, 408–415.

7 Wittig U, Kania R, Golebiewski M, Rey M, Shi L, Jong L, Algaa E, Weidemann A, Sauer-Danzwith H,

Mir S *et al.* (2012) SABIO-RK – database for biochemical reaction kinetics. *Nucleic Acids Res* **40**, D790–D796.

8 Henri V (1902) Théorie générale de l'action de quelques diastases. *Comptes Rendus l'Académie des Sci* **135**, 916–919.

9 Michaelis L & Menten ML (1913) Die Kinetik der Invertinwirkung. *Biochem Z* **49**, 333–369.

10 Gunawardena J (2012) Some lessons about models from Michaelis and Menten. *Mol Biol Cell* **23**, 517–519.

11 Bisswanger H (2002) Enzyme Kinetics: Principles and Methods. Wiley VCH, Weinheim, pp. 51–75.

12 Trautz M (1916) Das Gesetz der Reaktionsgeschwindigkeit und der Gleichgewichte in Gasen. Bestätigung der Additivität von Cv-3/2R. Neue Bestimmung der Integrationskonstanten und der Moleküldurchmesser. *Z anorg allg Chem* **96**, 1–28.

13 Segel IH (1993) Enzyme Kinetics: Behavior and Analysis of Rapid Equilibrium and Steady-State Enzyme Systems. Wiley Classics Library ed, New York.

14 Lineweaver H & Burk D (1934) The determination of enzyme dissociation constants. *J Am Chem Soc* **56**, 658–666.

15 Chen WW, Niepel M & Sorger PK (2010) Classic and contemporary approaches to modeling biochemical reactions. *Genes Dev* **24**, 1861–1875.

16 Cornish-Bowden A (2012) Fundamentals of Enzyme Kinetics. Wiley VCH, Weinheim, Germany, pp. 28–71.

17 Nomenclature Committee of the International Union of Biochemistry (NC-IUB) (1982) Symbolism and terminology in enzyme kinetics. Recommendations 1981. *Eur J Biochem* **128**, 281–291.

18 Johnson KA & Goody RS (2011) The original Michaelis Constant: translation of the 1913 Michaelis–Menten paper. *Biochemistry* **50**, 8264–8269.

19 Cornish-Bowden A (2001) Detection of errors of interpretation in experiments in enzyme kinetics. *Methods* **24**, 181–190.

20 Courtot M, Juty N, Knüpfer C, Waltemath D, Zhukova A, Dräger A, Dumontier M, Finney A, Golebiewski M, Hastings J *et al.* (2011) Controlled vocabularies and semantics in systems biology. *Mol Syst Biol* **7**, 543.

21 Berman HM (2008) The Protein Data Bank: a historical perspective. *Acta Crystallogr A* **64**, 88–95.

22 Nakamura Y, Cochrane G, Karsch-Mizrachi I; International Nucleotide Sequence Database Collaboration (2013) The International Nucleotide Sequence Database Collaboration. *Nucleic Acids Res* **41**, D21–D24.

23 Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, Kitano H, Arkin AP, Bornstein BJ, 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**, 524–531.

24 Ruebenacker O, Moraru II, Schaff JC & Blinov ML (2009) Integrating BioPAX pathway knowledge with SBML models. *IET Syst Biol* **3**, 317–328.

25 Kummer U & Sahle S (2007) Problems of currently published enzyme kinetic data for usage in modelling and simulation. In Proceedings of the 2nd International Symposium on 'Experimental Standard Conditions of Enzyme Characterizations', Ruedesheim am Rhein, Germany, 2006 (Hicks MG & Kettner C, eds), pp. 129–136. Logos-Verlag, Berlin.

26 Taylor CF, Field D, Sansone SA, Aerts J, Apweiler R, Ashburner M, Ball CA, Binz PA, Bogue M, Booth T *et al.* (2008) Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nat Biotechnol* **26**, 889–896.

27 Apweiler R, Cornish-Bowden A, Hofmeyr JH, Kettner C, Leyh TS, Schomburg D & Tipton K (2005) The importance of uniformity in reporting protein-function data. *Trends Biochem Sci* **30**, 11–12.

28 Swainston N, Golebiewski M, Messiha HL, Malys N, Kania R, Kengne S, Krebs O, Mir S, Sauer-Danzwith H, Smallbone K *et al.* (2010) Enzyme kinetics informatics: from instrument to browser. *FEBS J* **277**, 3769–3779.

29 Wolstencroft K, Owen S, Horridge M, Krebs O, Mueller W, Snoep JL, du Preez F & Goble C (2011) RightField: embedding ontology annotation in spreadsheets. *Bioinformatics* **27**, 2021–2022.

30 Krause F, Uhlendorf J, Lubitz T, Schulz M, Klipp E & Liebermeister W (2010) Annotation and merging of SBML models with semanticSBML. *Bioinformatics* **26**, 421–422.

31 Shi L, Jong L, Wittig U, Lucarelli P, Stepath M, Mueller S, D'Alessandro LA, Klingmüller U & Müller W (2013) Excemplify: a flexible template based solution, parsing and managing data in spreadsheets for experimentalists. *J Integr Bioinform* **10**, 220.

32 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. *Lect Notes Bioinformat* **4075**, 94–103.

33 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**, D355–D360.

34 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**, D249–D254.

35 Li P, Dada JO, Jameson D, Spasic I, Swainston N, Carroll K, Dunn W, Khan F, Malys N, Messiha HL *et al.* (2010) Systematic integration of experimental data

and models in systems biology. *BMC Bioinformatics* **11**, 582.

36 Funahashi A, Jouraku A, Matsuoka Y & Kitano H (2007) Integration of cell designer and SABIO-RK. *In Silico Biol* **7**, S81–S90.

37 Moraru II, Schaff JC, Slepchenko BM, Blinov ML, Morgan F, Lakshminarayana A, Gao F, Li Y & Loew LM (2008) Virtual Cell modelling and simulation software environment. *IET Syst Biol* **2**, 352–362.

38 Weidemann A, Richter S, Stein M, Sahle S, Gauges R, Gabdoulline R, Surovtsova I, Semmelrock N, Besson B, Rojas I *et al.* (2008) SYCAMORE – a systems biology computational analysis and modeling research environment. *Bioinformatics* **24**, 1463–1464.

39 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**, D507–D513.

40 Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Federhen S *et al.* (2011) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **39**, D38–D51.

41 The Gene Ontology Consortium (2000) Gene ontology: tool for the unification of biology. *Nat Genet* **25**, 25–29.

42 Kummer U (2007) Usage of reaction kinetics data stored in databases – a modeler's point of view. *In Silico Biol* **7**, S65–S71.