Journal of Biotechnology 261 (2017) 229–237

Contents lists available at ScienceDirect

Journal of Biotechnology

journal homepage: www.elsevier.com/locate/jbiotec

Data management and data enrichment for systems biology projects

Ulrike Wittig, Maja Rey, Andreas Weidemann, Wolfgang Müller*

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

ARTICLE INFO

Keywords:
Data management
Systems biology
SABIO-RK
Reaction kinetics
SEEK
FAIR

ABSTRACT

Collecting, curating, interlinking, and sharing high quality data are central to de.NBI-SysBio, the systems biology data management service center within the de.NBI network (German Network for Bioinformatics Infrastructure). The work of the center is guided by the FAIR principles for scientific data management and stewardship. FAIR stands for the four foundational principles *Findability, Accessibility, Interoperability,* and *Reusability* which were established to enhance the ability of machines to automatically find, access, exchange and use data.

Within this overview paper we describe three tools (SABIO-RK, Excemplify, SEEK) that exemplify the contribution of de.NBI-SysBio services to FAIR data, models, and experimental methods storage and exchange. The interconnectivity of the tools and the data workflow within systems biology projects will be explained. For many years we are the German partner in the FAIRDOM initiative (http://fair-dom.org) to establish a European data and model management service facility for systems biology.

1. Introduction

The increasing amount of data doesn't necessarily entail an increasing amount of knowledge. "The real problem is that we have failed to store and organize much of the rapidly accumulating information (whether in databases or documents) in rigorous, principled ways, so that finding what we want and understanding what's already known become exhausting, frustrating, stressful and increasingly costly experiences" (Attwood et al., 2009). To use and reuse the data, storage, organization and communication in a structured and standardized format is needed. Today, the FAIR principles sum up what data organization should be: Findable, Accessible, Interoperable, and Reusable (Wilkinson et al., 2016). All of these principles, except for *Accessibility* rely on data quality. Biocuration is a key to data quality (Bateman, 2010):

Findability is enhanced by using standard identifiers and annotations which point to standard ontologies and databases. The same applies for the use of controlled vocabularies. This allows answering questions that arise from ambiguous information, like for example: "Has the abbreviation 'Glu' in one document the same meaning in another document?". An identifier based on standards determines unambiguously that Glu represents either Glucose or Glutamate.

Interoperability is greatly enhanced by using common exchange formats. In systems biology, SBML (Systems Biology Markup Language) (Hucka et al., 2003) is a commonly and widely-used example. Standard exchange formats allow the automatic and machine-readable data exchange and enables the development of automatic data workflows between databases, data management systems and applications, e.g. simulation tools.

Finally, *Reusability* is greatly enhanced if a file carries metadata, i.e. descriptive information including information about its context. It includes information about the original data source (e.g. organism, laboratory sample), procedures how data were generated (e.g. experimental setup, environmental conditions), and further information about unique data attribution. This relevant parameters should be present in a data file or connected to it. The MIBBI (Minimum Information for Biological and Biomedical Investigations) standards initiative seeks to provide the minimum context needed for information exchange to "fully understand the context, methods, data and conclusions that pertain to an experiment" (Taylor et al., 2008).

All the information needed to represent and understand the data depend on the manual work of biological experts to annotate and curate the information. Biocuration is the transformation of biological data into an organized form (Bateman, 2010). And the main motivation goes in the direction of self-curation of data by experimentalists and authors to sensitize the owner of the data to standards, formats and controlled vocabularies.

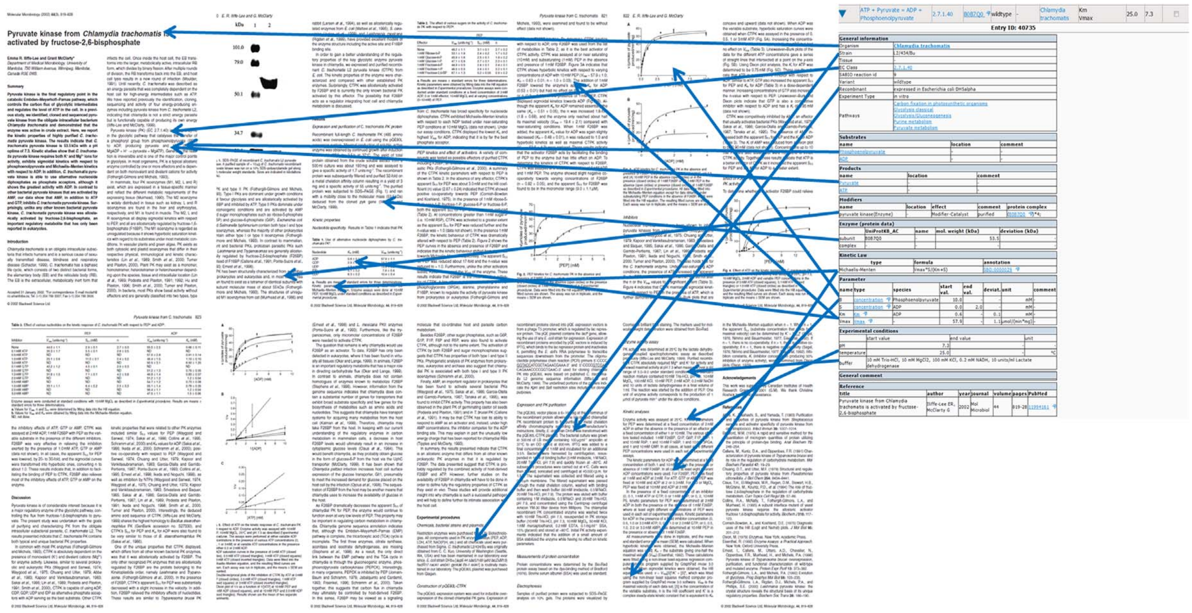
Within this paper we will first present the curated system SABIO-RK, a database for highly-structured, quickly reusable kinetics data for biochemical reactions.

We will also describe our HCI (human computer interaction) and algorithmic efforts towards increasing findability of data in everyday use (e.g see section *Excemplify*).

The paper will explain the FAIR principles applied to the tools we

* Corresponding author.
E-mail address: wolfgang.mueller@h-its.org (W. Müller).

http://dx.doi.org/10.1016/j.jbiotec.2017.06.007
Received 21 February 2017; Received in revised form 6 June 2017; Accepted 9 June 2017
Available online 10 June 2017



▼	ATP + Pyruvate = ADP + Phosphoenolpyruvate	2.7.1.40	B0B7Q0	wildtype	-	Chlamydia trachomatis	Km Vmax	25.0	7.3	□

Entry ID: 40735

General information	
Organism	Chlamydia trachomatis
Strain	L2/434/Bu
Tissue	-
EC Class	2.7.1.40
SABIO reaction id	9
Variant	wildtype
Recombinant	expressed in Escherichia coli DH5alpha
Experiment Type	in vitro
Pathways	Carbon fixation in photosynthetic organisms Glycolysis classical Glycolysis/Gluconeogenesis Purine metabolism Pyruvate metabolism

Substrates

name	location	comment
Phosphoenolpyruvate	-	-
ADP	-	-

Products

name	location	comment
Pyruvate	-	-
ATP	-	-

Modifiers

name	location	effect	comment	protein complex
pyruvate kinase(Enzyme)	-	Modifier-Catalyst	purified	(B0B7Q0)*4;

Enzyme (protein data)

	UniProtKB_AC	name	mol. weight (kDa)	deviation (kDa)
subunit	B0B7Q0	-	53.5	-
complex	-	-	-	-

Kinetic Law

type	formula	annotation
Michaelis-Menten	Vmax*S/(Km+S)	SBO:0000029

Parameter

name	type	species	start val.	end val.	deviat.	unit	comment
B	concentration	Phosphoenolpyruvate	10.0	-	-	mM	-
S	concentration	ADP	0.0	2.0	-	mM	-
Km	Km	ADP	0.6	-	0.1	mM	-
Vmax	Vmax	-	57.9	-	1.1	µmol/(min*mg)	-

Experimental conditions

	start value	end value	unit
pH	7.3	-	-
temperature	25.0	-	°C
buffer	10 mM Tris-HCl, 10 mM MgCl2, 100 mM KCl, 0.2 mM NADH, 10 units/ml Lactate dehydrogenase		

General comment

Reference

title	author	year	journal	volume	pages	PubMed
Pyruvate kinase from Chlamydia trachomatis is activated by fructose-2,6-bisphosphate	Iliffe-Lee ER, McClarty G	2002	Mol Microbiol	44	819-28	11994161

Fig. 1. SABIO-RK database entry (ID 40735) and arrows indicating extracted information distributed in the corresponding publication (PMID 11994161). On the right one can see a structured SABIO-RK entry that puts semantically close data close to each other. The arrows show that items that are close semantically do not need to be close to each other within the publication.

offer as the de.NBI-SysBio service center within the de.NBI initiative. The de.NBI-SysBio center focuses on standards-based management solutions for data and models with the special core area of systems biology. Our expertise and experience lies in data management, data curation and data standards.

We first review the SABIO-RK database (http://sabiork.h-its.org), then Excemplify, a tool for data collection, and SEEK (http://www.seek4science.org), a tool for data and model management, that is the basis for the FAIRDOMHub (http://www.fairdomhub.org) data management service.

Note that all services described here involve humans providing service. Using Excemplify and SEEK/FAIRDOMHub however, the users must be trained in order to understand benefits and pitfalls of data curation, understand how to use a variety of tools and standards for accomplishing their goals. This is the reason why in infrastructure projects like the present de.NBI project, teaching and training, consulting and advice play an important role.

2. SABIO-RK – reaction kinetics database

SABIO-RK (http://sabiork.h-its.org/) is a web-accessible, manually curated database for biochemical reactions and their kinetics (Wittig et al., 2012). The database has been developed to support scientists in modelling and understanding of complex biochemical networks by structuring kinetic data and related information from literature. SABIO-RK uses a reaction-oriented approach for representing kinetic data compared to most of the other biological databases with a focus on proteins or enzymes (e.g. BRENDA (Placzek et al., 2017), UniprotKB (The UniProt Consortium, 2011)). Reaction- or pathway-oriented databases are for example KEGG (Kanehisa et al., 2014), Reactome (Croft et al., 2014), or MetaCyc (Caspi et al., 2014). Compared to these other databases SABIO-RK summarizes not all available kinetic data for one reaction or enzyme in one database entry but separates kinetic parameters based on the environmental conditions and literature sources. It stores available kinetic parameters from publications together with kinetic rate equations, protein/enzyme information, biological source, and environmental conditions. Users can access the database via the web interface or automatically using Python scripts via web services. They can export their search results in XML-formats (SBML, BioPAX/SBPAX (Demir et al., 2010; Ruebenacker et al., 2009 Ruebenacker et al., 2009)) or in a spreadsheet format which is mainly preferred by experimentalists.

2.1. Collecting and curating data from the literature

The content of the SABIO-RK database originates predominantly from scientific publications containing kinetic data. As of February 2017 SABIO-RK comprises more than 55.000 database entries with data extracted from more than 5.500 publications. These data are related to 912 organisms, 7.196 reactions, 1.559 enzymes, and 4.283 UniprotKB accession numbers for proteins. The kinetic parameters include 43.308 substrate specific constants (e.g. Km), 37.395 velocity constants (Vmax, kcat), and 11.215 inhibition constants (Ki, IC50).

The selection of articles is based on user requirements within collaborative projects or external user contacts. SABIO-RK offers a public curation service accessible on the user interface website where users are encouraged to send requests for specific research interests. Especially users who could not get sufficient search results in SABIO-RK are automatically invited to add curation requests.

A typical data integration and curation workflow includes the selection of publications from literature search, the reading of the articles, the manual extraction of information by students or biological experts and the manual insertion of the data using a web-based input interface.

To avoid errors and inconsistencies SABIO-RK database curators read the paper a second time to validate the data and to adjust them to SABIO-RK data standards. It includes the annotation of data with external unique identifiers to ontologies, controlled vocabularies, and external databases (UniprotKB, KEGG (Kanehisa et al., 2010), ChEBI (de Matos et al., 2010), EC-Enzyme Classification (http://www.chem.qmul.ac.uk/iubmb/enzyme), BTO-Brenda Tissue Ontology (Gremse et al., 2011), SBO-Systems Biology Ontology (Courtot et al., 2011), GO-Gene Ontology (The Gene Ontology Consortium, 2000), NCBI taxonomy (Sayers et al., 2011) etc.). Finally, the data are transferred to the public online database.

Compliant to the FAIR principles mentioned above data in SABIO-RK are annotated to allow *Interoperability* and *Reusability*. Using the annotations mentioned above SABIO-RK is highly interlinked with other biological databases and ontologies. Currently about 20% of SABIO-RK users enter the database via external links from other databases (e.g. UniprotKB, KEGG, BRENDA, ChEBI).