Proposal for BLAHmuc 2016: Text Mining to Support Data Curation for SABIO-RK

Mariana Neves, Hasso Plattner Institute, Postdam (Germany)
mariana.neves@hpi.de
Maja Rey, Ulrike Wittig, Heidelberg Institute for Theoretical Studies, Heidelberg (Germany)
maja.rey@h-its.org,ulrike.wittig@h-its.org

Biological databases are manually curated knowledge bases which include valuable information that is necessary to support various research tasks for the biomedical domain. This information is usually obtained through careful reading of relevant publications and by manually including the curated data into the database. Text mining can efficiently support biological data curation in many of its steps, i.e., document triage and information extraction, by automatically processing large document collections [1].

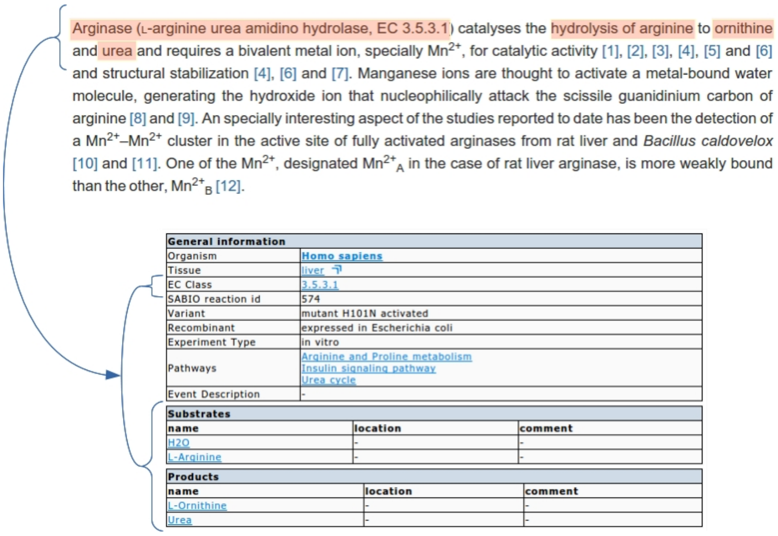
We propose a project for applying two existing text mining tools to support data curation for the SABIO-RK database (cf. Figure 1). SABIO-RK[1] is a curated database containing structured information about biochemical reactions and their corresponding kinetics. It describes participants and modifiers of the reactions, as well as measured kinetic data embedded in their experimental and environmental context. Data in SABIO-RK are manually extracted from published literature, curated by biological experts and enriched with additional information from other databases, biological ontologies, and controlled vocabularies [2, 3].

Only scientific publications containing kinetic data such as Km, kcat or Vmax values are included in SABIO-RK. Therefore a first effort in the proposed project would be to automate the method of document triage by mining full text articles for relevant keywords like kinetic parameters. Up to now there are no text mining solutions which support the SABIO-RK biocuration team in finding and extracting the relevant information out of a publication since they are scattered across the whole article. An implementation of tools highlighting and ideally interlinking relevant information within the publication would support the curators in their daily work.

We will investigate the use of two recently developed tools: (1) Medicate (not yet published), a search engine for navigating through publications from MEDLINE; and (2) TextAI[4], a machine learning-enriched extension of the TextAE annotation tool[2]. Medicate will provide support for document triage while Tex-

[1] http://sabio.h-its.org/

[2] http://textae.pubannotation.org/



Arginase (L-arginine urea amidino hydrolase, EC 3.5.3.1) catalyses the hydrolysis of arginine to ornithine and urea and requires a bivalent metal ion, specially Mn^{2+}, for catalytic activity [1], [2], [3], [4], [5] and [6] and structural stabilization [4], [6] and [7]. Manganese ions are thought to activate a metal-bound water molecule, generating the hydroxide ion that nucleophilically attack the scissile guanidinium carbon of arginine [8] and [9]. An specially interesting aspect of the studies reported to date has been the detection of a Mn^{2+}–Mn^{2+} cluster in the active site of fully activated arginases from rat liver and *Bacillus caldovelox* [10] and [11]. One of the Mn^{2+}, designated ${Mn^{2+}}_A$ in the case of rat liver arginase, is more weakly bound than the other, ${Mn^{2+}}_B$ [12].

General information	
Organism	Homo sapiens
Tissue	liver
EC Class	3.5.3.1
SABIO reaction id	574
Variant	mutant H101N activated
Recombinant	expressed in Escherichia coli
Experiment Type	in vitro
Pathways	Arginine and Proline metabolism Insulin signaling pathway Urea cycle
Event Description	-

Substrates		
name	**location**	**comment**
H2O	-	-
L-Arginine	-	-

Products		
name	**location**	**comment**
L-Ornithine	-	-
Urea	-	-

Figure 1: Curation of biochemical reaction data from the literature for the SABIO-RK database.

tAI will allow manual annotation assisted by supervised learning methods based on little training data. Integration between these two tools is currently under development and will be based on the BioC XML format [5], the current state-of-the-art interoperability format in the biomedical natural language processing community (BioNLP).

The tasks that we plan to perform during the hackathon includes the following: (a) evaluation of the curation needs for the SABIO-RK database; (b) reproducibility of existing data, and (c) curation of new data. We believe that participating in the hackathon will allow us to discuss our project and the technologies that we plan to use with the BioNLP community. Additionally, we also plan to involve the participants in the discussion of standards for the BioC elements, e.g., names for the annotations' attributes (infons).

References

[1] Neves, M. *et al.* Preliminary evaluation of the cellfinder literature curation pipeline for gene expression in kidney cells and anatomical parts. *Database* **2013** (2013).

[2] Wittig, U. *et al.* SABIO-RK - database for biochemical reaction kinetics. *Nucleic Acids Research* **40**, D790–6 (2012).

[3] Wittig, U. *et al.* Data extraction for the reaction kinetics database SABIO-RK. *Perspectives in Science* **1**, 33 – 40 (2014).

[4] Grundke, M. *et al.* TextAI: Enhancing textae with intelligent annotation support. In *Proceedings of the Seventh International Symposium on Semantic Mining in Biomedicine* ((accepted), 2016).

[5] Comeau, D. C. *et al.* Bioc interoperability track overview. *Database* **2014** (2014).