

# Systematic screening of yeast kinetic parameters for metabolic models using a text mining toolbox

Irena Spasic<sup>1,2,\*</sup>, Evangelos Simeonidis<sup>1,3</sup>, Norman Paton<sup>1,2</sup> and Douglas Kell<sup>1,4,\*</sup>

1. Manchester Centre for Integrative Systems Biology, Manchester Interdisciplinary Biocentre, 131 Princess Street, M1 7DN, Manchester, United Kingdom

2. School of Computer Science, The University of Manchester, Kilburn Building, Oxford Road, Manchester, M13 9PL, United Kingdom

3. School of Chemical Engineering and Analytical Science, The University of Manchester, PO Box 88, Sackville Street, M60 1QD, Manchester, United Kingdom

4. School of Chemistry, The University of Manchester, Oxford Road, Manchester, M13 9PL, United Kingdom

\* e-mail: {i.spasic, dbk}@manchester.ac.uk

## Abstract

At the heart of systems biology is a classical inverse problem, which requires the iterative interplay between mathematical/computational simulation of the system of interest and experimental measurements with which the model can be compared. Our goal is to develop and exploit appropriate methods for modelling metabolism. There are two distinctive branches in metabolic modelling: qualitative and quantitative. Qualitative (or structural) models are used to describe the relations between different components. Quantitative descriptions of these relations (e.g. binding and kinetic constants) are used to parameterise such models and represent dynamic aspects of biological systems (e.g. kinetic behaviour) *in silico*. While qualitative models are moderately well known, the new advances in biotechnology have given rise to an enormous production of quantitative omics measurements, which are described in the literature and need to be integrated into qualitative models. Here we present a text mining toolbox, KiPar, developed to aid a modeller in retrieving kinetic parameters of interest from the publicly available scientific literature. The modeller provides pathway-specific input specifications (consisting of pathway-specific enzymes and GO terms) and a list of the types of desired kinetic parameters (specified by their SBO terms). Given this high-level input specification, KiPar employs a range of

integrated bioinformatics strategies (relying heavily on Web Services) to harvest reaction-specific terms (e.g. an enzyme, compounds acting as substrates/products, and the gene encoding the given enzyme) from publicly available biological databases (KEGG, PubChem, ChEBI, SGD, CYGD), as depicted in Figure 1. The problem of terminological variability is further tackled by collecting additional synonyms from ontologies and controlled vocabularies (UMLS, MeSH, SBO, GO). The search terms collected in this manner are used to effect a transition from conceptual space to textual space. Using the collected synonyms, each concept is mapped to the matching documents in NCBI literature databases (PubMed and PubMed Central) using a web service of Entrez, a search and retrieval system, which enables a user to access information from many NCBI databases. Some concepts (or terms representing them) may be too broad and thus need to be removed from the search queries (targeting the pathway-related information). A high number of returned documents is used as an indicator of non-discriminatory concepts, which are removed from further consideration from the information retrieval point of view. A local database is used to store all information gathered about: concepts, synonyms and PubMed [Central] documents. The information collected is used to formulate pathway-related search queries over this database. Each document is scored using a formula combining the number of hits for each of the considered concept classes (i.e. enzymes, compounds, genes, pathways and kinetic parameters) weighted appropriately. The selected documents are presented to the modeller in HTML format. The results produced represent links to the original documents annotated with the matching concepts and quantitative data. The annotation aims to help the modeller determine which types of information each document contains so it can be incorporated into the model, which can be formally represented in SBML.

## References

- Galperin, M. (2005) The molecular biology database collection: 2005 update. *Nucleic Acids Res* **33**: D5-D24.
- Hearst, M.A. (1999) Untangling text data mining. Proc 37th Annual Meeting of the Association for Computational Linguistics (ACL 1999), 3-10.
- Hucka, M. et al. (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* **19**: 524-531.
- Kitano, H. (2002) Systems biology: a brief overview. *Science* **295**: 1662-1664.

Mendes, P. (1997) Biochemistry by numbers: simulation of biochemical pathways with Gepasi 3. *Trends Biochem Sci* **22**: 361-363.

Mendes, P. & Kell, D.B. (2001). MEG (Model Extender for Gepasi): a program for the modelling of complex, heterogeneous cellular systems. *Bioinformatics* **17**: 288-289.

Spasić, I. Ananiadou, S. McNaught, J. & Kumar, A. (2005) Text mining and ontologies in biomedicine: making sense of raw text. *Briefings in Bioinformatics* **6**: 239-251.

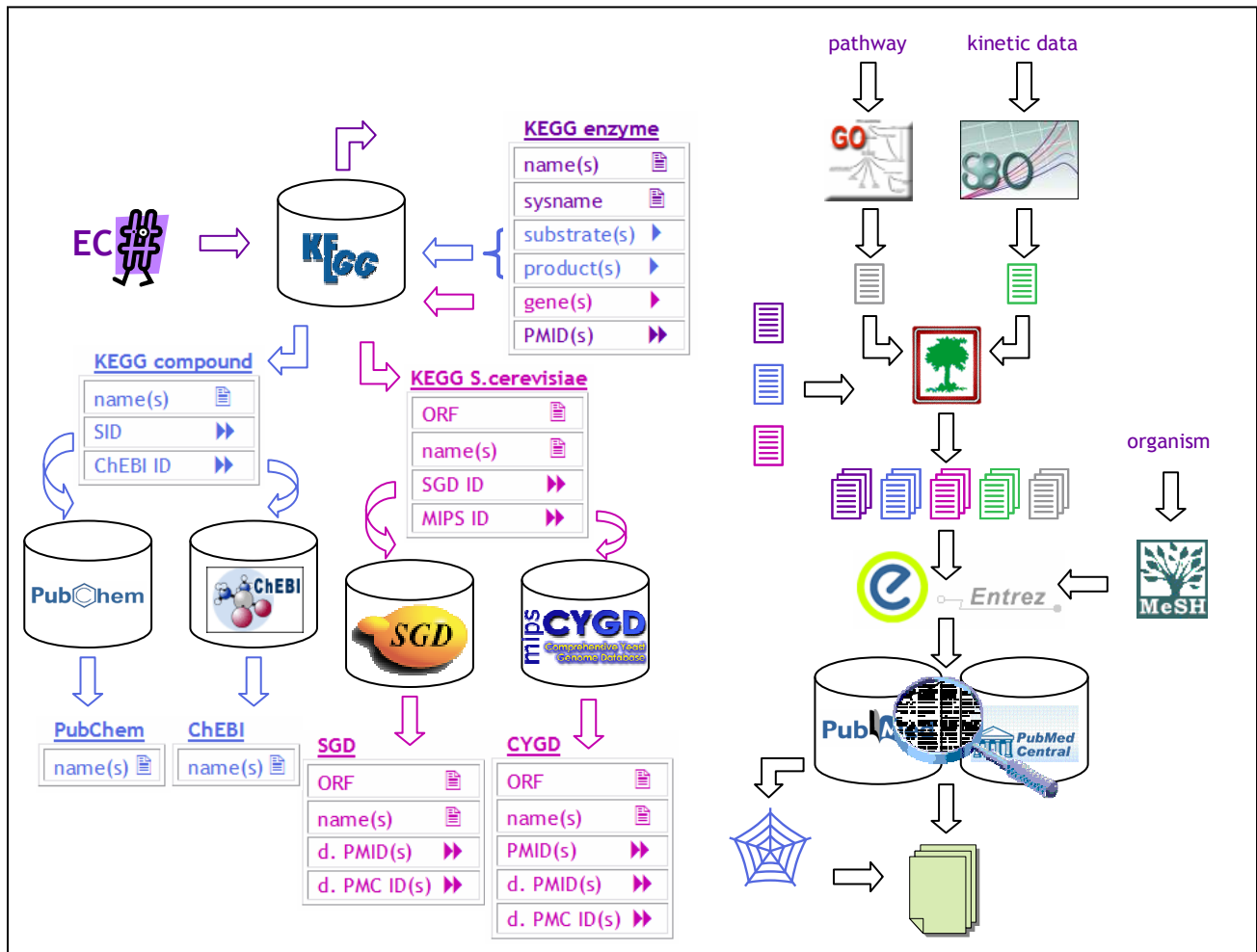


Figure 1: The key processing steps of mining kinetic parameters from the literature