Towards Pathway Prediction and Subcellular Localization by using FraMeTex

Thorben Wallmeyer*, Björn Sommer, Benjamin Kormeier

Bio-/Medical Informatics Department, Bielefeld University/Center for Biotechnology
Universitätsstr. 25, 33615 Bielefeld, Germany

1 Introduction

The visualization and localization of biochemical networks represent major tasks of bioinformatics. In the future, they could be used as additional criteria to find potential -omics interaction partners. Moreover, this visualization and localization can help scientists to find appropriate experiments or to create detailed problem-oriented pathways. Today, we can use database integration and text mining methods for the prediction of biological networks. However, we can use the same applications for the localization prediction of biological networks and their components [1]. Using this information, we are able to model and visualize biochemical networks in 2D or 3D. For this purpose, we use our database integration infrastructure BioDWH with the DAWIS-M.D., as well as ANDCell as the base for the subcellular localization of the CELLmicrocosmos 4.2 PathwayIntegration (CmPI) project [1–4]. Here, we discuss how this approach can be extended by using the new FraMeTex text mining framework.

2 Data integration

High-throughput methods generate, in short time, data concerning the whole genome of an organism. Hence, enormous, heterogeneous and versatile data is produced. The total number of databases, as well the data itself, is continuously increasing. Data distribution and heterogeneity causes big problems in biological data integration. Therefore, we introduced a bioinformatics data warehouse information system DAWIS-M.D. that can be accessed with the CmPI for tailor-made 3D cell and pathway visualization. DAWIS-M.D. is a platform-independent data warehouse approach for metabolic data that is based on the BioDWH data warehouse infrastructure. The information system combines a number of relevant biological databases. For the subcellular localization, BRENDA, Gene Ontology, Reactome, and UniProt are used. The content of the data warehouse is divided into 12 diverse domains, which can be queried via a web-bases graphical user interface that can be accessed with any common web-browser. The DAWIS-M.D. application provides search forms for the following domains: Compound, Disease, Drug, Transcription Factor, Enzyme, Gene, Glycan, Gene Ontology, Pathway, Protein, Reaction and Reaction Pair. Moreover, it is possible to identify relationships and interactions spanning multiple biological domains [3]. DAWIS-M.D. provides also a web-service to the CmPI. Networks can be localized and displayed in a cell environment, edited and extended by using the CmPI software application [1].

3 Text mining

The number of text mining algorithms that can be helpful in reconstructing biological networks from text data is unclear. Therefore, selecting and using the best text mining algorithm for a specific task is challenging. In addition, the heterogeneity of their interfaces and data structures for representing extracted facts is even complicating their

*thorben.wallmeyer@googlemail.com

application. We addressed this problem by developing a powerful framework. FraMeTex (Framework for Medical Text mining and knowledge engineering) offers a highly customizable analysis pipeline that is able to consult different text mining algorithms for knowledge extraction. Furthermore, it supports the adaption of text sources to be analyzed as well as the filtering of relevant data. Extracted knowledge like biological pathways or localizations can afterwards be persisted within a deductive knowledgebase. By using additional rules it is able to identify additional relations. The offered analysis pipeline of FraMeTex is divided into several logical modules that provide a convenient API (see figure). It allows a seamless and flexible integration of its functional modules within almost any application. Each module can be used standalone or chained with other modules to build a complex workflow. We developed a workflow that filters Medline abstracts by keywords and analyzes them with two different text mining algorithms. This enables us to extract reliable pathway and localization data from Medline. GENIAs Named Entity Recognition (NER) web service is responsible for tagging potential biomedical entities within previously selected Medline abstracts [5]. It distinguishes between proteins, cell lines, cell types, DNA and RNA terms. Additionally Enju performs a semantic parsing of the abstracts and returns Predicate Argument Structures (PAS). The analysis results of both algorithms are finally merged. PAS that deal with biomedical entities are most likely supposed to describe biological pathways or localizations.

4 Conclusion and Outlook

The FrameTex localization should be used to try to extend and verify the knowledge of CmPI. We will do this by using the described and a newly developed workflow. The additional workflow will use different text mining algorithms which allows an additional cross check of results. Especially the critical NER process may be performed better by other algorithms.

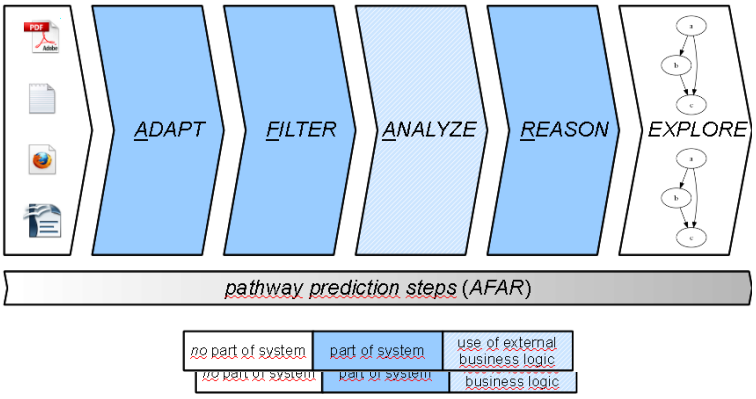

Figure 1: The FraMeTex pipeline

5 References

[1] B. Sommer, B. Kormeier, P. S. Demenkov, P. Arrigo, K. Hippe, Ö. Ates, A. V. Kochetov, V. A. Ivanisenko, N. A. Kolchanov, and R. Hofestädt. Subcellular Localization Charts: A new visual methodology for the semi-automatic localization of protein-related data sets. Journal of Bioinformatics and Computational Biology, 11(1):1340005, 2013.

[2] T. Töpel, B. Kormeier, A. Klassen, and R. Hofestädt. BioDWH: a data warehouse kit for life science data integration. Journal of Integrative Bioinformatics, 5(2):93, 2008.

[3] K. Hippe, B. Kormeier, T. Töpel, S. Janowski, and R. Hofestädt. DAWIS-MD-a data warehouse system for metabolic data. GI Jahrestagung, 2:720–725, 2010.

[4] O. A. Podkolodnaya, E. E. Yarkova, P. S. Demenkov, O. S. Konovalova, V. A. Ivanisenko, and N. A. Kolchanov. Application of the ANDCell computer system to reconstruction and analysis of associative networks describing potential relationships between myopia and glaucoma. Russian Journal of Genetics: Applied Research, 1(1):21–28, 2011.

[5] J.-D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii. GENIA corpus—a semantically annotated corpus for bio-textmining. Bioinformatics, 19(suppl 1):i180–i182, 2003.