Finding Kinetic Data Using Text Mining

Sebastian Schmeier[1], Axel Kowald[1], Jörg Hakenberg[2], Ulf Leser[2], and Edda Klipp[1]

[1] Max Planck-Institute for Molecular Genetics, Berlin
[2] Department of Computer Science, Humboldt-Universität zu Berlin, Berlin

Keywords: text mining, data mining, database

1 Introduction

Quantitative modeling of complex biological systems requires a large number of parameters such as kinetic constants and experimental conditions. Our group is developing a database (*Kinetikon* [7]) for the storage of such information. Most of the data is buried in a constantly and quickly growing number of scientific publications, thus not being available for computation. It is practically impossible for researchers to keep up with this data flood without support of computers. Due to this, we are also developing a text mining system, which aims at developing methods for supporting systems biology researchers in their information retrieval needs. The problem was divided into the retrieval of publications and the extraction of data relevant to kinetic modeling.

2 Method and Results

Two complementary approaches are under investigation: 1.) Classification of publications (see Figure 1A) relevant for kinetic modeling [2, 5] and 2.) automated extraction of biological entities from those publications (see Figure 1B) and semi-automated database filling (see Figure 1C and 1D).

For the first approach, we classify new documents based on linguistic analysis of their content. Text mining algorithms require an intensive preprocessing phase in which documents are retrieved and converted into ASCII text, word and sentence boundaries are recognized, and words are converted to their respective stems. Documents are eventually represented as a high dimensional document vector. Each dimension represents a word stem in the document collection, and the values of the vector components correspond to some measure for the frequency of this word in the document, weighted by the frequency of the word in the entire collection. In this approach we used a *vector space model* [1] for representing texts.

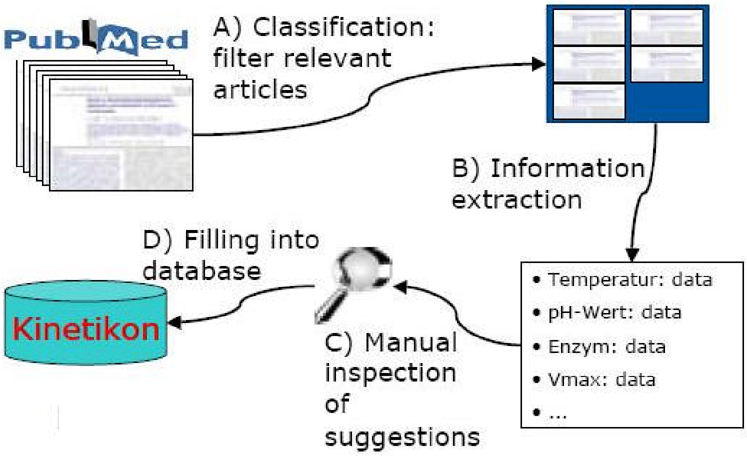
PubMed
A) Classification: filter relevant articles
B) Information extraction
• Temperatur: data
• pH-Wert: data
• Enzym: data
• Vmax: data
• ...
C) Manual inspection of suggestions
D) Filling into database
Kinetikon

Figure 1: Overall architecture of the retrieval step

We approached the problem of finding papers containing kinetic parameters by reformulating it as a classification problem, which we solved using support vector machines (SVM) [6]. The SVM is trained using a set of manually labeled documents. Using this set, the SVM learns to discriminate documents, which are relevant from those, which are not relevant by finding discriminatory properties of the document vectors. Using cross-validation, we find that our classifier reaches 60% precision at 50% recall. Comparing our approach to a pure keyword-style search procedure it was found to outperform the latter by a factor of three in terms of precision.

The main aim of the second task is to set up an automated process, which identifies, marks, and than extracts several entities related to enzyme kinetics, such as types of reactions, names of substrates and products, enzymes, or kinetic parameters, etc. from *PubMed* abstracts.

A dictionary approach will be used to tag enzyme and substance names mentioned in the text. These names will be gathered from the *Kyoto Encyclopedia of Genes and Genomes* (KEGG) [3, 4]. Likewise, we take all reactions stored in KEGG into account, to restrict ourselves to a well-known and accepted set of naming entities. After the entity-tagging step in the articles, we try to combine the tagged information and find relations between them. This means that the semantics and structure of the sentences in the abstract have to be analyzed to find proper rules for the information processing. Finally, we obtain a well defined set of extracted information, which we can represent for example using the *Hypertext Markup Language* (HTML) or which we can transfer semi-automatically to our database *Kinetikon*. Thus, a systems biologist can use this tool to get fast appropriate information about reactions, compounds and kinetic parameters from text sources. Additionally, our tool can also help database curators to improve the efficiency and quality of their work.

References

[1] Baeza-Yates, R. and Ribeiro-Neto, B., *Modern Information Retrieval*, Addison-Wesley, 1999.

[2] Hakenberg, J., Schmeier, S., Kowald, A., Klipp, E., and Leser, U., Finding Kinetic Parameters Using Text Mining, *OMICS* 8(2) Special Issue: Data Mining meets Integrative Biology – A Symbiosis in the Making: 131–152, 2004.

[3] Kanehisa, M., A database for post-genome analysis, *Trends Genet.* 13: 375–376, 1997.

[4] Kanehisa, M. and Goto, S., KEGG: Kyoto Encyclopedia of Genes and Genomes, *Nucleic Acids Res.*, 28: 27–30, 2000.

[5] Schmeier, S., Hakenberg, J., Kowald, A., Klipp, E., and Leser, U., Text Mining for Systems Biology Using Statistical Learning Methods, *3. Workshop des Arbeitskreises Knowledge Discovery (AKKD)*, Karlsruhe, Germany, 125–129, 2003.

[6] Vapnik, V.N., *The Nature of Statistical Learning Theory*, Springer-Verlag, 1995.

[7] http://www.molgen.mpg.de/~ag_klipp/projects/