Automated Recognition and Extraction of Enzyme Kinetic Entities from Text

Sebastian Schmeier

Proposal for a master's thesis

April-September 2005

Referees: Edda Klipp, Ulf Leser, Jörg Hakenberg

Developing accurate kinetic models depends on the availability of kinetic parameters, which can usually only be retrieved in the form of published articles. The extremely large number of publications makes it advantageous to develop tools for automated recognition and extraction of entities related to enzyme kinetics. The idea of this project is to provide a tool that can be used by database annotators as well as the kinetic modeling community to quickly gather appropriate data. The approach that will be used in this work is known as “Template filling” (Gaizauskas et al., 2003).

Objectives

The main aim of the project 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. In addition, a web tool for the *Kinetikon* database has to be implemented that provides the annotation features mentioned before, as a service for both users and database curators (Zhou et al., 2004).

Procedure

The basis of this work will be *PubMed* abstracts in XML format. Easy

accessibility of XML abstracts is the main motivation for choosing them instead of full-text publications, which would require the hard task of PDF-transformation and parsing. One of the first steps toward fact extraction will be to create an enriched corpus of *PubMed* abstracts that contains kinetic data. This will be done by several full text searches on a locally stored copy of *Medline*. Once the documents are gathered, they will form the basis for subsequent analysis steps, such as rule extraction for building regular expressions.

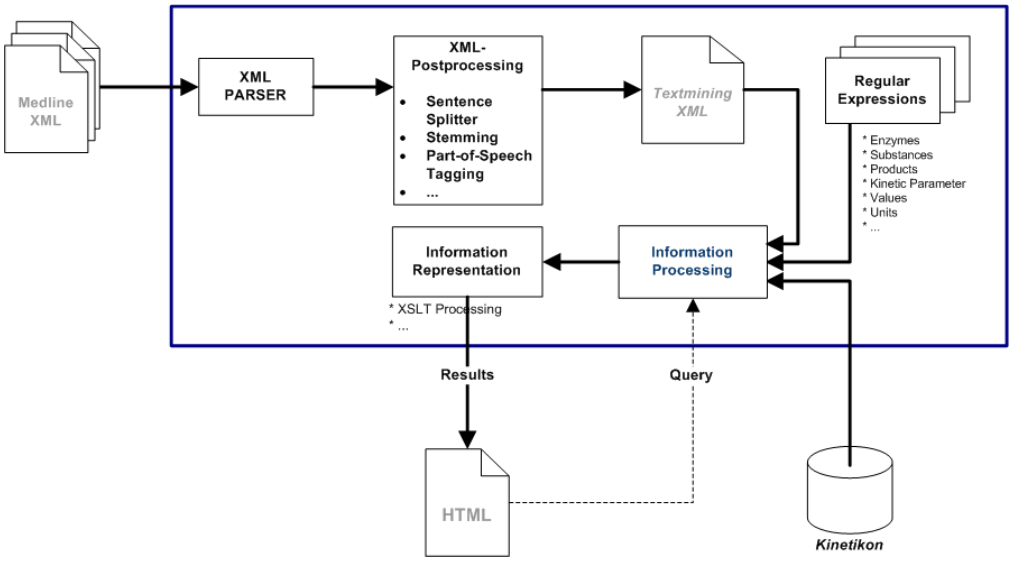

Figure 1: Overview of the information extraction pipeline.

After several post processing steps, such as sentence splitting and POS-tagging, a consistent XML format ("Textmining XML" in Figure 1) is achieved, which provides for example tagged paragraphs and sentences with meta-information such as enumerations of sentences. This ensures that the pipeline could later be used for different data formats, including full text publications. To do so, one only has to set up an appropriate parser to transform the format into "Textmining XML".

Specific regular expression modules will be used to mark different biological entities. To accomplish this aim, a large number of abstracts has to be studied to extract rules for building such regular expressions (e.g., which entities occur in the text; which entities were mentioned together in the same texts or

sentences). A dictionary approach will be used to tag enzyme and substance names mentioned in the text. These names will be gathered from the database *Kinetikon*. Therefore we can easily backtrack the primary-database-ID to their names. Synonyms, which are also stored in *Kinetikon,* will be taken into account as well. In addition, we try to enhance the synonyms by including information stored in several databases such as *UniProt* and *Brenda,* and in available corpora like *iProLink* (Hu et al., 2004). Afterwards, the core module, the "Information Processing" module (Figure 1), is used to combine tagged information and the information provided by *Kinetikon*. The semantics and structure of the sentences in the abstract has to be analyzed to find proper rules for the information processing. For instance, if an enzyme and a substance are mentioned together in one sentence, meaning that the enzyme interacts with the substance somehow, one has to find the proper unique IDs for the reaction, enzyme, and substance, from *Kinetikon*. Such handcrafted rules will be used to extract relations between biological entities in the tagged text and map the information onto the database structure. The tagged XML is then represented as HTML using XSLT processing, and a prespecified form will be filled with all the information extracted from it. After a manual inspection, it is possible to change data fields and transfer them to the database *Kinetikon*. To evaluate the performance of the extracted and combined information the system will be tested on a set of abstracts. The precision of the extracted information will be verified by hand.

References

Gaizauskas, R., G. Demetriou, P. J. Artymiuk, and P. Willett (2003). "Protein structures and information extraction from biological texts: the PASTA system." Bioinformatics **19**(1): 135-43.

Hu, Z. Z., I. Mani, V. Hermoso, H. Liu, and C. H. Wu (2004). "iProLINK: an integrated protein resource for literature mining." Comput Biol Chem **28**(5-6): 409-16.

Zhou, G., J. Zhang, J. Su, D. Shen, and C. Tan (2004). "Recognizing names in biomedical texts: a machine learning approach." Bioinformatics **20**(7): 1178-90.