

# Proposal for a joint tutorial: Information Extraction from Biomedical Literature

Dietrich Schuhmann, Lynette Hirschman,  
Alfonso Valencia

*LION Bioscience AG, Waldhofer Strasse 98, 69123 Heidelberg, Germany*  
*Tel.: +49-6221-4038-179, email: dietrich.schuhmann@lionbioscience.com*

April 10, 2002

## 1 Presenters

This is a joint proposal of three presenters, who come from the academic field and from a business environment. All three of them have a background in computer science, i.e. bioinformatics, computational linguistics or computer science. The majority of them has also long-term teaching experience and numerous publications. Last but not least, all three of them have already held a joint tutorial at the ISMB 2001 in Copenhagen on the same topic.

- Dr. Dietrich Schuhmann  
LION Bioscience AG, Heidelberg, Germany

Dr. Schuhmann is running a research project for text mining in molecular biology, which is funded by the German government. His team has developed various software modules, which are optimized to offer efficient information extraction from biomedical literature.

- Dr. Alfonso Valencia  
CNB, Universidad Autonoma Cantoblanco, Madrid, Spain

Dr. Valencia has long-term experience in information extraction from biomedical literature. His team has developed different prototypes to read out information from Medline. Furthermore, his team has investigate on the integration between bioinformatic tools and information extraction from Medline.

- Dr. Lynette Hirschman  
The Information Technology Center, MITRE Corporation, USA

Dr. Hirschman has long-term experience in the field of text mining. She has worked on the development of tools, which integrate syntactical analysis, semantic representations and handling of nomenclature and ontologies to offer text mining in different domains.

## 2 Outline of the tutorial

The following list shows the parts of the tutorial. Under the section heading follows a short outline of the content of this section. A few headings carry the name of the presenter. Lynette Hirschman is forseen for the section 3. For the remaining sections the other two presenters will give the presentation.

- 1 Overview

The motivation for the text mining tutorial will be discussed: Why is it important to improve document and information retrieval?

- 2 Introduction (D. Schuhmann)

This part explains, why there is a need in text mining. It shows, how text mining fits into the work of the biologist.

- 3 Natural Language Processing: Methods and Evaluation (L. Hirschman or D. Schuhmann)

Some usages of text mining in fields other than biology are considered. There have been activities over the past 20 years to do natural language processing and to extract information from publically available text. Especially in the field of message understanding quite some work has been done to filter out names, events and relationships. These approaches have been useful, but solve problems which are different from the ones in biology.

- 3.1 Information retrieval vs. information extraction
- 3.2 The use of Natural Language Process in fields other than biology

- 4 Term identification and classification for text mining (D. Schuhmann)

Term identification is a very important issue in biology. Biologists use abbreviations and special names to identify their objects. Such objects and their names undergo evolutionary changes. Due to these changes it is difficult and important to have efficient methods to reliably extract the entity names.

This part will cover work and ideas developed by Dr. Sophia Ananiadou from Salford University, Manchester, Uk and her team. She is a project partner of Dr. Schuhmann.

- 4.1 Importance of term identification
- 4.2 Term formation patterns and variability
- 4.2 Methods for Automatic Term recognition
- 4.3 Using term classifications for text mining

- 5 Information extraction in molecular biology (A. Valencia, D. Schuhmann)

Information extraction is used to reduce the text to the facts, that are of interest. There are different approaches. Some use profound natural language processing with special technology (context free grammar parsing, cascaded FSAs) while others use more basic regular expression matching. The advantages and disadvantages of the different techniques are going to be presented and discussed.

- 5.1 Frame-based information extraction (A. Valencia)
- 5.2 IE based on natural language processing (D. Schuhmann):
- 5.2 How is the information represented and how does the biologist work with it?

- 6 Document, sentence and fact classification tools

Apart from information extraction quite some techniques are available, which do extraction of features from the document and classify sentences and facts according to a training set. Different approaches will be presented and evaluated.

- 6.1 Classification techniques available
- 6.2 Tools and products for feature extraction and document classification
- 6.3 Approaches for sentence and fact classification in biology
- 7 Evaluation of extracted results
 

This section deals with a few examples where a data source has been analysed and compared to a reference information source, e.g. a database with such information. In this case it is possible to derive parameters, which prove the quality of the evaluation. On the other side, this evaluation offers ways to understand the limits of the extraction methods and their parts.

  - 7.1 Information retrieval using protein names from DIP (A.Valencia)
  - 7.2 Evaluation of synonym and abbreviation extraction (D.Schuhmann)
  - 7.3. Evaluation of identification of polymorphisms (D. Schuhmann)
- 8 Biological Applications
 

A few applications will be demonstrated, which show the biological benefit. This shows, how text mining can be integrated into the work of the biologist.

  - 8.1 Geisha (A.Valencia)
  - 8.2 Identification of relations between biological entities: gene, proteins, drugs
  - 8.3 Open problems in molecular biology
- 9 Paper, Hints and Links
 

Links and hints to find information.

  - 9.1 Bibliography
  - 9.2 Brief overview on software solutions available from companies

### 3 Legitimation

Information extraction needs expertise from different scientific fields: linguistics, computational linguistics, statistics, computer science, and domain knowledge, i.e. biology. Any approach to this scientific field has to consider state of the art technology of these different fields. This is the major reason for this joint proposal.

#### 3.1 Expected goals

The goals of this tutorial are the following ones:

- To demonstrate solutions to special problems, e.g. automatic term recognition, definition of language patterns, identification of name entities in molecular biology.
- To show results from information extraction: We will select a topic, which has already been worked on to demonstrate the results from information extraction.
- To give insight into the complexity of natural language representations and into the problems associated to natural language processing from scientific text.

- To offer hints at possible integration of such information extraction into other bioinformatic tools.

The presenters will stick to their core topic as they have done last year. In contrast to last year, this year's tutorial will be enriched with examples and evaluations of samples drawn from Medline. Different sets of information extracted from Medline will be completed till summer 2002.

### 3.2 Objectives

The most important objectives are:

- The tutorial will explain the importance of named entities in the field of molecular biology. This shows the need for classification of terms, for automatic term extraction and ontologies.
- The tutorial will explain the different components needed for natural language processing: tagging, term recognition, pattern definition and pattern recognition. Any of these components is a small step towards the extraction of parts of sentences after it has been specified in an abstract representation.
- The importance of annotation and annotated corpora to train information extraction tools.
- Ways to link information from text mining to biological experiments, e.g. expression profiling results.

### 3.3 Motivation

ISMB has become one of the main reference points for the work in text mining in the biomedical literature. Key papers and posters in this area have been presented in ISMB since 1996. The main teams working in this area used to be active in ISMB and the Pacific Symposium Conferences. All this scientific activity will benefit from a complementary activity in teaching and cross-fertilization between more theoretical and biological approaches. Something particularly needed in this area where the fusion of very different approaches has still not completely emerged.

### 3.4 Intended audience

The tutorial is primarily intended for biologists, who want to use such information extraction tools, and want to get an insight into the benefits, the drawbacks and the overhead linked to this technology. Since bioinformaticians and computer scientists offer such tools to biologists, the tutorial is also intended for them.

It is clear, that reasonable progress from information extraction can only be expected as a result from a team effort. The consequence is, that single researchers will not be able to put up a larger approach in information extraction. Therefore, any researcher or biologist interested in text mining would profit from more insight into the complexity of this topic.