

# ISMB'05 Tutorial Proposal

## Mining the Biomedical Literature : State of the Art, Challenges and Evaluation Issues

Hagit Shatkay

### Instructor's Background and Qualifications

Hagit Shatkay is an assistant professor at the School of Computing, Queen's University in Kingston, Ontario. Her research is in the area of machine learning as it applies to biomedical data mining – text mining and information retrieval in particular. She has authored some of the earliest publications in post-genomic biomedical text mining, and is an active member of the biomedical text-mining research community. Prior to joining Queen's University, she was an Informatics Research scientist with the Informatics Research group at Celera/Applied Biosystems, and before that a postdoctoral fellow at the National Center for Biotechnology Information (NCBI). She has a PhD in Computer Science from Brown University, and an MSc and BSc in Computer Science from the Hebrew University in Jerusalem.

*Previous Tutorials:* Hagit Shatkay has previously presented tutorials on biomedical literature mining at ISMB'04, ISMB'03, PSB'03 (the latter two jointly with Ronen Feldman), and in the Bioinformatics Summer School (BISS'2002) at the University of Tübingen in Germany. At BISS'2002 she also presented a tutorial on hidden Markov models (HMMs) and their bioinformatics applications.

### Tutorial Motivation and Goals

Almost every kind of known or postulated information pertaining to genes, proteins, and their role in biological processes is reported in the vast amounts of published literature. The advancement of biological techniques supporting large-scale genomics and proteomics, is accompanied by an overwhelming increase in the amount of literature discussing the biology of genes and proteins. The ability to rapidly survey the literature forms a necessary step in both the design and the interpretation of any large-scale experiment. Moreover, automated text mining offers a yet-untapped opportunity to integrate many fragments of information, gathered by researchers from multiple fields of expertise, into a complete picture exposing the interrelated roles of various genes, proteins and chemical reactions in cells and organisms.

For all these reasons, the past few years have seen a surge of interest in utilizing biomedical text for various purposes, ranging from identifying gene and protein names within sentences and articles, to trying to establish and predict regulatory networks. (See complete sessions in past ISMB and PSB conferences, as well as publications in biological and bioinformatics journals such as *Nature Reviews*, *Nature Genetics*, *Jouranal of Computational Biology* and *Bioinformatics*.) Several text-related disciplines are harnessed in such efforts, including natural language processing, information extraction, and information retrieval.

The objective of this tutorial is to provide a structured introduction to biomedical text mining, from both the biomedical-application and the text-mining perspectives. It will provide background, as well as build and raise the awareness of researchers – who arrived at the area

from either the biomedical or the text-mining domains – about the well-established and the more recently formed theory, techniques, data sources and tools.

The tutorial will present general and biomedical-specific text mining methods. It will discuss the kinds of text these methods can be applied to, and the work that was done so far towards such applications. Existing work in biomedical literature mining will be analyzed and put in the context of the explicit text mining disciplines, as well as examined from a machine learning perspective. Until recently, little has been done about objective assessment of text mining in biology. The tutorial will put a special emphasis on critical assessment, validation and evaluation methods that are used in text mining and information retrieval, and focus on their application in the context of emerging benchmarks. In particular we shall discuss recent evaluation efforts such as the KDD 2002, BioCreAtIvE 2004, and TREC Genomics.

## **Tutorial Length, Level and Target Audience**

**A half-day tutorial.** The overall level will be intermediate with some advanced material. However, the tutorial will provide sufficient background to accommodate both biomedical researchers and computer scientists who are incoming to the field.

The tutorial is intended for a broad audience of bioinformaticians, biologists, medical doctors or other biomedical researchers who use biomedical text sources and are interested in automated text mining. It will also benefit computer scientists working in text analysis (NLP, extraction, retrieval) who are entering the text-mining arena in bioinformatics.

**Prerequisites:** The tutorial is self-contained and general. It is expected though that the audience is comfortable with basic mathematical and statistical concepts and familiar with basic biological terminology.

## **Tutorial Outline**

### **1. Introduction to Methods in Text Analysis**

- a. Preface: A brief introduction to hidden Markov models as they pertain to natural language processing; A basis for a discussion of the methods presented later on
- b. Basics of natural language processing.  
Outline of the basic problems: ambiguity, polysemy, synonymy, large data volume.
- c. Information extraction from printed text – high level.
- d. Information retrieval from printed text – high level.

### **2. Information Extraction and Text Mining**

- a. Text representation, terms, parts of speech, parsing, tagging.
- b. Types of extraction: facts, relations, entities.
- c. Architecture and methods in IE systems: rule-based and machine-learning systems.

### **3. Information Retrieval and Text Categorization**

- a. Indexing by keywords, terms and contents; Boolean queries.
- b. The vector model and term weighting; Similarity queries.
- c. Probabilistic models and latent semantics
- d. Text categorization, clustering and classification

- e. Evaluation methods for extraction and retrieval: Precision, recall and beyond; Evaluation standards – TREC, MUC, ACE.

#### **4. Biomedical Text: Sources and Tasks**

- a. Documents: Pubmed abstracts ; Discussion of full-text vs. abstracts.
- b. Terms, hierarchies and knowledge representation; Ontologies such as GO, MeSH, UMLS.
- c. Annotated databases such as Swiss-Prot, LocusLink, GeneCards.
- d. Tasks: Some examples include identification of relevant literature for a specific gene/protein; connecting genes with disease; grouping genes/proteins by function; reconstructing and predicting gene networks.

#### **5. Research and Systems in Biomedical Literature Mining : Past, Present and Future**

- a. Information extraction for bioinformatics – mining for *facts and relations*:  
Gene location on chromosomes, protein-protein interaction, protein subcellular location, gene-disorder association, discovery of putative indirect and new relationships among entities through *link-analysis*.
- b. Information retrieval for bioinformatics: Text categorization and summarization for finding functional relations among genes, abstracts classification for identifying papers discussing protein-protein interaction, abstracts classification for identifying sub-cellular location of proteins.
- c. Hybrid methods: Combining multiple techniques and data sources.
- d. Evaluation: Creations of evaluation standards and corpora, including the GENIA corpus, the KDD cup, TREC Genomics, BioCreAtIvE. We shall also discuss the critical feasibility assessment for specific tasks.