AM1: ISMB2006 Tutorial

Title: Biological literature mining – from information retrieval to biological discovery

Topic Area: please select preferably one from the following and delete the rest:

- Text Mining

Main Presenter:

- Title: Dr.
- Full name: Lars Juhl Jensen
- Affiliation: European Molecular Biology Laboratory (EMBL)
- Mailing address: Meyerhofstrasse 1, D-69117 Heidelberg, Germany
- Email address: jensen@embl.de
- Telephone number: +49 6221 387296 (work), +49 171 7706178 (cell)
- Fax number: +49 6221 387517
- Homepage URL: http://www.embl.de/~jensen

Other contributors to the tutorial presentation:
Jasmin Saric and Peer Bork have contributed to the presentation.

50-word abstract: To most biologists, hands-on literature mining is currently limited to using PubMed. However, methods for extracting facts from the biomedical literature have improved considerably, and the associated tools will likely soon be used in many laboratories to interpret large-scale experimental data sets and thereby to make biological discoveries.

Tutorial level: Introductory

Prior knowledge required: The participants are expected to have basic knowledge on molecular biology.

Suitability of this tutorial for ISMB:
Biological literature mining is a highly interdisciplinary research topic, but has so far been dominated by computational linguists. As the basic methodologies for automatic extraction of fact from literature are currently maturing, biological knowledge is becoming increasingly important, both for improving the tools that have been developed and, especially, for applying them in combination with the many different types of high-throughput data. The ISMB participants, who typically have a mixed background in computational and biological sciences, are thus in a prime position to benefit from the recent progress in the field of biological literature mining.

Profile of Presenter
I received the M.Sc. degree in chemistry from the Technical University of Denmark (DTU) in 1999. There I continued as a graduate student in Prof. Brunak's group and received the Ph.D. degree in bioinformatics in 2002 for my work on non-homology based protein function prediction. During this time, I also developed methods for visualization of microbial genomes, pattern recognition in promoter regions, and microarray analysis. In 2003 I joined Prof. Peer Bork’s group at the European

Molecular Biology Laboratory (EMBL), where I work on literature mining and on integration of literature and large-scale experimental data.

During my graduate studies I were teaching on several different graduate and undergraduate bioinformatics courses. I have also been giving oral presentations at numerous conferences, meetings, and workshops. In the past couple of months, I have been been giving tutorial-like lectures on biological literature mining, both for industry partners (BioSys, Denmark, November 2005) and for undergraduate students (Technical University of Denmark, January 2006).

Finally, I am the first author on a new review on biological literature mining, which will soon be published in Nature Reviews Genetics.

Tutorial Outline:

- Introduction (~10 min)
    - What is literature mining?
    - Why do we need it?
    - Where do we stand at the moment?
- Information retrieval (~30 min)
    - *Ad hoc* information retrieval
    - Document clustering
    - Text categorization
- Entity recognition and entity identification (~20 min)
    - Recognition vs. identification
    - Machine learning approaches
    - Dictionary-based methods
    - Disambiguation
- Information extraction (~40 min)
    - Statistical co-occurrence methods
    - Combining text categorization and co-occurrence
    - Natural language processing (NLP)
    - Example: extraction of molecular interactions
- Text mining (~20 min)
    - Mining text for overlooked "golden nuggets"
    - Using literature to reveal temporal trends (example: buzzwords)
    - Discovery of global trends from literature
- Text/data integration (~30 min)
    - Molecular networks as an integration platform
    - Discovering candidate genes for genetic diseases
    - Linking genotype to phenotype
- Outlook (~10 min)
    - What is the future of biological literature mining?
    - Which are the main challenges at the moment?