Call for Papers

SIGIR'03 Workshop on Text Analysis and Search for Bioinformatics

August 1, 2003
Toronto, Canada

Bioinformatics is generally defined as the application of information technology to help solve problems in cellular and molecular biology. This covers a broad range of topics from computational models of protein folding to the storage, search, and retrieval of gene sequence data. An emerging topic of interest in this area is automatic analysis of the bio-medical scientific literature. The goals in this area generally include providing easy access to specific textual information from a potentially very large corpus, and automatically extracting information from the text in a form amenable to further, possibly more structured analysis.

The bio-medical literature is full of papers that describe clinical and experimental results, many of which are expressed at the cellular or molecular level as interactions between genes, proteins, and other molecules, or as signal pathways through the cell. Scientists typically describe these results using complex natural language. If this information can be accurately extracted and represented in a more structured form, it can be used to facilitate locating the source document, and, perhaps more interestingly, it can form the basis of a richer knowledge representation and analysis system.

Techniques developed for the scientific literature may also be applicable to the Medical Informatics domain, which includes clinical patient records. Clinical records contain the observations of clinicians as well as the results of medical tests. This may include coded or structured information, but important details often reside in textual notes. Applying text analysis and information extraction techniques can help automate tasks currently performed manually, enable various statistical analyses on individual and large groups of records, and allow connections back to the bioinformatics world. This last task will become more important as personalized medicine (e.g., individually customized drugs) evolves.

Over the last several years interest in the application of text analysis and natural language processing techniques to bio-medical text has grown rapidly, and a research community of biomedical scientists, computer scientists, and computational linguists has emerged. The goals of this workshop are twofold. First, we want to provide a forum where the latest problems, techniques, and results in Bioinformatics for text can be discussed. Second, we want to bring together the Bioinformatics and SIGIR communities to share their insights and results and build on each other's work.

Workshop topics assume a bio-medical text domain and may include:

- Named entity identification (e.g., genes, proteins, interactions, etc.)
- Relationship extraction (e.g., protein-protein interactions)

- "Full-paper techniques" that exploit document structure, tables, figure captions, etc. (i.e., techniques that go beyond PubMed abstracts)
- Indexing and search
- Meta-data extraction and exploitation (e.g., cross-linkage to other structured databases)
- Visualization/results presentation
- Opportunities for standardization (e.g., structured representations of information extracted from text)
- Evaluation criteria and benchmarking data

The workshop will include paper presentations and discussion. The organizers will arrange the presentations and discussion based on the interests of the attendees. All attendees may submit a short abstract on why this topic is of interest to them, and those wishing to make presentations should submit a 5-8 page position paper. The papers should describe recent work and may be preliminary in nature. The organizers will select position papers for presentation, and may invite other presentations as well. See http://www.sigir.org/sigir2003 for more information.

Important Dates

Position paper submission	June 16, 2003
Acceptance notification	July 1, 2003
Final papers due	July 14, 2003
Workshop	August 1, 2003

Submission Instructions

Position papers should be no more than 4000 words (5-8 pages). The standard ACM conference style (see http://www.acm.org/sigs/pubs/proceed/template.html) is recommended. Submissions must be sent electronically in PDF or PostScript format to:

Eric Brown
ewb@us.ibm.com
IBM TJ Watson Research Center
PO Box 704
Yorktown Heights, NY 10598
+1-914-784-7708

Co-organizers

Eric Brown, IBM T.J. Watson Research Center
William Hersh, Oregon Health & Science University
Alfonso Valencia, CNB-CSIC

Program Committee

Christopher Chute, Mayo Clinic
Vasileios Hatzivassiloglou, Columbia University
Lynette Hirschman, MITRE Corporation
Lawrence Hunter, University of Colorado Health Sciences Center
James Pustejovsky, Brandeis University