Natural Language Processing and Information Extraction in Biology

Jun-ichi Tsujii
Department of Information Science
Graduate School of Science
University of Tokyo
7-3-1 Hongo, Bunkyo-ku
Tokyo, 113-0033, Japan

Limsoon Wong
Kent Ridge Digital Labs
21 Heng Mui Keng Terrace
Singapore 119613

A significant proportion of the information required for biology research is currently recorded as free-text such as MEDLINE abstracts and comment fields of relevant reports like GenBank feature table annotations. Such information is important for many types of analysis, including classification of proteins into functional groups, extraction of protein-protein interaction facts, discovery of new functional relationships, maintaining information of material and methods, and increasing the precision and relevance of hits returned by information retrieval systems.

Unfortunately, information in free-text form or in comment fields is difficult for use by automated systems. For example, the annotation of biological function of different proteins is a time-consuming process currently performed by human experts because genome analysis tools encounter great difficulty in performing this task. The ability to extract information directly from MEDLINE abstracts and other sources can directly help in such a task.

Four papers were accepted under peer-review in this session. Early work in automated understanding of biomedical papers tended to concentrate on analytical tasks such as identifying protein names. More recent work considered problems of finding relationships and contexts using simple lexical means. We are delighted that two of the accepted papers considered substantially more advanced natural language processing techniques for extracting protein interaction events. The remaining two accepted papers considered the application of document similarity to improve homology search and to concept discovery.

The paper by Park *et. al.* introduces a bidirectional incremental parsing technique based on combinatory categorical grammar. The paper by Yakushiji *et. al.* uses a full parser with a large-scale general-purpose grammar to analyse the argument structure in Medline abstracts. Both are pioneering papers on the

use of very advanced natural language processing techniques to automatically extract precise protein interaction information and other events from Medline abstracts.

The paper by Chang *et. al.* modifies the PSI-BLAST algorithm to use literature similarity in each iteration of its database search. It shows that supplementing sequence similarity with information from biomedical literature can increase the accuracy of homology search result. The paper by Iliopoulos *et. al.* presents an algorithm for large-scale clustering of Medline abstracts based on statistical treatment of terms, stemming, 'go-list', and unsupervised machine learning. In spite of the minimal semantic analysis, clusters constructed in this paper provide a shallow description of the documents and support concept discovery.

This is the second time a special session at PSB has been devoted to the application of natural language processing and information extraction in bioinformatics. The response to the call for papers and the quality of the submitted papers are extremely encouraging. This is an area with emerging interest and much research and development remains to be carried out. We look forward to seeing results in this field in novel bioinformatics products.

Acknowledgements

We would like to thank the reviewers of this session for their comments: Sophia Ananiadou, Key-Sun Choi, Toru Hisamitsu, Jong C. Park, Ah-Hwee Tan, Loong Cheong Tong, Alfonso Valencia.