Pacific Symposium on Biocomputing 5:488-489 (2000)

Natural Language Processing for Biology

Tatsuhiko Tsunoda
Institute of Medical Science
University of Tokyo
4-6-1 Shirokane-dai, Minato-ku,
Tokyo, 108-8639, Japan
tatsu@ims.u-tokyo.ac.jp

Limsoon Wong
Kent Ridge Digital Labs
21 Heng Mui Keng Terrace
Singapore 119613
limsoon@krdl.org.sg

A large part of the information required for biology research can only be found in free-text form, as in MEDLINE abstracts, or in comment fields of relevant reports, as in GenBank feature table annotations. Such information is important for many types of analysis, such as classification of proteins into functional groups, extraction of protein-protein interaction facts, discovery of new functional relationships, maintaining information of material and methods, increasing the precision and relevance of hits returned by information retrieval systems, and so on.

However, information in free-text form or in comment fields is very difficult for use by automated system. For example, 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.

Five papers were accepted under peer-review in this session. Previous work in automated understanding of biomedical papers tended to concentrate on analytical tasks such as identifying protein names. We are delighted that all five accepted papers considered substantially less constrained problems that involved finding relationships and contexts.

The paper by Baclawski *et. al.* describes a diagrammatic knowledge representation technique called keynets. The rich ontology of the Unified Medical Language System was used to construct and index keynets. Fully using the domain-independent and domain-specific knowledge, keynets parses texts and resolve references to construct new relationships between entities. The paper by Humphreys *et. al.* describes two information extraction applications in bioinformatics based on templates. The first application is EMPathIE, which is able to extract details of enzyme and metabolic pathways from journal articles. The second application is PASTA, which is able to extract information on the roles of amino acids and active sites in protein molecules from journal articles. They clarified how important template matching is in this field, and applied the technique to terminology recognition. The paper by Rindflesch *el. al.* describes EDGAR, a natural language processing system that extracts relationships between cancer-related drugs and genes from biomedical literature. EDGAR draws on a combination of technologies: a stochastic part of speech tagger, an underspecified syntactic parser, a rule-based system, as well as semantics information from the Unified Medical Language System. The metathesaurus and the lexicon in the knowledge base are used to identify the structure of noun phrases in MEDLINE texts. The paper by Stapley *et. al.* describes a system for extracting and visualizing genes that might have related biological function. The system extracts co-occurrences of gene names from MEDLINE documents and predicts their connection based on their joint and individual occurrence statistics. The paper by Thomas *et. al.* presents the customization of an existing information extraction system called Highlight for the task of gathering data on protein interactions from MEDLINE abstracts. They developed and applied templates to every part of the texts and calculated the confidence for each match. The resulting system could be a cost-effective means for populating a database of protein-protein interactions.

This is one of the first opportunities devoted to the application of natural language processing in bioinformatics. The response to the call for papers and the quality of the submitted papers are extremely encouraging. We feel that this is an area with emerging interest and much research and development remains to be carried out. We also feel that results in this field have good potential in generating many novel commercial products.

Acknowledgements

We would like to thank the reviewers of this session for their comments: Kenichiro Fukuda, Satoru Kuhara, Cindy Lim, Osamu Maruyama, Hiroshi Matsuno, Yuji Matsumoto, Satoru Miyano, Jong C. Park, Kanagasabai Rajaraman, Jian Su, Masayuki Takeda, Ah-Hwee Tan, Loong Cheong Tong, Jun-ichi Tsujii.