Pacific Symposium on Biocomputing 7:323-325 (2002)

Literature Data Mining for Biology

Lynette Hirschman
The MITRE Corporation

Jong C. Park
KAIST

Junichi Tsujii
University of Tokyo

Cathy Wu
Georgetown University

Limsoon Wong
Kent Ridge Digital Labs

Even though the number and the size of sequence databases are growing rapidly, most new information relevant to biology research is still recorded as free text in journal articles and in comment fields of databases like the GenBank feature table annotations. As biomedical research enters the post-genome era, new kinds of databases that contain information beyond simple sequences are needed, for example, information on cellular localization, protein-protein interactions, gene regulation and the context of these interactions. The forerunners of such databases include KEGG [1], DIP [2], BIND [3], among others. Such databases are still small in size and are largely hand curated. A factor that can accelerate their growth is the development of reliable literature data mining technologies.

This year is the third time the Pacific Symposium on Biocomputing has devoted an entire session to natural language processing and information extraction for biology. Compared to the last two years, the field has made tremendous strides. Most of the early work on automated understanding of biomedical papers concentrated on analytical tasks such as identifying protein names [4] or relied on simple techniques such as word co-occurrence [5] and pattern matching [6]. Last year, we began to see work based on more general natural language parsers that could handle considerably more complex sentences [7,8]. This year, we see the emergence of more sophisticated natural language technologies that can handle anaphora, as well as extracting a broader range of information.

Six papers were accepted under peer-review out of a total of seventeen submissions reviewed for this session. We briefly introduce them here:

• The paper by Ding *et al.* examines an issue that is fundamental to literature

data mining based on term co-occurrence methods. It systematically compares the impact on recall and precision of mining interaction information when an abstract, a sentence, or a phrase is used as the unit in which to check for term co-occurrence.

• The paper by Hahn *et al.* describes the MEDSYNDIKATE natural language processor designed for acquiring knowledge from medical reports. The system is capable of analysing co-referring sentences and is also capable of extracting new concepts given a set of grammatical constructs.

• The paper by Leroy *et al.* presents the medical parser of the GeneScene system. An interesting aspect of this parser is that it uses prepositions as entry points into phrases in the text, in contrast to earlier approaches which used verbs as entry points. It then fills in a set of basic templates of patterns of prepositions around verbs and nominalized verbs. It also has a set of rules for combining these templates to extract information from more complex sentences.

• The paper by Pustejovsky *et al.* gives us a robust parser for identifying and extracting inhibition relations from biomedical literature. The system is founded on corpus-based linguistics. A particularly interesting feature of this system is its anaphora resolution module. The results reported in this paper focus on *inhibition* relations and demonstrate that it is possible to extract biologically important information from free text with high reliability using a classical approach.

• The paper by Stapley *et al.* is an interesting combination of text processing and machine learning technologies to predict the cellular location of proteins. The performance of the classifier on a benchmark of proteins with known cellular locations is better than a support vector machine trained on amino acid composition and is comparable to an expertly hand-crafted rule-based classifier. [9]

• The paper by Wilbur formalizes the idea of a "theme" in a collection of documents as a subset of the documents and a subset of the indexing terms such that each element of the latter has a high probability of occurring in all elements of the former. An algorithm is then given to produce themes and to cluster documents according to these themes in an optimal way. Results of applying this method to over fifty thousand documents on AIDS are given as an illustration.

The response to the call for papers and the quality of the submitted papers mark this as an emerging field which combines bioinformatics and natural language processing in innovative and productive ways. We find this very encouraging, but we also feel that much research and development remains to be carried out. In particular, the papers in this session illustrate both the

promise of literature data mining and the need for challenge evaluations. On the one hand, they show how current language processing approaches can be successfully used to extract and organize information from the literature. On the other, they illustrate the diversity of applications and evaluation metrics. By defining several biologically important challenge problems and by providing the associated infrastructure (annotated data and a common evaluation framework), we can accelerate progress in this field. This will allow us to compare approaches, to scale up the technology to tackle important problems, and to learn what works and what areas still need work. For this purpose, we have organized an additional special session on literature data mining at this Pacific Symposium on Biocomputing to specifically discuss these challenges and benchmarks.

References

1. H. Ogata et al. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, 27(1):29–34, January 1999.
2. I. Xenarios, D.W. Rice, L. Salwinski, M.K. Baron, E.M. Marcotte, and D. Eisenberg. DIP: The database of interacting proteins. *Nucleic Acid Res*, 28(1):289–291, January 2000.
3. G.D. Bader, I. Donaldson, C. Wolting, B.F. Ouellette, T. Pawson, and C.W. Hogue. BIND–the biomolecular interaction network database. *Nucleic Acids Res*, 29(1):242–245, January 2001.
4. K. Fukuda, A. Tamura, T. Tsunoda, and T. Takagi. Toward information extraction: Identifying protein names from biological papers. *Proc. of PSB*, pp. 707–718, Maui, Hawaii, January 1998.
5. B.J. Stapley and G. Benoit. Biobibliometrics: Information retrieval and visualization from co-occurrences of gene names in MEDLINE abstracts. *Proc. of PSB*, pp. 529–540, 2000.
6. S.-K. Ng and M. Wong. Toward routine automatic pathway discovery from on-line scientific text abstracts. *Genome Informatics*, 10:104–112, December 1999.
7. J.C. Park, H.S. Kim, and J.J. Kim. Bidirectional incremental parsing for automatic pathway identification with combinatory categorial grammar. *Proc. of PSB*, pp. 396–407, 2001.
8. A. Yakushiji, Y. Tateisi, Y. Miyao, and J. Tsujii. Event extraction from biomedical papers using a full parser. *Proc. of PSB*, pp. 408–419, 2001.
9. F. Eisenhaber and P. Bork. Evaluation of human-readable annotation in biomolecular sequence databases with biological rule libraries. *Bioinformatics*, 15:528–535, 1999.