Gaps in Text-based Knowledge Discovery for Biology

In the post-genome era, the emphasis on the use of bioinformatic technology in pharmaceutcal research is increasingly shifting from target identification to target ranking and due diligence [1]. New kind of databases that contain information beyond simple sequences are needed, such as information on subcellular localization, protein interactions, gene regulation, and the context of these interactions. The forerunners of such databases include KEGG, DIP, and BIND. These databases are still small in size and are largely curated by hand. The development of reliable text-based knowledge discovery or literature datamining technologies can accelerate their growth.

Many example applications of text-based knowledge discovery technologies in Biology were described in [3, 4]. These examples demonstrated significant progress both in terms of depth and in terms of breadth. Text-based knowledge discovery in biology has advanced from simple recognition of terms to extraction of interaction relationships from complex sentences. It has also broadened from recognition of protein interactions to a range of problems such as improving homology search, identifying subcellular location, or recognizing themes in the literature. The techniques employed have spanned word co-occurrence statistics, to pattern matching of linguistic constructs in limited contexts, to powerful natural language processing techniques capable of extracting relations that span multiple sentences through the use of coreference. These results mark this as an emerging field that provides a synergistic combination of bioinformatics and natural language processing.

In spite of the enormous potential for the application text-based knowledge discovery techniques to Biology, few of these techniques have made it into routine use to help manage biological information. We list below some issues that need to be addressed in order to accelerate progress and acceptance of the field:

- Abstracts could generally be obtained for free, whereas full papers could generally only be obtained after payment of a fee. It is thus tempting to consider applying a literature mining tool to abstracts. It is crucial to assess, for each type of information that are to be extracted from the literature, whether there is a significant loss if only abstracts are processed, as opposed to full papers. To date, it seems that no single group has investigate this issue to any great extent.

- A number of papers [5, 6, etc.] focused on extracting interactions of proteins, drugs, and other molecules from the literature. They variously reported specificity figures from 60%–90%. The sensitivity of these systems remains an issue. Moreover, these performance studies were done on sample sets that were small in size and also different sample sets were used by different researchers. In order for an impartial assessment and comparison of the performance of these systems, as well as to understand what works and what does not, it is crucial to do a systematic evaluation. Such an evaluation should be based on a biologically important challenge problem, should have extensive training data and blind test data, and should have a clear repeatable evaluation metric. To date, this issue seems to have begun to catch the attention of researchers—the "KDD Cup 2002" competitions have included a task on "In-

formation extraction from Biomedical articles", jointly organized by Alex Yeh of MITRE and the Flybase group of Harvard (see `http://www.biostat.wisc.edu/~craven/kddcup/tasks.html`).

- It is reasonable to assume that the completeness and reliability of the outcome of text-based knowledge discovery in biology are dependent on the input documents. Should the selection of input documents be based on keywords, based on papers chosen by expert biologists, based on well-cited articles and their cited references therein, or based on some other methods? To date, it seems that no single group has considered it in this context.
- It is also unclear how well a text-based knowledge discovery system has to perform in order for it to be useful in biology. To know how good a system has to be, working systems must be given to biologists in user-centered evaluations. To date, it seems no single group has conducted such a study in any extensive way. We acknowledge however that, from experience with previous evaluations in the information retrieval community [2], it is hard to extrapolate from results of batch experiments to predict complex issues of utility and user acceptance.

Many such issues have remained unaddressed to date. Nevertheless, text-based knowledge discovery for biology have significant potential, because even imperfect tools are useful if they give improved functionality at low cost.

References

[1] Boston Consulting Group. A revolution in R&D—The impact of genomics. *BCG Focus*, pages 1–15, June 2001.

[2] W. Hersh, A. Turpin, S.Price, D. Kraemer, D. Olson, B. Chan, and L. Sacherek. Challenging conventional assumptions of automated information retrieval with real users: Boolean searching and batch retrieval evaluations. *Information Processing and Management*, 37:383–402, 2001.

[3] L. Hirschman, J.C. Park, J. Tsujii, L. Wong, and C.H. Wu. Accomplishments and challenges in literature data mining for biology. *Bioinformatics*, to appear.

[4] R. Mack and M. Hehenberger. Text-based knowledge discovery: Search and mining of life-sciences documents. *Drug Discovery Today*, 7(11 Supplement):S89–S98, 2002.

[5] J.C. Park, H.S. Kim, and J.J. Kim. Bidirectional incremental parsing for automatic pathway identification with combinatory categorial grammar. *Proc. Pacific Symposium on Biocomputing*, pages 396–407, 2001.

[6] J. Putejovsky and J.M. Castano. Robust relational parsing over biomedical literature: Extracting inhibit relations. *Proc. Pacific Symposium on Biocomputing*, pages 362–373, 2002.

Limsoon Wong
Laboratories for Information Technology
21 Heng Mui Keng Terrace, Singapore 119613
Email: limsoon@lit.a-star.edu.sg