LINKING BIOLOGICAL LANGUAGE, INFORMATION AND KNOWLEDGE

L. HIRSCHMAN
The MITRE Corporation, 202 Burlington Road, Bedford, MA 01730, USA

C. FRIEDMAN
Columbia University, 622 West 168 St, VC-5, New York, NY 10032,USA

R. MCENTIRE
GlaxoSmithKline Pharmaceuticals, 709 Swedeland Road King of Prussia, PA 19406-0939, USA

C. WU
Georgetown University Medical Center, Box 571414 , 3900 Reservoir Road, NW Washington, DC 20057-1414, USA

Information access is a major challenge for biologists today. Results are pouring in from microarray experiments, more model organisms are being sequenced and results are being used to expedite drug discovery. There is a growing demand to combine information from different sources and across multiple disciplines, such as clinical medicine, pharmacology, and molecular biology. The volume of literature is increasing exponentially, making it almost impossible for biologists to keep up with current research or to find the particular pieces of information that they need. This makes linkage among existing biological information resources a critical problem.

Information resides in the biological literature. It resides in biological databases that distill information on specialized topics. Such databases include genomic databases (Genbank[a]), model organism databases (FlyBase[b], Mouse[c], Yeast[d]) and protein databases (e.g., SWISS-PROT[e], PIR[f]). Linking the literature and the databases are nomenclatures and ontologies that provide standardized references to biological entities and topics – e.g., gene names and symbols, protein names, and standard terminology for diseases and symptoms or biological function.

This session focuses on techniques for managing the linkages among the literature, the databases, and existing terminologies (e.g., UMLS[1]) and ontologies

---

[a] http://www.ncbi.nlm.nih.gov/Genbank/
[b] http://flybase.bio.indiana.edu/
[c] http://www.informatics.jax.org/
[d] http://genome-www.stanford.edu/Saccharomyces/
[e] http://kr.expasy.org/sprot/
[f] http://pir.georgetown.edu/

such as the Gene Ontology.[2] All of these sources mediate the information through natural language. The literature consists of free text, with accompanying figures and tables. The databases are typically a mixture of structured information (for example, DNA, RNA and protein sequences and structures), pointers to the underlying primary source in the literature, and text annotations describing function, form, location, organism, and other relevant information. These annotations are expressed in a controlled vocabulary, or, for more complex information, in short phrases or even short paragraphs of text. Thus typical biological databases contain significant amounts of content expressed in natural language. The nomenclatures and ontologies also rely on natural language – they contain entries expressed as terms (a word or phrase) with an accompanying unique identifier.

There is an urgent need for tools to maintain these linkages. New knowledge is being added very rapidly. From a linguistic point of view, this means that the language of biology is changing. For example, the Mouse Genome issues a weekly report on "Nomenclature Events." For the week of 8/25/01, there were 166 such events, reporting name additions or name withdrawals for the Mouse Genome database[g]. Tools are needed that can recognize mentions of new entities or new relations in text, to capture these new pieces of information for inclusion in ontologies and entry into databases.

Since genes and proteins are often named by their function, they can have lengthy names, which are then abbreviated. Abbreviations are a major source of ambiguity, especially when searching across multiple subdisciplines. Two papers in this session address the issue of identifying abbreviations and their expansions when they appear in parenthetical expressions. The paper by Liu and Friedman uses collocations to determine the appropriate expansion; they report a precision of 96.3% and an estimated recall of 88.5% for 380,000 parenthetical expressions automatically extracted from MEDLINE abstracts. The paper by Schwartz and Hearst addresses the same problem. Their approach identifies candidate "long forms" in the neighborhood of the short form and defines some simple rules to map from long form to short form. This method was evaluated on a small corpus of 1000 abstracts and 871 abbreviation/expansion pairs, with a reported recall of 83% and precision of 96%.

There are many databases (over 280 according to recent estimates), and also multiple nomenclatures and ontologies. To provide a uniform view across information sources, it would be useful to "translate" between different nomenclatures and/or ontologies. This is explored in the paper by Sarkar *et al.* who describe several methods for mapping from GO terminology to UMLS. They report a precision of 89% and recall of 90% by normalizing text strings and removing stop words from both terminologies before matching.

---

[g] (ftp://ftp.informatics.jax.org/pub/informatics/reports/index.html#statistics

Other tools can assist by identifying mentions of relevant biological entities in running text. The names of these entities serve as indices into the article, to characterize the subject of the article. Extraction of more detailed information, such as interaction information, depends critically on the ability to identify the underlying participants in an interaction relation. Two papers in this session describe two different approaches to identifying biological entities. The paper by Hanisch *et al.* describes the process of assembling and correcting a large-scale lexical resource to identify gene and protein names in text; their method analyzes complex names into token classes, which can then be selectively matched against the lexicon. The paper by Narayanaswamy *et al.* describe a linguistically motivated approach to capturing names of key biological entities, including genes, proteins, and chemical names. Their approach uses features, such as case, or suffix to define semantic classes for individual words. These are then assembled into complex terms, which can be labeled with the appropriate class.

Finally, the paper by Glenisson *et al.* makes use of the multiple information sources including terminology lists from databases, reference pointers, and journal articles, to generate clusters of related proteins. The paper presents several evaluation methods of the clusters, including both measures of internal cluster cohesion and an external method based on comparison to a gold standard.

These papers provide significant contributions to linking the literature, databases and ontologies. Progress in this area is accelerating, but evaluation remains a major stumbling block. There are still no standard measures or test sets that can be used to compare the efficacy of one approach to another. Each research group is still forced to spend a significant amount of time creating training and test corpora and defining appropriate evaluation measures. The associated special session will focus on the creation of common resources and challenge evaluations to facilitate cross-system comparison and accelerate progress in this important area.

References

1. Lindberg DA, Humphries BL, McCray AT. The Unified Medical Language System. Methods Info Med. 32(4):281-291 (1993)
2. Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. Nature Genetics 25: 25-29 (2000); also see http://www.geneontology.org/ for further information.