

## Promises of text processing: natural language processing meets AI

We were pleased to see the timely review by Mack and Hehenberger on methods for analyzing biomedical literature [1]. Text mining is becoming increasingly important in biology and medicine. These fields possess large electronically accessible bodies of text that act as the main repositories of new knowledge. For example, the MEDLINE database currently contains over six million abstracts for articles (going back to 1966), and initiatives such as PubMed Central (<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi>) promise the availability of full articles [2].

The opportunity for text analysis to benefit biology is particularly compelling. Experimental biology primarily involves characterizing 'things' such as proteins, cells or tissues, and synthesizing observations to build conceptual models of processes such as reactions, pathways and networks of interactions. As Mack and Hehenberger discuss, the text processing community is building tools that search for relevant documents (information retrieval), identify facts (information extraction) and find implicit patterns in the literature (text mining). These tools are currently useful and will also support the long-term goal of developing computer systems that accelerate research progress. By leveraging the information in the literature, computers might be able to generate hypotheses and propose experiments that are only apparent when combining knowledge from multiple disparate fields. For example, Blagosklonny has argued that the information necessary to understand feedback control of p53 function was implicitly available in

MEDLINE in 1990, 10 years before it was finally elucidated [3].

To realize such goals, the text processing community has looked toward related work in artificial intelligence (AI). This community has been developing data structures (i.e. ontologies and knowledge bases) to encode knowledge in a computable format and algorithms that enable computers to 'understand' it [4]. In some sense, the text processing work is 'bottom up' (looking primarily at the raw data of textual communication) and the AI community is 'top down' (looking at the conceptual cognitive structures that humans use to organize information). As they meet, a powerful new set of capabilities should emerge.

In this context, many laboratories (including our own) are investigating methods of transferring information from the free text of scientific literature into ontologies and knowledge bases. The ambiguities in free text must be reconciled with the rigorous structure required by computers. This problem is unsolved and difficult. Synonyms abound in free text, and there are multiple ways of expressing the same idea. Even more challenging is the fact that knowledge itself is fluid. As our understanding of living systems increases, definitions of words and conceptual paradigms change and adapt.

Therefore, perhaps the easiest solution would be to circumvent text processing entirely and to report knowledge gained from research as structured formats directly. This would be similar to the current practices of depositing sequence information into GenBank, which have enabled computational analysis while enhancing human communication through increased searchability. Unfortunately, this goes against hundreds of years of scientific tradition. Although transmitting data in standard formats is routinely accepted, it is

much more difficult to transmit knowledge (particularly new, partial or speculative theories) in a structured manner. Scientific communication still requires the ability to express subtlety, ambiguity and uncertainty. For the foreseeable future, we are stuck with the legacy of textual communication and will be hard at work developing methods to understand it.

### References

- 1 Mack, R. and Hehenberger, M. (2002) Text-based knowledge discovery: search and mining of life-sciences documents. *Drug Discov. Today* 7 (Suppl.), S89-S98
- 2 Roberts, R.J. *et al.* (2001) Building a 'GenBank' of the published literature. *Science* 291, 2318-2319
- 3 Blagosklonny, M.V. and Pardee, A.B. (2001) Conceptual biology: unearthing the gems. *Nature* 416, 373
- 4 Stevens, R. *et al.* (2000) Ontology-based knowledge representation for bioinformatics. *Brief. Bioinform.* 1, 398-414

**Jeffrey T. Chang\* and Russ B. Altman**

*Department of Genetics, Stanford Medical Informatics  
Stanford School of Medicine  
MSOB X-215  
251 Campus Drive  
Stanford, CA 94305, USA  
\*e-mail: jchang@smi.stanford.edu*