The *Discussion Forum* provides a medium for airing your views on any issues related to the pharmaceutical industry and obtaining feedback and discussion on these views from others in the field. You can discuss issues that get you hot under the collar, practical problems at the bench, recently published literature, or just something bizarre or humorous that you wish to share. Publication of letters in this section is subject to editorial discretion and company-promotional letters will be rejected immediately. Furthermore, the views provided are those of the authors and are not intended to represent the views of the companies they work for. Moreover, these views do not reflect those of Elsevier, *Drug Discovery Today* or its editorial team. Please submit all letters to Joanna Owens, Acting News & Features Editor, *Drug Discovery Today*, e-mail: Joanna.Owens@elsevier.com

The semantic web and biology

The vision of a semantic web was first put forward in the late 1990s by Tim Berners-Lee, the father of the World Wide Web (http://www.w3.org/2001/sw/). Since then, the idea of a semantic web, loosely defined as a web with machine understandable content, has grown to become a significant driving force in the development of standards for describing web content with the use of XML (eXstensible Mark-up Language)-based languages. Several concepts related to this vision are likely to be useful for bioinformatics, in particular for the organization and discovery of knowledge from life science documents. Not only has the XML format largely been adopted by publishers and citation databases (e.g. MEDLINE) for the encoding of documents, but XML-specified ontologies are now also used for defining and describing data in biological databases. The efficient use of an ontology requires wide acceptance of the terminology of concepts and the relationships between concepts described therein. Broad acceptance generally requires either a central defining authority or a large concerted effort in deriving a consensus over the global view.

A central question is whether the different scientific communities that are involved have the will to speak the same language. Even in science, which is generally much less controversial than politics, the necessary level of agreement will take time to achieve. The extent of present knowledge that has been recorded within biology, as refereed literature and its digitally available summary MEDLINE, is organized as free-text. Several efforts, including the Gene Ontology Consortium (http://www.geneontology.org/) and our own effort PubGene (http://www.pubgene.org), are in development to move science toward the semantic web vision. However, such efforts are severely hindered by the lack of digitally usable standards, even for simple items such as gene and protein names, where a name is often imprecise, changes over time or has several synonyms pointing to different entities. Some progress is being made but significant strides forward need new strategies in many disciplines. We suggest two specific areas of bioinformatics where semantic web ideas could prove useful:

(1) Knowledge representation and systems biology. Information stored as facts and relationships could support simulation of larger systems and complex interactions. This requires ontologies and logic. The Gene Ontology effort represents a first demonstration of what could be achieved in knowledge representation in the biology domain. With the advent of increasingly advanced high-throughput data-harvesting techniques, biology is rapidly being transformed to knowledge handling and conceptually to mathematically based systems biology.

(2) Knowledge extraction. It is an important but difficult task to extract existing knowledge. Semantically annotated (tagged) documents could significantly ease the extraction of facts and/or relationships, and would provide a way to increase the precision and specificity of digitally represented biological knowledge. Today, most scientific journals provide digital versions of their content. However, semantic content annotation has not been achieved and would be difficult to realize with the present barriers to literature access. The advent of the Open Archive Initiative (http://www.openarchives.org/), which aims to free the full body of scientific literature from the Gutenberg restraints, could provide opportunities for obtaining useful tags, possibly in association with publishing houses. To our knowledge, no literature annotation initiatives have been declared so we take this opportunity to declare the need for such an initiative.

With the advent of such work, the hope is that the semantic biology web will become more than an idea for the future.

Tor-Kristian Jenssen and Eivind Hovig
Dept of Tumor Biology
Institute for Cancer Research
The Norwegian Radium Hospital
0310 Oslo, Norway

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

  PII: S1359-6446(02)02458-3

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 transfering 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 (http://www.ncbi.nlm.nih.gov/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