Editorial

Text mining — from technology to biological applications

Advances in technology and the growth of the life sciences are generating ever-increasing amounts of data. High-throughput techniques are regularly used to capture thousands of data points in a single experiment, and many well-established low-throughput experiments are performed in thousands of laboratories. The results of these experiments mostly end up in scientific databases and in scientific publications. Although there have been concerted efforts to capture more scientific data in specialist databases, it is generally acknowledged that only 20 per cent of biological knowledge and data is available in a structured format or a database. The remaining 80 per cent of biological information is hidden in the unstructured, free text of scientific publications. The life sciences, as a knowledge-driven discipline, currently produces more publications than any other scientific field. This can be seen in the annual rate of growth of the literature database PubMed, which grows by about 700,000 publications each year. This means that, even for scientists who specialise in a specific subdiscipline, it is difficult to keep track of publications in their field of research.

Over the past decade, the internet, and the World Wide Web, have revolutionised the way scientists access the scientific literature online, and the growing movement towards Open Access publishing promises to further accelerate the use of electronic journals. Alongside this growth in volume and simplified access, there have also been important developments in the sophistication of the search and retrieval tools used by scientists to help them find the articles that interest them quickly and deliver them in a user-friendly way to their desktop. More recently, there has been an increasing interest in methods that do more than retrieve the scientific texts, but mine them for the information they contain in a way that can be used directly to support the research goals of life scientists. Text mining, the subject of this special issue, promises to support many useful activities that are currently challenging to biologists. These include: database curation, the analysis and interpretation of high-dimensional OMICs data (microarrays, NMR, protein chips, etc), building models of biological systems as well as deriving novel hypotheses by combining knowledge from different publications. Although much research on biomedical text mining is going on, only a relatively small number of applications exist that can directly address realistic biological problems. This special issue sets out the current state of text-mining research, starting with basic computational techniques and principles, and continuing with articles that explore the potential applications, even though in some areas these are still in active development and not yet mature, for example, the biomedical semantic web.

The special issue begins with an overview on text mining by Hagit Shatkay, which focuses on information retrieval and its application in biomedicine, and discusses several ways of using information retrieval techniques. Currently, much research is driven by text-mining competitions that are focused on certain types of problems. These competitions enable scientists to compare the success of different approaches in an objective way. Such competitions and evaluation frameworks also have a significant influence on the research of text mining, since in many cases hundreds of scientists tailor their competing systems over several months to address specific tasks of the competitions. The contribution of William Hersh gives an overview on the different text mining competitions. One of the most common evaluation tasks is the identification of named entities in free texts (NER) which is introduced by Ulf Leser

and Jörg Hakenberg. Due to limitations of space, the two contributions on text mining evaluations will appear in the next issue.

The next publication from Irena Spasic *et al*. illuminates the role of ontologies as conceptual frameworks for the semantic representation of textual information, where the principal link between text and an ontology is terminology which maps terms to domain-specific concepts. This paper summarises different approaches where ontologies have been used for text-mining applications in biomedicine. The issue also includes a tutorial on the Gene Ontology, which provides a good overview of the basic principles, applications and tools that are associated with GO.

Ontologies are also the basic data structures behind the semantic web. Berners-Lee, who invented the World Wide Web, has the vision that the semantic web will become the 'next generation' of the internet, where data on the web are defined and linked in a way that can be used by machines. The life sciences, being a discipline that accumulates many different kinds of data in many different places, could be one of the first areas in which the semantic web will realise its potential. The contribution of Sougata Mukherjea surveys current research efforts in this domain and discusses the challenges that must be addressed to make the vision of a biomedical semantic web a reality.

The extraction of biological networks (protein interactions, metabolic networks gene regulation, etc) from text is reviewed in the paper from Andre Skusa *et al*. This requires the integration of several different computational disciplines such as NER, relation mining and graph-based analysis. This paper summarises the most important steps in network extraction and reviews common approaches and solutions to the extraction of biological networks from scientific literature. Such extracted relations can then be used for hypothesis generation, where information is extracted from several publications to generate hypotheses about new interactions. The contribution of Marc Weeber *et al*. presents an overview of such literature-based discovery and discusses methodology, results and online tools that are available to the scientific community.

Whereas the papers up to this point address how information published in textual publications can be mined for information, the paper from Matthias Scherf *et al*. describes the next step in the development of text-mining systems where results from literature analysis are combined with evidence from experiments (eg transcriptome analysis) to improve the accuracy of the results and to generate additional knowledge beyond that recorded in the literature.

In summary, this special issue gives a good overview of text mining as applied to biologically relevant topics. We hope that presenting these papers will encourage the reader to think of text mining not only as a challenging academic discipline, but also as a technique that can draw from general data-mining techniques in order to be applied to realistic biological problems.

Jacob Koehler
Bioinformatics, BAB
Rothamsted Research
Harpenden, UK