European Journal of Human Genetics (2007) 15, 816–817

www.nature.com/ejhg

BOOK REVIEWS

# Mining the literature for relationships between genes

**Computational Text Analysis for Functional Genomics and Bioinformatics**

Soumya Raychaudhuri, Oxford University Press, New York, UK. 2006.
ISBN: 0198567405 (hardcover), 0198567413 (paperback)

Martin Oti

*European Journal of Human Genetics* (2007) **15,** 816;
doi:10.1038/sj.ejhg.5201849

As its title states, this book covers the application of text mining to functional genomics and bioinformatics, which is a little different from the daily activities of the average clinical geneticist. However, while this may not be core material for geneticists, it should still appeal to the more broadly oriented researcher who wants a deeper look into how bioinformatics – and in particular text mining – can be applied to medical research. This is facilitated by the clear writing style used, which is not unnecessarily verbose but not also too dense and dry. The chapters are also well organised, with the sections generally being preceded by a brief tabular overview. It is also well illustrated with black-and-white figures and charts and there are a number of colour plates at the end of the book. Each chapter ends with a list of references for those who absolutely must investigate further. A specialised bioinformatics background is not required to be able to follow the book, though a familiarity with basic mathematics is helpful, and it should be quite accessible to a non-specialist readership.

The book is divided into 11 chapters, although the final chapter is no more than a one-page conclusion. After a couple of chapters introducing functional genomics and text mining, there are several chapters that describe various applications of text mining to functional genomic research. The first chapter discusses the uses of text mining, also pointing out its potential use for candidate disease gene identification. Unfortunately, the work of a few groups in this area goes unacknowledged as the author considered this to be a completely uninvestigated area of research at the time of writing (there were actually at least three web-based tools available at the time the book was written that utilised literature mining to some extent to facilitate candidate disease gene identification – Genes-to-Diseases from Perez-Iratxeta *et al*, BITOLA from Hristovski *et al* and GeneSeeker from Van Driel *et al*).

Chapter two gives some basic biological, statistical and bioinformatic background information, while subsequent chapters cover text mining techniques and applications. Chapter three explains the conversion of text into more usable word vectors, while chapter four discusses the use of text mining to improve sequence searches, for instance to prevent PSI-BLAST search iterations from drifting away from the desired protein family. Other chapters explain how literature mining can be used to automatically identify functionally coherent groups of genes, for gene function annotation and for identifying protein–protein interactions. The application of text mining to microarray data analysis is more comprehensively treated and gets two chapters. The automatic extraction of gene names from text is also not neglected, though surprisingly it is one of the last chapters rather than one of the first.

A potential pitfall of literature mining is that different genes have different amounts of associated literature and many interesting genes may not be mentioned at all. Variation in literature coverage between genes can be compensated for to some extent, and this is also treated in the book, but undescribed genes – of which there are many and which are frequently the most interesting to a researcher – remain outside the scope of this approach.

In summary, this book presents several applications of text mining to functional genomics analyses which should be quite interesting to bioinformaticians and broadly interested geneticists, but which would probably be of less utility to geneticists hoping to apply it to their own work. Text mining can only tell you about what is already known in the literature, and for individual genes, manual literature mining by the geneticist (sometimes referred to as 'searching PubMed') will probably be far more effective than relying on automated approaches. Automated text mining is most effective when analysing large numbers of genes, such as with microarrays, but available microarray analysis software and the various web-based gene functional analysis tools will probably suffice in practice for the geneticist. Nevertheless, this field does hold a lot of promise, and it is still a relatively underdeveloped field in bioinformatics. For those who do have an interest in biological text mining, this book is certainly worth reading■

*Martin Oti is at the Centre for Molecular and Biomolecular Informatics, Nijmegen Centre for Molecular Life Sciences, Radboud University Nijmegen Medical Centre, PO Box 9101, 6500 HB, Nijmegen, The Netherlands*
*E-mail: m.oti@ncmls.ru.nl*