
Preface

The genomic era of biomedicine has been defined by unprecedented growth of data sampling capacity and increasing publication rates discussing it. While such technical advances have heralded a period of intensive scientific discovery, the associated deluge of biomedical literature has reached a volume exceeding the capacity of any researcher to process and assume, critically limiting the ability to realize the full benefit of these findings.

The need to rapidly survey the published literature, synthesize, and discover the embedded knowledge without compromising the integrity of published data is critical if researchers are to conduct “informed” work, avoid repetition, and generate new hypotheses. It is therefore unsurprising that within the scientific community a great deal of interest and effort is focused on the development of techniques that can identify, extract, and exploit this knowledge in a meaningful manner. To do so in an efficient way requires methods that can reduce complexity without compromising the integrity of published data. Consequently, over the last two decades one has seen a surge of publications related to biomedical text mining with the primary intent of aiding scientific researchers cope with the information overload.

This volume of *Methods in Molecular Biology* discusses the multiple facets of modern biomedical literature mining and its many applications in genomics and systems biology. The volume has been designed as a useful bioinformatics resource in biomedical literature text mining for both those long experienced in and entirely new to the field. As such, this book serves two purposes: (a) to provide a timely and comprehensive overview of the current status of this field, including a survey of present challenges; (b) to empower researchers to decide how and when to integrate text-mining tools to facilitate their own research. It comprises 15 chapters including an introductory chapter giving the fundamental definitions and some important research challenges. The 15 chapters are organized in three sections encompassing information retrieval, integrated text-mining approaches, and domain-specific mining methods.

Saffer and Burnett introduce the volume by providing a current perspective on the role of text mining in biomedical research and health care. While addressing the importance of text mining in drug discovery the authors also outline the continuing challenges relating to improved search methodologies, discovering hidden information, and improved rate of discovery.

The first section of the book reviews information retrieval methods:

- *Khare et al.* describe the current state of practice of biomedical literature access and state-of-the-art information retrieval systems in areas related to text and data mining, text similarity search, and semantic search. The authors discuss emerging trends in improving biomedical literature access using portable devices and the adoption of open access policy systems.
- One of the first steps towards making full use of the information encoded in biomedical text is the task of recognizing biological terms, such as gene and protein names. *Bada* provides a detailed survey of the various lexical terminological resources currently available and how best to utilize them to improve entity recognition tasks in biomedical text-mining applications.

- Generating useful Pharmacokinetics (PK), Pharmacodynamics (PD) models to understand Drug-Drug Interaction (DDI) is a critical step during drug development process. However, an appropriate PK ontology and a well-annotated PK corpus which provide the background knowledge for determining DDI have been lacking. To overcome this information gap, Wu et al. developed comprehensive pharmacokinetics ontology capable of encompassing in vitro and in vivo pharmacokinetics studies.
- Once biological entities have been identified within the text fragments, the next step consists of identifying the potential relationships among them. *Pavlopoulos et al.* describe how relationships between bioentities are detected by co-occurrence analysis of single sentences and/or entire abstracts.

The second section outlines how, through the integration of text-mining efforts, hidden or implicit functional information leading to new biological hypotheses generation can be discovered. Key examples are described.

- *Verspoor* describes how the application of novel biomedical text-mining strategies is being utilized for novel protein function prediction, a problem at the forefront of modern biology.
- The advent of high-throughput “omics” approaches to generate data has outstripped the ability to interpret and assign biological relevance. *Heinzel et al.* outline a method for interlinking omic data and biomedical literature towards identifying markers as representatives for a specific disease-relevant pathophysiological (mechanistic) process.
- *Czarnecki and Shepherd* present a practical guideline for constructing a text-mining pipeline from existing code and software components capable of extracting protein-protein interaction networks from full text articles. Their approach demonstrates how literature mining can be used to identify functionally coherent gene groups to facilitate the reconstruction of protein interaction networks in the formulation of novel biological processes.
- The combination of scientific knowledge and experience is the key success for biomedical research. *Jonnalagadda et al.* outline some of the strategies used to identify key scientific opinion leaders in order to support increased collaborative biomedical research.
- *Petric et al.* demonstrate the use of creative literature-mining methods to advance valuable new discoveries from existing literature and provide application examples from their research findings.

The third section of the book focuses on the utility of specialized text-mining applications that are suited to address particular domains or purposes related to drug discovery. The use of literature-mining approaches to extract novel but not yet recognized associations between concepts such as genes, diseases, drugs, and cellular processes can aid the discovery of novel drug targets and increase insight into the mode of action of a drug or find novel applications for known drugs.

- The ability to identify accelerating areas of science for a given disease area highlights scientific advancements in aspects of biology, and offers opportunities for both near and long-term strategy development for innovative medicines with early translational possibilities. *Rajpal et al.* present a literature-mining methodology that evaluates trends, and points to gene-disease associations that can be employed in making various important scientific and strategic decisions during drug development.

- Discovering novel disease genes is a key step in the drug discovery pipeline and requires not only the identification, prioritization, and selection of reliable druggable targets. *Wu et al.* review recent advances in literature- and data-mining approaches for gene prioritization, and describe a computational approach to identify and rank candidate genes by finding associations between known disease genes and disease relevant pathways.
- Drug toxicity remains a major reason why new drug candidates which enter clinical trials fail to ever reach the market. However, there are vast amounts of information in the public domain concerned with pharmacological interactions, biomedical literature, consumer posts in social media, and narrative electronic medical records (EMRs); all of which can be relevant and informative for predicting the safety of novel drugs. *Lin et al.* describe the use of text-mining techniques from these diverse document resources to uncover hidden knowledge and help predict their toxicity profiles.
- Systematically seeking novel associations between existing drugs and new indications has recently emerged as an alternative to the limited productivity issues associated with traditional drug discovery. *Tari and Patel* describe various strategies including application examples that use biomedical literature as a source for systematic drug repositioning.
- In the concluding chapter of this section, *Chen and Sarkar* present a knowledge discovery framework for mining the electronic health records (EHR) to gather phenotypic descriptions of patients from medical records in a systematic manner to identify comorbidities occurring in patients more often than expected. Currently available resources and their caveats are also discussed.

We are very grateful to the authors for contributing to this volume. The editors would like to thank Professor John Walker (Series Editor) for suggesting this project and guiding us through till the end so that we can produce this important scientific work. We hope that the reader will share our excitement to present this volume on “Biomedical Literature Mining” and will find it useful.

King of Prussia, PA
Hitchin, UK

Vinod D. Kumar
Hannah Jane Tipney

Biomedical Literature Mining

Kumar, V.D.; Tipney, H.J. (Eds.)

2014, XII, 288 p. 51 illus., 36 illus. in color., Hardcover

ISBN: 978-1-4939-0708-3

A product of Humana Press