Summary of the BioLINK SIG 2013 meeting at ISMB/ECCB 2013

Karin Verspoor
University of Melbourne
karin.verspoor@unimelb.edu.au

Hagit Shatkay
University of Delaware
shatkay@cis.udel.edu

Lynette Hirschman
MITRE Corporation
lynette@mitre.org

Christian Blaschke
Spanish Institute of Bioinformatics
blaschke.christian@gmail.com

Alfonso Valencia
Spanish National Centre for Cancer Research
valencia@cnio.es

Abstract—The ISMB Special Interest Group on Linking Literature, Information and Knowledge for Biology (BioLINK) organized a one-day workshop at ISMB/ECCB 2013 in Berlin, Germany. The theme of the workshop was “Roles for text mining in biomedical knowledge discovery and translational medicine”. This summary reviews the outcomes of the workshop. Meeting themes included concept annotation methods and applications, extraction of biological relationships, and the use of text-mined data for biological data analysis. All papers are available at http://biolinksig.org/proceedings-online/

I. INTRODUCTION

With the increasing availability of text data related to biology and medicine in the scientific literature, database annotations, the electronic health record, clinical trials data, and health information online, exciting opportunities arise to provide access to pertinent biomedical information and to advance biomedical knowledge. An evolving research direction is the integration of information from diverse data sources, including textual data, to support deeper understanding of biological systems, the genomic basis of disease, and genotype-phenotype relationships. During the BioLINK SIG 2013 meeting, we explored the current state of the field with respect to information extraction from the wealth of available textual sources, and application areas where text mining is deployed to support biological data analysis.

II. MEETING THEMES

A. Concept Annotation Methods and Applications

Many text mining systems incorporate annotation of core concepts such as genes, diseases, and Gene Ontology (GO) terms as a fundamental component. We learned about several case studies making use of a patient-feature matrix derived from the application of the NCBO Annotator tool to clinical text to answer specific biomedical questions. The workshop also included presentations aimed at improved biomedical concept recognition: Neji is an open source framework addressing annotation of a range of biomedical concept categories and Enerto is a method targeted specifically at GO terms.

The source of concept annotations was also discussed. The presentation by Bada, which examined manual annotations of GO concepts specifically relevant to biocuration, highlighted the potential role of key evidential sentences in automatically identifying GO concepts that are likely to lead to curated gene or gene product annotations. Jimeno and Verspoor identify text mining of supplementary material in addition to narrative text as critical for finding information on genetic variants.

Several posters introduced applications making available concept annotations over text. The PubAnnotation tool supports sharing and interoperability of annotations over the biomedical literature. The Botea system provides annotations over PubMed in terms of a Resource Description Framework (RDF) model. PDFJailbreak provides a framework for direct semantic annotation of PDF files.

B. Extraction of Biological Relationships

Beyond recognition of biological entities and concept terms, there is significant interest in extracting biological events, e.g. interactions involving multiple entities or entity components. This involves modeling of myriad linguistic realizations of a given event type, and typically uses either high-precision linguistic patterns or machine learning techniques. Verbs play an important role in connecting entities in event descriptions; Roberts explored differences in the lexical semantic patterns of causal verbs as compared to associative verbs. The SIG also included research aiming to extract some relatively under explored event types: transcription regulation events, including biologically relevant context such as experimental conditions and host tissue, histone modification, and the expression of particular genes in particular cells for the CellFinder database. In addition, a poster was presented describing initial efforts to produce Biological Expression Language statements using text mining.

C. Use of text-mined data for biological data analysis

Our keynote speaker, Lars Juhl Jensen, presented his work on “pragmatic” text mining of clinical records as well as the biomedical literature, in which he highlighted the important role that text mining can play in biomedical applications when it enables large-scale identification of concept co-occurrences. Mining of such relationships among entities can identify significant hidden associations.

Kissa et al developed a novel relatedness measure for chemical compounds which integrates a text-based similarity measure for nominal features (terms from a controlled vocabulary) or free text fields. Hettne et al explore the use...
of background knowledge in pathway databases and literature to make sense of SNP-metabolite pairs from genome-wide association studies [23].

Two contributions to the workshop addressed the interplay between text and structured knowledge resources, from distinct perspectives: Kim and Cohen [24] propose a natural language-based query interface to an underlying structured resources, while Garcia et al. [25] propose aggregation of context extracted from multiple documents into problem-dependent self-describing research objects.

III. Conclusions and Future Outlook

Textual sources such as clinical text and the biomedical literature are rich sources of knowledge, and text mining tools aim to extract and structure that knowledge. The diversity of topics in the BioLINK SIG 2013 reflects a field in transition: the increasing availability of concept annotations, alongside the increasing specificity and complexity of events being tackled for relation extraction, are making possible the use of text-mined data in problems where the primary objective is understanding a biological mechanism or a clinical relationship. We anticipate seeing text mining play an increasing role translational bioinformatics and systems biology models.

IV. Acknowledgements

We thank our keynote speaker, Lars Juhl Jensen for his stimulating talk. We acknowledge the significant contributions of the authors who submitted papers or abstracts, and those who presented their work at the workshop. We would like to thank the members of the BioLINK SIG program committee, listed in full at https://sites.google.com/site/biolinksig2013/organizers who provided excellent reviews of the submitted articles and enabled us to build a nice program for the day.

V. Funding

F1000 Research provided poster award prizes. National ICT Australia (NICTA) supported KV’s time contribution.

References