BIOINFORMATICS MESSAGE FROM THE ISCB
Vol. 31 no. 2 2015, pages 297–298
doi:10.1093/bioinformatics/btu412

Advance Access publication July 14, 2014

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

Karin Verspoor[1,*], Hagit Shatkay[2], Lynette Hirschman[3], Christian Blaschke[4] and Alfonso Valencia[5]

[1]Department of Computing and Information Systems, The University of Melbourne, Parkville VIC 3010, Australia [2]Department of Computer and Information Sciences, University of Delaware, Newark, DE 19716, USA, [3]MITRE Corporation, Bedford, MA 01730, USA, [4]Spanish Institute of Bioinformatics, Madrid, Spain and [5]Spanish National Centre for Cancer Research, Madrid, Spain

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.

Availability and implementation: All articles are available at http://biolinksig.org/proceedings-online/.

Contact: karin.verspoor@unimelb.edu.au

1 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.

2 MEETING THEMES

2.1 Concept annotation methods and applications

Many text-mining systems incorporate annotation of core concepts such as genes, diseases and Gene Ontology (GO) terms (Ashburner *et al.*, 2000) 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 (Shah *et al.*, 2009) to clinical text (Pendu *et al.*, 2013) to answer specific biomedical questions. The workshop also included presentations aimed at improved biomedical concept recognition: Neji (Campos *et al.*, 2013) is an open-source framework addressing annotation of a range of biomedical concept categories, and t4rgot (Jacob *et al.*, 2013) is a method targeted specifically at GO terms.

The source of concept annotations was also discussed. The presentation by Bada *et al.* (2013), 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 (2013) 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 (Kim, 2013). The Biotea system provides annotations over PubMed in terms of a Resource Description Framework (RDF) model (Garcia *et al.*, 2013a). PDFJailbreak (Garcia *et al.*, 2013b) provides a framework for direct semantic annotation of PDF files.

2.2 Extraction of biological relationships

Beyond recognition of biological entities and concept terms, there is a 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 (Cohen *et al.* 2011; Kilicoglu and Bergler, 2012) or machinelearning techniques (Kim *et al.*, 2012; Bjorne *et al.*, 2012). Verbs play an important role in connecting entities in event descriptions; Roberts (Roberts, 2013) explored differences in the lexical semantic patterns of causal verbs as compared with associative verbs. The SIG also included research aiming to extract some relatively underexplored event types: transcription regulation events, including biologically relevant context, such as experimental conditions and host tissue (Leitner, 2013), histone modification (Thomas and Leser, 2013) and the expression of particular genes in particular cells for the CellFinder database (Neves *et al.*, 2013). In addition, a poster was presented describing initial efforts to produce Biological Expression Language (http://www.openbel.org) statements using text mining (Liu *et al.*, 2013).

*To whom correspondence should be addressed.

2.3 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 (Jensen *et al.*, 2012) as well as the biomedical literature (Jensen *et al.*, 2006), 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 (Kissa *et al.*, 2013). Hettne *et al.* explore the use of background knowledge in pathway databases and literature to make sense of single nucleotide polymorphism –metabolite pairs from genome-wide association studies (Hettne *et al.*, 2013).

Two contributions to the workshop addressed the interplay between text and structured knowledge resources, from distinct perspectives: Kim and Cohen (Kim and Cohen, 2013) propose a natural language-based query interface to an underlying structured resources, whereas Garcia *et al.* (2013c) propose aggregation of context extracted from multiple documents into problem-dependent self-describing research objects.

3 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 in translational bioinformatics and systems biology models.

ACKNOWLEDGEMENTS

The authors thank their keynote speaker, Lars Juhl Jensen for his stimulating talk. They acknowledge the significant contributions of the authors who submitted papers or abstracts, and those who presented their work at the workshop. The authors 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 them to build a nice program for the day.

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

Conflict of interest: none declared.

REFERENCES

Ashburner,M. *et al.* (2000) Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Bada,M. *et al.* (2013) Occurrence of gene ontology, protein ontology, and NCBI taxonomy concepts in text toward automatic gene ontology annotation. In: *Proceedings of BioLINK SIG 2013*. Berlin, Germany, pp. 13–19.

Bjorne,J. *et al.* (2012) University of Turku in the bionlp'11 shared task. *BMC Bioinformatics*, **13** (**Suppl. 11**), S4.

Campos,D. *et al.* (2013) Neji: a tool for heterogeneous biomedical concept identification. In: *Proceedings of BioLINK SIG 2013*. Berlin, Germany, pp. 28–31.

Cohen,K. *et al.* (2011) High-precision biological event extraction: effects of system and data. *Comput. Intell.*, **27**, 681–701.

Garcia,A. *et al.* (2013a) Biotea. In: *Proceedings of BioLINK SIG 2013*. Berlin, Germany, p. 59.

Garcia,A. *et al.* (2013b) PDFJailbreak—a communal architecture for making biomedical pdfs semantic. In: *Proceedings of BioLINK SIG 2013*. Berlin, Germany, p. 63.

Garcia,A. *et al.* (2013c) In the pursuit of open science. In: *Proceedings of BioLINK SIG 2013*. Berlin, Germany, p. 61.

Hettne,K.M. *et al.* (2013) Explaining genome-wide association study results using concept profile analysis and the kyoto encyclopedia of genes and genomes pathway database. In: *Proceedings of BioLINK SIG 2013*. Berlin, Germany, p. 60.

Jacob,C. *et al.* (2013) Comprehensive benchmark of Gene Ontology concept recognition tools. In: *Proceedings of BioLINK SIG 2013*. Berlin, Germany, pp. 20–26.

Jensen,L.J. *et al.* (2006) Literature mining for the biologist: from information retrieval to biological discovery. *Nat. Rev. Genet.*, **7**, 119–129.

Jensen,P.B. *et al.* (2012) Mining electronic health records: towards better research applications and clinical care. *Nat. Rev. Genet.*, **13**, 395–405.

Jimeno Yepes,A. and Verspoor,K. (2013) Towards automatic large-scale curation of genomic variation: improving coverage based on supplementary material. In: *Proceedings of BioLINK SIG 2013*. Berlin, Germany, pp. 39–43.

Kilicoglu,H. and Bergler,S. (2012) Biological event composition. *BMC Bioinformatics*, **13** (**Suppl. 11**), S7.

Kim,J. (2013) Pubannotation—a storage system for sharing of literature annotation. In: *Proceedings of BioLINK SIG 2013*. Berlin, Germany, p. 62.

Kim,J.D. and Cohen,K. (2013) Natural language query processing for sparql generation: a prototype system for snomed-ct. In: *Proceedings of BioLINK SIG 2013*. Berlin, Germany, pp. 32–38.

Kim,J.D. *et al.* (2012) The genia event and protein coreference tasks of the BioNLP shared task 2011. *BMC Bioinformatics*, **13** (**Suppl. 11**), S1.

Kissa,M. *et al.* (2013) Towards an integrated compound to compound relatedness measure. In: *Proceedings of BioLINK SIG 2013*. Berlin, Germany, pp. 44–47.

Leitner,F. (2013) Mining cis-regulatory transcription networks from literature. In: *Proceedings of BioLINK SIG 2013*. Berlin, Germany, pp. 5–12.

Liu,H. *et al.* (2013) Automatic generation of BEL statements from text-mined biological events. In: *Proceedings of BioLINK SIG 2013*. Berlin, Germany, p. 58.

Neves,M. *et al.* (2013) Text mining for characterizing cells and tissues. In: *Proceedings of BioLINK SIG 2013*. Berlin, Germany, p. 64.

Pendu,P.L. *et al.* (2013) Case studies in making sense of clinical text. In: *Proceedings of BioLINK SIG 2013*. Berlin, Germany, pp. 48–51.

Roberts,P. (2013) A quantitative analysis of causal and associative events involving genes and proteins. In: *Proceedings of BioLINK SIG 2013*. Berlin, Germany, p. 57.

Shah,N.H. *et al.* (2009) Comparison of concept recognizers for building the open biomedical annotator. *BMC Bioinformatics*, **10** (**Suppl. 9**), S14.

Thomas,P. and Leser,U. (2013) Histoner: histone modification extraction from text. In: *Proceedings of BioLINK SIG 2013*. Berlin, Germany, pp. 52–55.