BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 19 2014, pages 2840–2842
doi:10.1093/bioinformatics/btu383

Databases and ontologies

Advance Access publication June 17, 2014

Literome: PubMed-scale genomic knowledge base in the cloud

Hoifung Poon*, Chris Quirk, Charlie DeZiel and David Heckerman

Microsoft Research, One Microsoft Way, Redmond, WA 98052, USA

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: Advances in sequencing technology have led to an exponential growth of genomics data, yet it remains a formidable challenge to interpret such data for identifying disease genes and drug targets. There has been increasing interest in adopting a systems approach that incorporates prior knowledge such as gene networks and genotype–phenotype associations. The majority of such knowledge resides in text such as journal publications, which has been undergoing its own exponential growth. It has thus become a significant bottleneck to identify relevant knowledge for genomic interpretation as well as to keep up with new genomics findings.

Results: In the Literome project, we have developed an automatic curation system to extract genomic knowledge from PubMed articles and made this knowledge available in the cloud with a Web site to facilitate browsing, searching and reasoning. Currently, Literome focuses on two types of knowledge most pertinent to genomic medicine: directed genic interactions such as pathways and genotype–phenotype associations. Users can search for interacting genes and the nature of the interactions, as well as diseases and drugs associated with a single nucleotide polymorphism or gene. Users can also search for indirect connections between two entities, e.g. a gene and a disease might be linked because an interacting gene is associated with a related disease.

Availability and implementation: Literome is freely available at literome.azurewebsites.net. Download for non-commercial use is available via Web services.

Contact: hoifung@microsoft.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on May 1, 2014; revised on June 2, 2014; accepted on June 5, 2014

1 INTRODUCTION

The cost of sequencing a full human genome has recently dropped below $1000 (www.illumina.com/systems/hiseq-x-sequencing-system.ilmn). The impending broad availability of genomics data holds the promise of making personalized medicine a reality. There has been a rapid rise of interest in applying sequencing in clinical applications such as cancer treatment and pediatrics, as well as in consumer genomics such as 23AndMe. However, it remains a significant challenge to interpret genomics data for identifying causal genes for diseases and proposing novel drug targets. For complex diseases like cancer, the causal mechanism is complex and involves the synergistic interactions of many genetic components (Hanahan and Weinberg, 2011). Consequently, there has been increasing interest in adopting a systems approach to boost the signal-to-noise ratio by integrating prior knowledge such as gene networks (Ideker *et al*., 2011). Of equal importance is to keep up with new genomics findings and automate the process of genomic interpretation, as the evidence for diagnosis and treatment has grown from a handful of symptoms into hundreds or even thousands of genome-scale markers.

The majority of biomedical knowledge resides in free-text publications from online repositories such as PubMed (www.ncbi.nlm.nih.gov/pubmed), which currently contains >22 million published articles and is growing by >1 million each year. This makes it a priority to automate curation of genomic knowledge for scientific and translational research.

In the Literome project, we have developed a natural-language processing (NLP) system to automatically extract biomedical entities and relations from PubMed abstracts and made the resulting knowledge base available on the Internet. Currently, Literome focuses on entities and relations most pertinent to genomic medicine: genes, single nucleotide polymorphisms (SNPs), diseases and drugs, related through genomic interactions (e.g. transcription factors and kinases) and genome-wide associations (e.g. SNP–drug or gene–disease associations). Users can browse and search the resulting knowledge base in the Literome Web site, which will be kept up to date as new abstracts become available. The Web site and the knowledge base are deployed on Azure (azure.microsoft.com). It is straightforward to apply Literome to full texts (e.g. PubMed Central). We will pursue this in future work.

2 KNOWLEDGE BASE AND FEATURES

The Literome knowledge base currently features two types of knowledge particularly relevant for genomic medicine. First, it curates directed genic interactions by extracting transcriptional and regulatory events from text. Each interaction consists of three parts: a type indicating the regulatory direction (positive, negative or unspecified), a theme that undergoes a specific change (e.g. the gene being transcribed into messenger RNA or the protein being phosphorylated) and a cause that brings forth the change (the transcription factor or the kinase). Second, it extracts genotype–phenotype associations from abstracts to curate findings from genome-wide association studies (GWAS). Genotype refers to either a gene or an SNP. Phenotype refers to the presence of a disease or a drug reaction. An association is a potential correlation between the two. Literome retains the provenance information, so that users can read the sentences or abstracts behind each extraction for verification and context.

*To whom correspondence should be addressed.

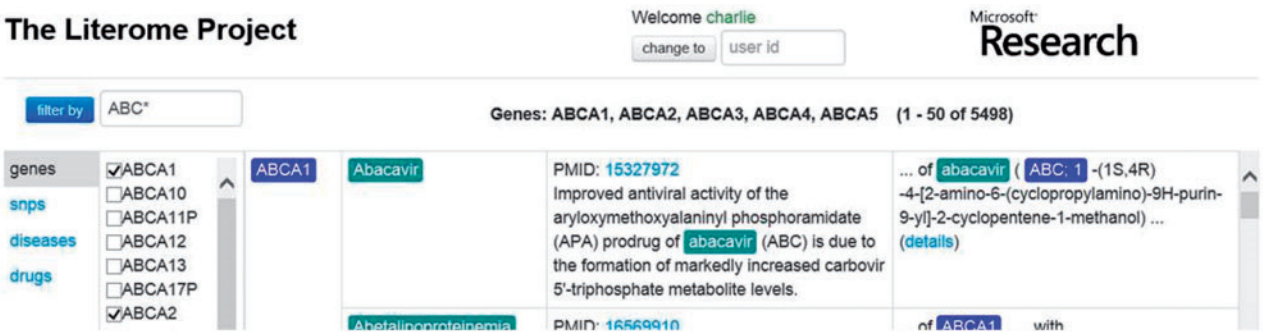

Fig. 1. Screenshot of Literome browsing: left pane allows users to explore entities by category, right pane shows association findings with provenance

Browsing: Browsing is available for users interested in exploratory studies. Figure 1 shows a sample browsing process: the user selected entity type 'genes', entered a query, received a list of genes, selected a subset (ABCA1, . . . , ABCA5) and was shown drugs and diseases associated with that gene, along with the PubMed abstracts, viewable from 'details'.

Searching: Users can search for relations between two specific entities. These could be two genes (as in pathway interactions) or a genotype and a phenotype (as in genome-wide associations). All instances supporting a relation are shown for users to review.

Reasoning: Users can search for relations that are not explicitly mentioned in text but can be inferred from other relations. For example, if A regulates B, and B regulates C, then it is likely that A can regulate C. Search results for genic interactions can thus be augmented by transitivity. Likewise, GWAS search results can be augmented by considering associations among related genotypes and phenotypes. Currently, Literome uses these types of relation expansion: SNP–gene (SNP belongs to gene region), gene–gene (directed interaction between the two genes) and disease–disease [neighbors or cousins in the Disease Ontology (Schriml *et al*., 2012)].

Feedback: Users can reinforce, refine or contradict each extraction by clicking buttons next to it. As Literome is based on a machine-learned system, feedback could potentially be used to improve the quality of future extraction.

3 APPLICATIONS

GWAS Imagine a researcher has found a tentative association between SNP `S` and disease `D` in a new study. This evidence alone might not be meaningful, but if prior findings suggest associations between `S` and `D` (or between related SNPs and diseases), the possibility that this finding is significant substantially increases. In addition to association findings, Literome also catalogs genic interactions and known phenotype relations such as the Disease Ontology, enabling powerful related search (e.g. from SNP to nearby gene, then to interacting gene, finally to related disease).

Network analysis Literome offers a comprehensive and up-to-date collection of published genic interactions, including causality and direction (e.g. A upregulates B versus B downregulates A). Such a gene network can serve as input in a systems biology or network-based approach. This collection of published findings can also be used to validate *de novo* pathway predictions.

Knowledge curation Literome can be used to increase productivity in manual curation. By biasing toward recall and eliminating the vast majority of obvious non-findings, Literome enables curators to concentrate on the small fraction of text likely to contain relevant findings. Moreover, it is much more efficient to curate by correcting errors with a button click, rather than annotating from scratch.

4 DISCUSSION

Literome used an NLP pipeline to extract entities and relations from text; see Supplemental Method for details.

The Literome knowledge base has several distinct advantages over existing related resources. With automatic curation, Literome offers more comprehensive and up-to-date coverage compared with manually curated databases such as Welter *et al*. (2014) and Whirl-Carrillo *et al*. (2012). By extracting interaction types and directions, Literome provides important detailed information generally missing in network resources such as Thomas *et al*. (2012). Like Landeghem *et al*. (2013), Literome recognizes much more diverse patterns for genic interactions than systems such as Chen and Sharp (2004) that used a small verb lexicon. Additionally, by assimilating both genetic and medical knowledge, such as pathway interactions and GWAS findings, Literome enables powerful reasoning that is only possible by integrating the two. Finally, sharing the growing interest in accelerating curation with automatic tools (Wei *et al*., 2012), Literome includes a feedback mechanism that could drive eCuration.

In conclusion, Literome provides a cloud-based knowledge base for genomic medicine, featuring knowledge automatically curated from PubMed abstracts by an NLP system. It offers powerful search and exploration capabilities as well as a feedback mechanism to continuously improve annotation and extraction.

ACKNOWLEDGMENT

We give warm thanks to Jennifer Listgarten, Anthony Gitter, Lucy Vanderwende, Pallavi Choudhury, Bob Davidson, and Simon Mercer for their helpful comments.

Conflict of Interest: none declared.

REFERENCES

Chen,H. and Sharp,B.M. (2004) Content-rich biological network constructed by mining PubMed abstracts. *BMC Bioinformatics*, **5**, 147.

Hanahan,D. and Weinberg,R.A. (2011) Hallmarks of cancer: the next generation. *Cell*, **144**, 646–674.

Ideker,T. *et al.* (2011) Boosting signal-to-noise in complex biology: prior knowledge is power. *Cell*, **144**, 860–863.

Landeghem,S.V. *et al.* (2013) Large-scale event extraction from literature with multi-level gene normalization. *PLoS One*, **8**, e55814.

Schriml,L.M. *et al.* (2012) Disease ontology: a backbone for disease semantic integration. *Nucleic Acids Res.*, **40**, D940–D946.

Thomas,P. *et al.* (2012) Geneview: a comprehensive semantic search engine for PubMed. *Nucleic Acids Res.*, **40**, W585–W591.

Wei,C.-H. *et al.* (2012) Accelerating literature curation with text-mining tools: a case study of using pubtator to curate genes in PubMed abstracts. *Database*, **2012**, bas041.

Welter,D. *et al.* (2014) The NHGRI GWAS catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.*, **42**, D1001–D1006.

Whirl-Carrillo,M. *et al.* (2012) Pharmacogenomics knowledge for personalized medicine. *Clin. Pharmacol. Ther.*, **92**, 414–417.