BIOINFORMATICS APPLICATIONS NOTE Vol. 21 no. 10 2005, pages 2560–2562
doi:10.1093/bioinformatics/bti381

Data and text mining

Using the biological taxonomy to access biological literature with PathBinderH

J. Ding[1,2], K. Viswanathan[2,3,4], D. Berleant[1,2,5,*], L. Hughes[1,2], E. S. Wurtele[2,5,6], D. Ashlock[2,5,7,†], J. A. Dickerson[1,2,5], A. Fulmer[8] and P. S. Schnable[2,4,6,9]

[1]Department of Electrical and Computer Engineering, [2]Iowa State University, Ames, IA 50011, USA, [3]Department of Industrial Engineering, [4]Center for Plant Genomics, [5]Virtual Reality Applications Center, and Lawrence Baker Center for Bioinformatics and Biological Statistics, [6]Department of Genetics, Development, and Cell Biology, [7]Department of Mathematics, [8]Miami Valley Laboratories, The Procter & Gamble Co., 11810 E. Miami River Rd., Ross, Ohio 45061, USA and [9]Department of Agronomy

Received on November 29, 2004; revised on February 20, 2005; accepted on March 7, 2005
Advance Access publication March 15, 2005

ABSTRACT

Summary: PathBinderH allows users to make queries that retrieve sentences and the abstracts containing them from PubMed. Another aspect of PathBinderH is that users can specify biological taxa in order to limit searches by mentioning either the specified taxa, or their subordinate taxa, in the biological taxonomy. Although the current project requires this function only for plant taxa, the principle is extensible to the entire taxonomy.

Availability: www.plantgenomics.iastate.edu/PathBinderH. Source code and databases on request.

Contact: berleant@iastate.edu

Supplementary information: A tutorial is at the tool Website. A longer paper is at class.ee.iastate.edu/berleant/s/paperPathBinderHreport.pdf

INTRODUCTION

Automated text mining in biology has grown dramatically in recent years, fueled by its potential to support efforts to understand and control biological processes (Barnes, 2002; Blagosklonny and Pardee, 2002; Dickman, 2003). Mined information can be used for such applications as gene annotation, curation support and improved literature access.

The goal of mining the biological literature for interactions has inspired a number of efforts to generate public resources. Major resulting systems include MedMiner (Tanabe *et al*., 1999), PreBIND (Donaldson *et al*., 2003) which feeds the curated BIND (Bader *et al*., 2002), Arrowsmith (Swanson, 2004, http://kiwi.uchicago.edu) and iHOP (http://www.pdg.cnb.uam.es/UniPub/iHOP) (2004). Automatic mining need not be labor intensive, and thus can provide resources that are larger than online interaction database projects relying on manual input of interactions, such as MINT (Zanzoni *et al*., 2002), DIP (Marcotte *et al*., 2001; Xenarios *et al*., 2002) and HPRD (Peri *et al*., 2003).

Few mining and retrieval systems integrate the biological taxonomy (sometimes termed the Linnaean taxonomy owing to its invention by Carl Linnaeus circa. 1735) into their operation. Yet, not using the biological taxonomy in biological literature access hinders the full utilization of the literature by systems biologists, students and others. It also hinders computer-generated gene annotation using passages from the literature, because passages typically apply to particular classes of organisms. We expect the integration of the biological taxonomy into literature access to have important benefits for two complementary reasons. First, it will help to reduce the amount of taxonomically irrelevant information brought to the attention of users. Second, it becomes possible to automatically find documents relevant to various different, but closely related, taxa without explicitly specifying all taxa names of interest.

BIOLOGICAL TAXONOMY-BASED LITERATURE MINING AND ACCESS

PathBinderH incorporates the biological taxonomy into literature mining and retrieval using the plant kingdom as an example. With PathBinderH, users can search for sentences, each of which not only matches a query but is also in a PubMed entry (and thus, is embedded in a context) indicating that the sentence is likely to be relevant. A PubMed entry contains key information about an article in the biological literature, usually including its abstract and MeSH (2005, http://www.nlm.nih.gov/mesh/meshhome.html) descriptor terms assigned to it by the US National Library of Medicine (UMLS, 2004, http://www.nlm.nih.gov/research/umls). A PubMed entry is deemed to indicate that the sentence is relevant if it contains a relevant plant taxonomy term anywhere in it, such as in its title, abstract or MeSH descriptors. A relevant plant taxonomy term is one specified by the user, its synonym or a taxon subordinate to it in the biological taxonomy. For example, specifying ‘Poaceae’ (the grass family) as the taxon will cause PathBinderH to search for sentences matching a given query within a set of PubMed entries that mention Poaceae, a synonym of Poaceae or any taxon subordinate to Poaceae (i.e. below it in the taxonomic tree), such as wheat, rice, maize or corn. Likewise, a user might wish to search PubMed abstracts relevant to Viridiplantae (green plants), thus considering

*To whom correspondence should be addressed.

†Present address: The Department of Mathematics and Statistics, University of Guelph, Ontario, Canada N1G 2W1.

Arabidopsis and many other organisms besides Poaceae. Since the MeSH headings assigned to a paper rarely if ever, give taxa either subordinate or superordinate to those mentioned in paper, PathBinderH must deduce these itself.

Approach

To address this important need, PathBinderH uses the biological taxonomy database at the NCBI Entrez Taxonomy Homepage (2004) portal available at: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Taxonomy. The NCBI biological taxonomy database contains the names of species and other taxa, their synonyms and their locations in the taxonomic tree. For each taxon in the plant portion of the biological taxonomy, a list of PubMed entries that mention the taxon was automatically generated by querying PubMed with its scientific and common names. Then each PubMed abstract is indexed under any plant taxa it explicitly mentions as well as, additionally, any plant taxa above (i.e. superordinate to) any explicitly mentioned ones. This enables restricting the retrieval of sentences matching the query to those that occur in a PubMed entry mentioning a user-specified taxon or any of its subordinate taxa.

For example, a PubMed abstract mentioning maize (or equivalently *Zea mays* or corn) will also be indexed under its superordinate taxa, which include, e.g. *Zea*, Andropogoneae, Panicoideae, Poaceae, Poales, Magnoliophyta, Tracheophyta, Embryophyta and Viridiplantae. Later, if a user specifies some taxon superordinate to maize, such as one of these, the abstract will be searched for sentences matching the query. If the user makes no taxonomic specification, then all of PubMed is searched up to February 2005 (as of the date of this writing). Regular updates are planned.

Analysis of an example

Analysis of a sample query helps to illustrate the advantages that the biological taxonomy-aware, sentence-based access provided by PathBinderH can have compared with standard literature access, such as that provided by PubMed. For this query, the taxon Viridiplantae was specified, restricting retrieval to sentences in PubMed entries that mention a species or other taxon in the green plant portion of the biological taxonomy. Next, a query was made using the terms ‘embryo’ and ‘development’. The query had two terms because the PathBinderH user interface currently permits only two-term queries, a design decision made to support searches for interactions. Interactions between pairs of biomolecules are usually stated within a single sentence (Ding *et al*., 2002, http://psb.stanford.edu) and preliminary results suggest that this finding also holds for pairs involving other biological entities. This query, made in summer of 2004, caused retrieval of sentences (defined to include titles) that both matched the query and were in taxonomically eligible PubMed entries.

PathBinderH returned 651 sentences contained in 542 PubMed entries. The standard PubMed interface returned 890 entries in response to the query `embryo[Text Word] development [Text Word] plant[Text Word]`. This query was chosen as the most comparable PubMed query. The plainer query “`embryo development plant`” causes PubMed to apply a complex (but not taxonomically aware) query expansion process because of the absence of the `[Text Word]` qualifier. The query `plant` “`embryo development`” immediately rules out sentences in which the terms `embryo` and `development` have intervening words or are in a different order, yet which the user would probably want to see. An examination of the 542 PubMed entries containing sentences returned by PathBinderH and the 890 returned by PubMed revealed the following salient points.

- Only 159 entries were returned by both PathBinderH and PubMed, perhaps a surprisingly small overlap. The PubMed query missed the other 383 entries found by PathBinderH because these entries contained a plant taxon name other than ‘plant,’ a recall of just 0.29 of the PathBinder result.
- When we enabled PubMed to increase its recall such that it could find all 542 entries found by PathBinder by eliminating the term ‘plant’ from the query it returned not only the 542 desired entries but many others—a total of over 56 000. The vast majority of these were not about embryo development in any kind of plant. The dilution of useful entries by this large number of entries not related to plants indicates a precision of the order of 1%, which is quite low.

In this example, PathBinderH found a significant number of relevant sentences which were in PubMed entries that either PubMed did not find or could find only at the price of also returning tens of thousands of irrelevant entries and consequently a precision that would typically be considered unacceptably low.

CONCLUSION

PathBinderH provides sentence-focused access to the large PubMed literature database. It also demonstrates the use of biological taxonomy to focus searches. An example illustrated the significant effect this can have. Although the current implementation focuses on plants, the principle extends to the entire biological taxonomy. In addition, although PathBinderH retrieves sentences, the principle of biological taxonomy-aware retrieval could be similarly applied by the standard Entrez interface to PubMed provided by NCBI (PubMed, 2004, http://www.ncbi.nlm.nih.gov/PubMed/) or by any other typical system for literature access within a biological context.

ACKNOWLEDGEMENTS

This research was funded in part by PubMed Central has very recently done this funding from The Procter & Gamble Co. Support was also provided by Hatch Act and State of Iowa funds. Computer support was provided in part by the Virtual Reality Applications Center (VRAC) at Iowa State University.

REFERENCES

Bader,G.D. *et al*. (2002) BIND: the biomolecular interaction network database. *Nucleic Acids Res*., **31**, 248–250.

Barnes,J.C. (2002) Conceptual biology: a semantic issue and more. *Nature*, **417**, 587–588.

Blagosklonny,M.V. and Pardee,A.B. (2002) Conceptual biology: unearthing the gems. Nature, **416**, 373.

Dickman,S. (2003) Tough mining. *PLoS Biol.*, **1**, 144–147.

Ding,J., Berleant,D., Nettleton,D. and Wurtele,E. (2002) Mining MEDLINE: abstracts, sentences, or phrases? *Pac. Symp. Biocomput.*, **7**, 326–337.

Donaldson,I. *et al*. (2003) PreBIND and Textomy—mining the biomedical literature for protein–protein interactions using a support vector machine. *BMC Bioinformatics*, **4**, http://www.biomedcentral.com/1471-2105/4/11

iHOP. Information Hyperlinked Over Proteins. National Center of Biotechnology (CNB), Madrid. Last accessed in November, 2004.

Marcotte,E. *et al*. (2001) Mining literature for protein–protein interactions. *Bioinformatics*, **17**, 359–363.

MeSH. Medical Subject Headings, U.S. National Library of Medicine. Last accessed on January 20, 2005.

NCBI Entrez Taxonomy Homepage. U.S. National Library of Medicine. Last accessed on November, 2004.

Peri,S. *et al*. (2003) Development of human protein reference database as a initial platform for approaching systems biology in humans. *Genome Res.*, **13**, 2363–2371.

PubMed. U.S. National Library of Medicine. Last accessed in November, 2004.

Swanson,D.R. (2004) Welcome to Arrowsmith 3.0. Last accessed in November, 2004.

Tanabe,L. *et al*. (1999) MedMiner: an internet text-mining tool for biomedical information, with application to gene expression profiling. *BioTechniques*, **27**, 1210–1217.

UMLS. Unified Medical Language System. U.S. National Library of Medicine. Last accessed in November, 2004.

Xenarios,I. *et al*. (2002) DIP, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.*, **30**, 303–305.

Zanzoni,A. *et al*. (2002) MINT: a Molecular INTeraction database. *FEBS Lett.*, **513**, 135–140.