

PathBinderH: a Tool for Sentence-Focused, Plant Taxonomy-Sensitive Access to the Biological Literature¹

Jing Ding^{2,3}, Karthik Viswanathan^{3,4}, Daniel Berleant^{2,3,*}, Eve Wurtele^{3,5}, Daniel
Ashlock^{3,6,†}, Julie Dickerson^{2,3}, Andy Fulmer⁷, and Patrick S. Schnable^{3,5,8}

²Department of Electrical and Computer Engineering; ³Iowa State University, Ames,
Iowa 50011; ⁴Department of Industrial Engineering; ⁵Department of Genetics,
Development, & Cell Biology; ⁶Department of Mathematics; ⁷Miami Valley
Laboratories, The Procter & Gamble Co., 11810 E. Miami River Rd., Ross, OH 45061;
⁸Center for Plant Genomics, and Department of Agronomy

Additional Footnotes

¹This research was funded in part by a competitive grant from the National Science Foundation Plant Genome Program (award number: DBI-0321711), and by funding from The Procter & Gamble Co. Support was also provided by Hatch Act and State of Iowa funds.

***Corresponding author**; email berleant@iastate.edu; fax (515) 294-8432

[†]Currently at Department of Mathematics and Statistics, University of Guelph, Ontario, Canada
N1G 2W1

PathBinderH: a Tool for Sentence-Focused, Plant Taxonomy-Sensitive Access to the Biological Literature

ABSTRACT

Mining the biological “literaturome” promises significant advancements in genome annotation, literature access, curation support, and other applications. Standard tools allow users to identify scientific abstracts containing one or more query terms. In contrast, PathBinderH is a Web-served text mining tool that allows users to search PubMed (including MEDLINE) for sentences with co-occurring terms and their containing abstracts. The most novel and distinguishing feature of PathBinderH, however, is that the set of abstracts to be searched can be constrained by user-specified plant taxa. This enables (1) screening out abstracts dealing with species of less interest while retrieving sentences from abstracts about any of the potentially many species within the specified taxa, and (2) identifying abstracts that are more likely to prove relevant to a user than abstracts that contain the query terms in different sentences because the query terms are more likely to be used in a coordinated way.

PathBinderH may be run over the Web at www.plantgenomics.iastate.edu/PathBinderH. By making it easier to access relevant literature, PathBinderH not only enables the plant community to efficiently zero in on existing works, enhancing their dissemination and hence their contributions, it also demonstrates a literature access model that can be directly applied to the literatures of other biological research communities.

INTRODUCTION

Automated text mining in biology has grown dramatically in importance and activity since the late 1990s, motivated by the expectation that this will enhance efforts to understand and control biological processes (Barnes, 2002; Blagosklonny and Pardee, 2002; Dickman, 2003). Mined information can then be used for applications such as gene annotation and streamlined literature access.

The goal of mining the biological literature for interactions has inspired several efforts to generate public resources. Prominent among these are MedMiner at the National Library of Medicine (Tanabe et al., 1999), PreBIND (Donaldson et al., 2003), an ongoing project feeding the curated BIND (Bader et al., 2002) system at the University of Toronto, Arrowsmith at the University of Chicago (Swanson, 2004), and iHOP (2004). The scale of such mined resources is much greater than that of on-line interaction database projects that rely 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). This attests to the present and potential future value of automatically mining the scientific literature.

However, the mining-based works do not presently integrate information from the biological taxonomy into the resources they provide. Species relatedness as expressed by the biological taxonomy promises significant improvement in gene annotation and literature access. The lack of taxonomy sensitivity unnecessarily hinders access to the scientific literature by systems biologists, students, and many others. It also renders unavailable the ready annotation of genes with relevant passages from the literature, thus hindering full use of existing knowledge.

PathBinderH is designed to demonstrate incorporation of the plant taxonomy into literature access. Users can view sentences and their containing abstracts relevant to specified plant taxa even when the taxon in the query is not the one mentioned in the abstract. For example, specifying poaceae as the taxon will enable an abstract mentioning maize or corn to be retrieved.

This is significantly more powerful than requiring users to state specific species and other taxa names.

RESULTS

PathBinderH is a Web-based tool that allows biologists enhanced access to an important segment of the scientific literature, PubMed, an on-line literature search service provided by the U.S. National Library of Medicine that contains MEDLINE and some smaller resources. PathBinderH itself consists of five modules: dictionary, PubMed crawler, sentence database, taxonomy filter, and Web-based user interface (Figure 1). The *dictionary* contains terms collected from the Enzyme Nomenclature Database (ENZYME, 2004), The Gene Ontology (GO, 2004), the Plant Ontology (PO, 2004), and MeSH (2004). The *crawler* populates the sentence database by pulling in each sentence containing two (or more) different terms in the dictionary from each PubMed entry. A PubMed entry usually consists of a title, an abstract, and some ancillary information. Often, multiple terms exist for the same or very similar concepts. Thus, PathBinderH returns sentences containing either the terms specified by the user or their synonyms. The Web interface provides a convenient interface for retrieving and presenting sentences. Each presented sentence is accompanied by a clickable link pointing to its containing on-line PubMed entry. The interface also provides an easy way to set up taxonomy filters, so that sentences are retrieved only from PubMed entries that concern the plant species or other taxa of interest.

The Taxonomy Filter

This useful capability is apparently not currently available in other biomedical text retrieval systems. For example, a user may wish to search only those PubMed entries relevant to grasses (including maize, wheat, rice, etc.), or to green plants in general (thus also considering *Arabidopsis* and many other species), or perhaps only to maize or some other individual species.

To address this important challenge, PathBinderH uses of a synonym database and the Linnaean taxonomy database available via the NCBI Entrez Taxonomy Homepage (2004). This database contains the names of organisms represented in the genetic databases in a hierarchical format. For every taxon in the plant portion of the tree, a list of PubMed entries that mention the taxon was automatically created by querying PubMed with the entry's scientific name plus its common names. Each PubMed entry indexed in PathBinderH is classified according to any plant taxa it explicitly mentions, as well as all plant taxa above it in the NCBI taxonomy. Consequently if a user sets the taxonomy filter when querying, sentences from PubMed entries in lists associated with the same or lower taxa are retrieved. Hence, a PubMed entry mentioning maize (or *Zea mays*, or corn) will also be flagged as being associated with *Zea*, Andropogoneae, Panicoideae, Poaceae, etc. This feature allows users to search, for example, all PubMed entries that mention a species or other taxon below Poaceae in the taxonomy (which will cover those that mention maize, wheat and rice, but not Arabidopsis) or below Magnoliophyta (which will additionally cover Arabidopsis and many other plants).

The value of the sentence-based, taxonomy-sensitive literature access provided by PathBinderH may be illustrated by an example comparing it to the access provided by PubMed. The topic of interest in this example is embryo development in plants. The queries to PubMed that were tested were as follows.

1. Embryo AND development, internally converted by PubMed into ("embryo"[MeSH Terms] OR embryo[Text Word]) AND ("growth and development"[Subheading] OR "human development"[MeSH Terms] OR development[Text Word]). (The internal conversion is easily viewed by clicking the “Details” menu item on the PubMed results page.) This query returned 63,606 hits, mostly about animals. For clarification, a variant of this query with a simpler internal representation was typed into the input box directly, shown next.

2. Embryo[Text Word] AND development[Text Word]. This returned 55,863 hits. Because this and the previous query had such low precisions, simple taxonomic filtering was included in the next query.
3. Embryo AND development AND plants. This query filtered out many irrelevant PubMed entries, resulting 1,731 hits. A variant query containing plant in singular form was also tried, described next.
4. Embryo AND development AND plant. After internal conversion this returned 1,838 hits. To try a simple internal expansion directly, the following query specifying the internal representation was typed directly into the query input window.
5. Embryo[Text Word] AND development[Text Word] AND plant[Text Word]. This query returned 890 PubMed entries. This and the preceding two queries have the similar problems of returning hits which are about development of things other than embryos, and not returning hits on specific kinds of plants when the term “plant” is not present. These problems help motivate the new tool we have developed, PathBinderH.

PathBinderH (www.plantgenomics.iastate.edu/PathBinderH) provides *sentence-focused, taxonomy-sensitive* searches in contrast to the abstract-focused, taxonomy-insensitive search provided by PubMed. To use PathBinderH for the embryo development in plants example, the taxon of interest was specified as Viridiplantae (green plants, Figure 2). This restricts results to sentences in those PubMed entries that contain the name of a species or other taxon at or below green plants in the biological taxonomy. Next, two terms were chosen to request retrieval of sentences (including titles) from PubMed entries in which those two terms or their synonyms co-occur (Figure 3), resulting in 651 such sentences contained within 542 PubMed entries.

A revealing contrast exists between the 651 PubMed entries returned by this query, and the 890 returned by query 5 above using the standard PubMed interface: only 159 PubMed entries were returned by both queries. The next section discusses this further.

DISCUSSION

The numbers just given for the query example are summarized in Table I. These numbers have some notable characteristics, as follows.

- The numerical contrast between the cells containing 159 and 383 shows that most (383) preponderance of plant-related entries do not actually contain the term “plant.” This illustrates the advantage of retrieval that is sensitive to the biological taxonomy, thereby enabling retrieval of material about groups of related organisms. In this example entries about individual crop species, about grain plants in general, about the well-studied plant *Arabidopsis*, etc., are retrieved, while those about humans, animals in general, specific animal species, and so on are filtered out.
- The numerical contrast between the cells containing 159 and 731 shows that most (731) collocations of “embryo” with “development” are not in the same sentence. This is significant because Ding et al. (2002) showed that a large majority of relationships, at least between biomolecule names, are described within single sentences. Therefore single-sentence collocations will often be a significantly richer source of information on the relationships between the co-occurring terms than more widely spaced collocations.
- The lower right cell does not contain a number. This is because of limitations in both PubMed and PathBinderH. PubMed does not currently provide taxonomy-sensitive retrieval (although as the present report shows, this would certainly be feasible), and PathBinderH does not retrieve PubMed entries containing the query terms but not in the same sentence.

The text mining process behind PathBinderH is limited by the difficulty of highly dependable analysis of natural language, which is due to its flexible, human-oriented character. This general

problem has pervaded automatic natural language analysis-based applications for decades. It appears in the present application as discussed next.

Second-order considerations. In language processing problems not characterized by highly constrained text structures, successful solutions typically also highlight second-order limitations. These can serve to motivate further advances and thus are useful to catalog, so we note the following limitations in results provided by PathBinderH.

- The crawler module uses a fuzzy string-matching algorithm based on a tokenizer which ignores spaces in multiple-word terms. While this successfully merges terms like *UCP 3*, *UCP3*, and *UCP-3*, it sometimes leads to incorrect mergings. For example, in the following example, the phrase “act in” was mistakenly labeled as the molecule “actin.” plants.”

*The calcium channel blocker verapamil and arsenite **act in** synergy in cells exhibiting the efflux system.* (PMID: 7838183)

- Some PubMed entries may slip through the taxonomy filter even though they have nothing to do with the species of interest. This is because some taxa have ambiguous synonyms. For example, the “plants” in the following passage are not “green plants” (*Viridiplantae*), but “water-treatment plants.”

*The efficiency of the **plants** in removing nonylphenolic compounds from drinking water is highly variable, ranging from 11% to 99%.* (PMID: 15172597)

However, existing works suggest that the presence of an ambiguous taxon name in the same PubMed entry as a query term meaningfully related to it will tend to be used in its relevant sense (i.e., as a taxon name). Although those works did not investigate PubMed texts, they did find that collocations separated by large distances (as much as 10,000 words) had significant disambiguating effects in texts concerning a range of

different topics, including medicine (Gale et al., 1992; Yarowsky, 1993). Fortunately the query strategy, which requires specifying a second term to co-occur in the same sentence as another possibly ambiguous term of interest, is typically sufficient to screen out sentences in which the ambiguous term is used in an undesired sense (Yarowsky, 1993).

- In some cases, plant taxa names can legitimately be extracted from an abstract even though the abstract is mainly about something else, as in the following examples.

*Pregnant dams received either subcutaneous injections of 1 microg of E on Day 2 of pregnancy only (vaginal plug = Day 1), or 5.0mg of MXC on Days 2-4 of pregnancy in **sesame** oil.* (PMID: 15013069)

*Detailed expression analysis from gastrula to neurula stages showed that these four genes named *crescent*, *P7E4* (homologous to human hypothetical genes), *P8F7* (an unclassified gene), and *P17F11* (homologous to human and ***Arabidopsis*** hypothetical genes) demarcate spatiotemporally distinct subregions of the AEM corresponding to the head organizer region.* (PMID: 11784032)

As in the case of ambiguous taxon names, the additional query terms required to be present should often filter out such irrelevant PubMed entries.

MATERIALS AND METHODS

The Dictionary

This database of concepts and terms was built from a variety of sources (Table II). For the current system, terms from the Gene Ontology (GO, 2004), the Plant Ontology (PO, 2004), the Enzyme Nomenclature (ENZYME, 2004), and MeSH (2004) are included. In the future, additional terms from the Unified Medical Language System (UMLS, 2004), may be added. The **Gene Ontology** is provided by the Gene Ontology Consortium to help annotate gene products based on molecular

function, biological process and cellular component. The **Plant Ontology** from the Plant Ontology Consortium (POC) gives a controlled vocabulary for plant anatomy and growth stages. Both the plant and gene ontology updates for PathBinderH are through Gramene (2004), the Comparative Mapping Resource for Grains. The **Enzyme Nomenclature** was obtained from ExPASy Enzyme. This includes the Enzyme number (EC#) and enzyme names. **MeSH** gives a lexicon of concepts and synonyms which is continually updated by subject specialists in various bioscience areas.

PubMed

The source of all the text mined by PathBinderH is the U.S. National Library of Medicine's PubMed (2004) service, which provides all of the information including titles and abstracts available from the MEDLINE database at NLM, in addition to lesser amounts of other texts. PubMed includes over 14 million citations to articles in the bioscience literature. The most recent update of the PathBinderH database was based on PubMed as of June 22, 2004.

CONCLUSION

PathBinderH is a resource available for public use over the Web. It supports a novel approach to focused access to the biological literature, using keyword queries limited to those PubMed entries that concern both specified and implied plant species and other taxa. For example, the user can browse the taxonomy and click on "poaceae" (grasses), thereby delimiting the pool of PubMed entries within which to search to those containing the name of any species of grain (e.g. wheat, maize, and rice) or other grass, any genus below the poaceae family in the biological taxonomy, or the term poaceae itself or its common synonym "grasses." The plant taxonomy-sensitive approach used by PathBinderH forms a model for the analogous treatment of other taxonomies

and biological ontologies. Taxonomy-sensitive retrieval supports the needs of biologists and is expected to contribute to a range of useful applications.

An additional outcome of the present work is support for literature searches seeking relationships between key terms, because requiring terms to co-occur within a single sentence enhances the likelihood that they are conceptually explicitly connected in the retrieved PubMed entries. Another potential outcome that is now within reach is conveniently accessible gene annotations mined from the literature. Thus, PathBinder provides an effective entrée to the literature based on the concepts of (1) query term collocation within sentences, and (2) biological taxonomy.

ACKNOWLEDGEMENTS

The authors are grateful to the Virtual Reality Applications Center at Iowa State University for computer support, and to Wenxin Qi for coding an early experimental version of the system.

LITERATURE CITED

Bader GD, Betel D, Hogue CWV (2002) BIND: the Biomolecular Interaction Network

Database. *Nucleic Acids Research* **31** (1): 248-250

Barnes JC (2002) Conceptual biology: a semantic issue and more. *Nature* **417**: 587-588

Blagosklonny MV, Pardee AB (2002) Conceptual biology: unearthing the gems. *Nature* **416**: 373

Dickman S (2003) Tough mining. *PLoS Biology* **1** (2): 144-147

Ding J, Berleant D, Nettleton D, Wurtele E (2002) Mining MEDLINE: abstracts, sentences, or phrases? *Pacific Symposium on Biocomputing* **7**, Hawaii, January, pp. 326-337, <http://psb.stanford.edu>

Donaldson I, Martin J, de Bruijn B, Wolting C, Lay V, Tuekam B, Zhang S, Baskin B, Bader GD, Michalickova K, Pawson T, Hogue CWV (2003) PreBIND and Textomy – mining the biomedical literature for protein-protein interactions using a support vector machine. *BMC Bioinformatics* **4** (11) <http://www.biomedcentral.com/1471-2105/4/11>

ENZYME (as of 9/04) Enzyme Nomenclature Database. ExPASy server of the Swiss Institute of Bioinformatics (SIB), <http://us.expasy.org/enzyme/>

Gale W, Church K, Yarowsky D (1992) A method for disambiguating word sense in a large corpus. *Computers and the Humanities* **26** (5): 415-439

GO (as of 9/04) The Gene Ontology. The Gene Ontology Consortium, <http://www.geneontology.org/>

Gramene: a Comparative Mapping Resource for Grains (as of 9/04). <Http://www.gramene.org>

iHOP (as of 9/04) Information Hyperlinked Over Proteins. National Center of Biotechnology (CNB), Madrid, <http://www.pdg.cnb.uam.es/UniPub/iHOP>

Marcotte E, Xenarios I, Eisenberg D (2001) Mining literature for protein-protein interactions. *Bioinformatics* **17** (4): 359-363

MeSH (as of 9/04) Medical Subject Headings. U.S. National Library of Medicine, <http://www.nlm.nih.gov/mesh/meshhome.html>

NCBI Entrez Taxonomy Homepage (as of 9/04). U.S. National Library of Medicine, <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Taxonomy>

Peri S and 51 additional authors (2003) Development of human protein reference database as a initial platform for approaching systems biology in humans. *Genome Research* **13**: 2363-2371

PO (as of 9/04) Plant Ontology. The Plant Ontology Consortium, <http://www.plantontology.org/>

PubMed (as of 9/04). U.S. National Library of Medicine, <http://www.ncbi.nlm.nih.gov/PubMed/>

Swanson DR (as of 9/04) Welcome to Arrowsmith 3.0, <http://kiwi.uchicago.edu>

- Tanabe L, Scherf U, Smith LH, Lee JK, Hunter L, Weinstein JN** (1999) MedMiner: an Internet text-mining tool for biomedical information, with application to gene expression profiling. *BioTechniques* **27**: 1210-1217
- UMLS** (as of 9/04) Unified Medical Language System. U.S. National Library of Medicine, <http://www.nlm.nih.gov/research/umls>
- Xenarios I, Salwinski L, Duan XJ, Higney P, Kim S-M Eisenberg D** (2002) DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Research* **30** (1): 303-305
- Yarowsky D** (1993) One sense per collocation. ARPA Human Language Technology Workshop Proceedings, Princeton, New Jersey, pp. 266-271
- Zanzoni A, Montecchi-Palazzi L, Quondam M, Ausiello G, Helmer-Citterich M, Cesareni G**, (2002) MINT: a Molecular INTeraction database. *FEBS Letters* **513**: 135-140

FIGURE CAPTIONS

Figure 1. Architectural overview of PathBinderH.

Figure 2. Selecting green plants (and therefore its constituent species and other sub-taxa).

Figure 3. Sentences (including titles) containing two specified terms, and that are found in PubMed entries containing the name of a green plant species or other taxon. Clicking the PMID displays the full entry.

TABLES

Query terms:	PubMed entries containing→	“plant” or “plants”	another plant taxon name
“embryo”	...when query terms are↓		
and	in one sentence	159	383
“development”	not in one sentence	731	undetermined

Table I. Numbers of PubMed entries retrieved by three categories of query. The numbers refer to the quantity of PubMed abstracts in each category.

Source	# of concepts	# of terms
Enzyme Nomenclature Database	3,978	12,944
Gene Ontology	15,959	20,128
Plant Ontology	551	551
Medical Subject Headings	22,584	79,873
Total	43,072	113,496

Table II. Sources of the concepts and terms in the PathBinderH dictionary.

FIGURES

Figure 1

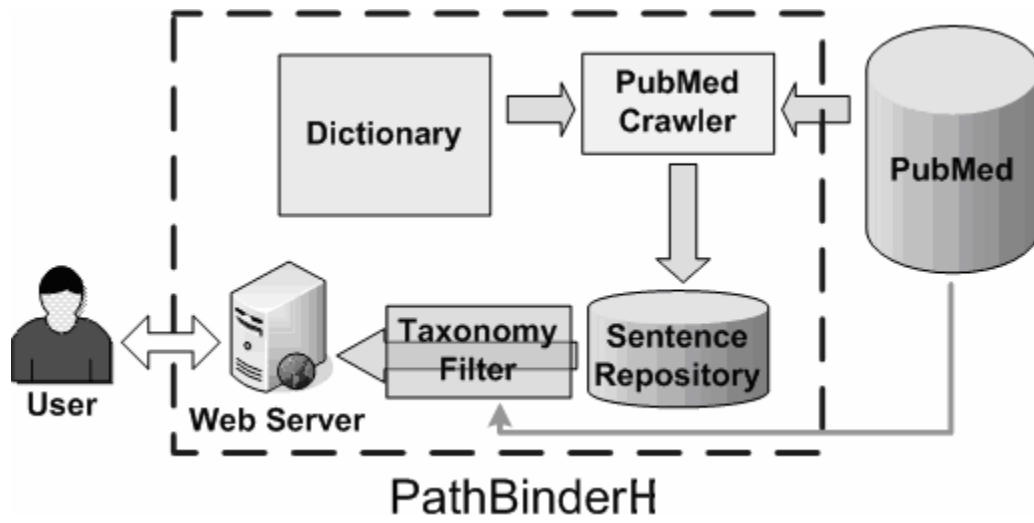


Figure 2

PathBinderH

[home](#) | [help](#) | [contact us](#) |

Hierarchical taxonomy filter

NCBI : Lineage > Eukaryota ->

- **Viridiplantae** (Green plants) ▼ <- Click funnel icon to add filter
 - **Chlorophyta** ▼
 - Chlorophyceae ▼
 - Chlorophyta incertae sedis ▼
 - Pedinophyceae ▼
 - Picocystophyceae ▼
 - Prasinophyceae ▼
 - Trebouxiophyceae ▼
 - Ulvophyceae ▼
 - unclassified Chlorophyta ▼
 - **Streptophyta** ▼
 - Charophyta/Embryophyta group ▼
 - Chlorokybophyceae ▼
 - Klebsormidiophyceae ▼
 - Mesostigmatophyceae ▼
 - Zygnemophyceae ▼
 - unclassified Viridiplantae ▼
 - environmental samples ▼

To add filter click funnel (▼) icon next to taxonomy term in the tree at left.

Filter List

☒ Viridiplantae (Green plants)

Apply filter

Note: Applying filter process might take few minutes depending on the taxonomy terms in the filter list, please be patient

Figure 3

PathBinderH: Result - Microsoft Internet Explorer

Address <http://129.186.85.49/PathBinderH/pbout.php?term1=17315&term2=13112&page=3> Go Links »

PathBinderH

[home](#) | [help](#) | [contact us](#) |

Here are the abstracts you have requested!

You requested abstracts that concern the taxon '*Viridiplantae (Green plants)*'

Your search terms are,

Term 1: '*development*'
Synonyms included in the search: No synonyms

Term 2: '*embryo*'
Synonyms included in the search: No synonyms

No of results: **651**

Pages : **Pages (65):** [[«First](#)] [[«Prev](#)] [[2](#)] [[3](#)] [[4](#)] [[5](#)] [[6](#)] [[Next](#)»] [[Last](#)»]

PMID	Year	Sentence
15136745	2004	Knockout of <i>A. thaliana</i> <i>gyrA</i> is embryo -lethal whereas knockouts in the <i>gyrB</i> genes lead to seedling-lethal phenotypes or severely stunted growth and development .
15169931	2004	The development of the embryos was arrested at the globular stage concomitant with a transient increase in <i>ATS2</i> gene expression.