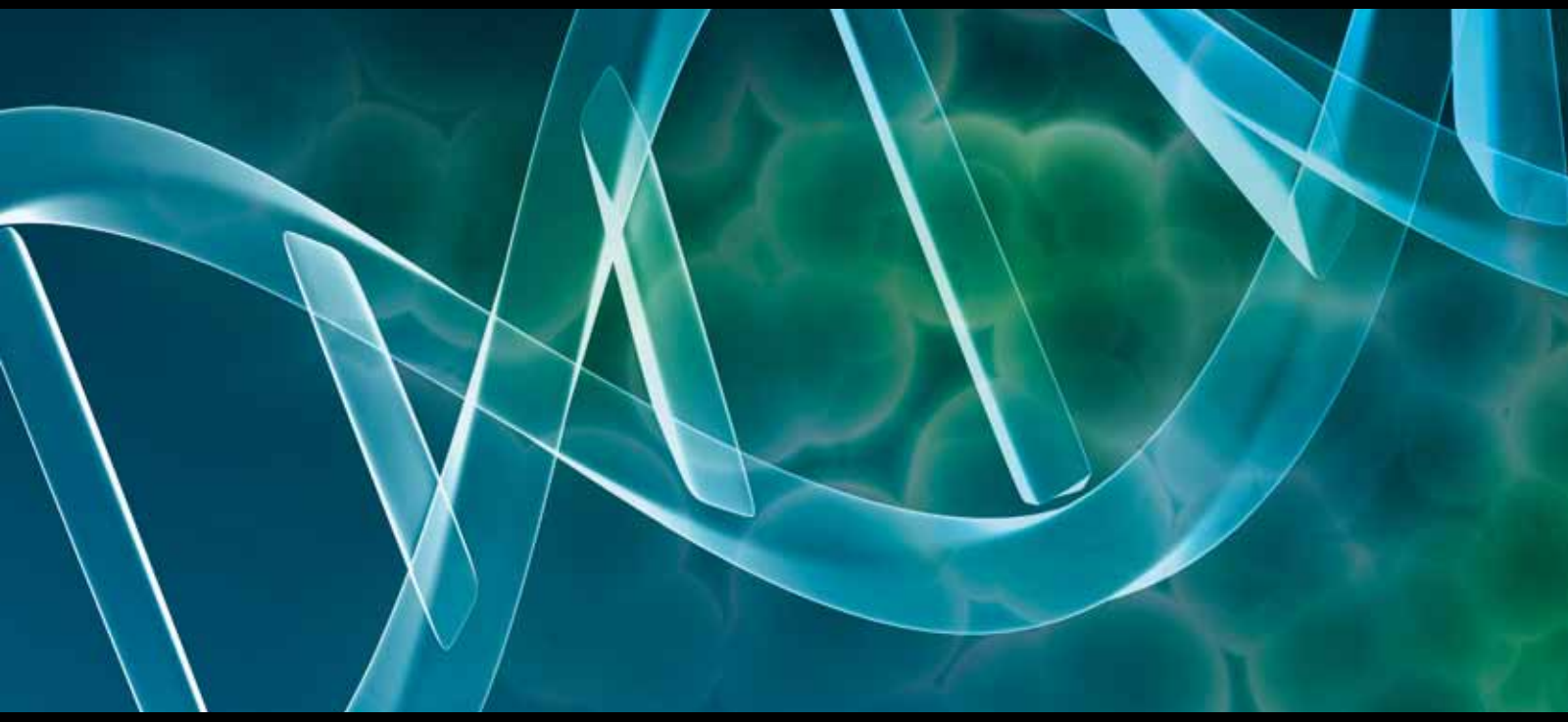
Advances in Bioinformatics

Literature-Mining Solutions for Life Science Research

Guest Editors: Jörg Hakenberg, Goran Nenadic, Dietrich Rebholz-Schuhman, and Jin-Dong Kim



Literature-Mining Solutions for Life Science Research

Advances in Bioinformatics

Literature-Mining Solutions for Life Science Research

Guest Editors: Jörg Hakenberg, Goran Nenadic, Dietrich Rebholz-Schuhman, and Jin-Dong Kim

Editorial Board

Contents

Hindawi Publishing Corporation
Advances in Bioinformatics
Volume 2013, Article ID 320436, 2 pages
http://dx.doi.org/10.1155/2013/320436

Editorial

Literature Mining Solutions for Life Science Research

Jörg Hakenberg,[1] Goran Nenadic,[2] Dietrich Rebholz-Schuhmann,[3,4] and Jin-Dong Kim[5]

[1] *Disease Translational Informatics, F. Hoffmann-La Roche Inc., Nutley, NJ 07110, USA*
[2] *School of Computer Science and Manchester Institute of Biotechnology, University of Manchester, Manchester M13 9PL, UK*
[3] *European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK*
[4] *Institut für Computerlinguistik, Universität Zürich, 8050 Zürich, Switzerland*
[5] *Database Center for Life Science (DBCLS), Research Organization of Information and Systems (ROIS), Tokyo 113-0032, Japan*

Correspondence should be addressed to Jörg Hakenberg; jorg.hakenberg@roche.com

Received 11 December 2012; Accepted 11 December 2012

Research and development in the area of biomedical literature analysis aims at providing life science researchers with effective means to access and exploit knowledge contained in scientific publications. Virtually all journal publications and many conference proceedings are nowadays readily available in an electronic form—for instance, as abstracts through the MEDLINE citation index or as full-text articles through PubMed Central. Nevertheless, keeping up to date with and searching for recent findings in a research domain remains a tedious task hampered by inefficient and ineffective means for access and exploitation. Biomedical text analysis aims to improve access to unstructured knowledge by alleviating searches, providing auto generated summaries of documents and topics, linking and integrating publications with structured resources, visualizing content for better understanding, and guiding researchers to novel hypotheses and into knowledge discovery.

Focused research over recent years has improved fundamental solutions for biomedical text mining, such as document retrieval, named entity recognition, normalization and grounding, and extraction of relationships, with levels of accuracy that reach human annotators when considering inter annotator agreement. Consequently, more and more integrative analysis tools were put forward by the text mining community targeting a broad audience of end users: generic and task-specific search engines for life science researchers, interfaces for networks synthesis based on textual evidences, or more specialized tools searching for transcription factors, or primer sequences.

This special issue of Advances in Bioinformatics presents overviews and examples of end-user-oriented biomedical text mining tools for bioinformaticians, molecular biologists, biochemists, clinicians, pharmacologists, and other researchers in life sciences.

We start with A. Manconi et al. survey on "*Literature retrieval and mining in bioinformatics: state of the art and challenges*." The authors introduce the major concepts that life science researchers should be familiar with getting the best out of existing text mining solutions, and survey key tools and research. In a dedicated second part of their survey, the authors address the major challenges both life science researchers and solution developers are facing at this point. The reader will find plenty of references to existing search tools, resources, and research papers.

A. E. Thessen et al. focus on a particular domain, presenting an overview of "*Applications of natural language processing in biodiversity science*." The authors review the application of natural processing and machine learning for biological information extraction regarding cellular processes, taxonomic names, and morphological characters. You will find detailed examples, a summary of all steps involved in information extraction, and lots of references to existing tools and resources.

S. T. Ahmed et al. introduce their semantic faceted search engine, BioEve, in "*A novel framework to facilitate interactive literature search*." They couple an automated extraction system with a cognitive search and navigation service, to alleviate the process of searching and browsing huge amounts of literature such as provided/delivered by MEDLINE. BioEve

enables interactive query refinement and suggests concepts and entities (like genes, drugs, and diseases) to quickly filter and modify search directions, thereby achieving semantic enrichment that improves insight gains while searching literature.

S. V. Landeghem et al. present their EVEX resource in "*Exploring biomolecular literature with EVEX: connecting genes through events, homology, and indirect associations.*" The authors extracted more than 20 million biomolecular events involving genes and proteins, such as phosphorylation and gene regulation, from MEDLINE. The online tool generates a summary on the searched gene denoting all regulated genes, binding partners, subcellular locations, and other related data linked to the searched gene.

We conclude this special issue with the paper by A. Divoli et al., discussing whether "*Do Peers see more in a paper than its authors.*" In a meta-analysis using automatic text analysis, they address questions such as how informative an abstract is compared to the full text; and how peers and authors might view the major contributions of a paper differently. Their analyses are comparing the information content of an abstract, as written by the paper's authors, to sentences that mention the paper as a reference, written by peers. Using this strategy, A. Divoli et al. found, for example, that citing sentences contain 20% additional concepts (likely important contributions) that were not mentioned in the abstract of the paper referred to, but maybe should have been to help attract even more peers.

Acknowledgment

The guest editors wish to thank all authors for their contributions to this special issue, as well as the numerous reviewers who supported the authors and us with their invaluable feedback.

Jörg Hakenberg
Goran Nenadic
Dietrich Rebholz-Schuhmann
Jin-Dong Kim

Hindawi Publishing Corporation
Advances in Bioinformatics
Volume 2012, Article ID 750214, 15 pages
doi:10.1155/2012/750214

Research Article

Do Peers See More in a Paper Than Its Authors?

Anna Divoli,[1] Preslav Nakov,[2] and Marti A. Hearst[3]

[1] *Pingar Research, Pingar, Auckland 1010, New Zealand*
[2] *Qatar Computing Research Institute, Qatar Foundation, Tornado Tower, Floor 10, P.O. Box 5825, Doha, Qatar*
[3] *School of Information, University of California at Berkeley, CA 94720, USA*

Correspondence should be addressed to Anna Divoli, annadivoli@gmail.com

Received 16 December 2011; Revised 17 March 2012; Accepted 5 June 2012

Academic Editor: Goran Nenadic

Recent years have shown a gradual shift in the content of biomedical publications that is freely accessible, from titles and abstracts to full text. This has enabled new forms of automatic text analysis and has given rise to some interesting questions: How informative is the abstract compared to the full-text? What important information in the full-text is not present in the abstract? What should a good summary contain that is not already in the abstract? Do authors and peers see an article differently? We answer these questions by comparing the information content of the abstract to that in *citances*—sentences containing citations to that article. We contrast the important points of an article as judged by its authors versus as seen by peers. Focusing on the area of molecular interactions, we perform manual and automatic analysis, and we find that the set of all citances to a target article not only covers most information (entities, functions, experimental methods, and other biological concepts) found in its abstract, but also contains 20% more concepts. We further present a detailed summary of the differences across information types, and we examine the effects other citations and time have on the content of *citances*.

1. Introduction

Text mining research in biosciences is concerned with how to extract biologically interesting information from journal articles and other written documents. To date, much of biomedical text processing has been performed on titles, abstracts, and other metadata available for journal articles in PubMed[1], as opposed to using full text. While the advantages of full text compared to abstracts have been widely recognized [1–5], until relatively recently, full text was rarely available online, and intellectual property constraints remain even to the present day. These latter constraints are loosening as open access (OA) publications are gaining popularity and online full text is gradually becoming the norm. This trend started in October 2006, when the Wellcome Trust[2], a major UK funding body, changed the conditions of grants, requiring that "research papers partly or wholly funded by the Wellcome Trust must be made freely accessible via PubMed Central[3] (PMC) (or UK PubMed Central once established) as soon as possible, and in any event no later than six months after publication" [6]. Canadian Institutes of Health Research followed, as did the National Institute of Health (NIH) in the USA in April 2008.[4] Moreover, many publishers founded and promoted OA initiatives, namely, BioMed Central[5] (BMC) and the Public Library of Science[6] (PLoS). PubMed now offers access to all OA publications via PMC. The availability of OA publications has allowed several recent text mining and information retrieval competitions turning to use full-text corpora, for example, BioCreAtIvE since 2004, the TREC Genomics Track since 2006, and the BioNLP shared task since 2011.

The rise of full text, which differs in length (both overall length and average sentence length), structure (e.g., use of parenthesized text, tables, and figures), and content from abstracts, has posed many new challenges for biomedical text processing, for example, standard tools like part-of-speech and gene mention taggers were found to perform much worse on article bodies than on abstracts [7]. The availability of full text has further opened up some more general interrelated questions.

(1) *How informative is the abstract compared to the full text?*

(2) *What important information in the full text does not appear in the abstract?*[7]

(3) *What should an ideal summary of the full text contain that is not already in the abstract?*

(4) *What are the differences in the way authors and peers see an article?*

We explore these questions indirectly, using an under-explored information source: the sentences containing the citations to a target article or *citances*. While cocitation analysis is commonly-used for determining the popularity, and by association, the importance of a publication [8–15], our focus here is on the *contents* of the sentences containing the citations, that is, the citances.

In particular, we compare the information content of the abstract of a biomedical journal article to the information in all citances that cite that article, thus contrasting the important points about it as judged by its authors versus as seen by peer researchers over the years following its publication. Put another way, we use citances as an indirect way to access important information in the full text[8]. The idea is that (1) any information not mentioned in the abstract but referred to in citances should be coming from the full text, and (2) entities and concepts mentioned in a citance should be important and somewhat representative of their source.

To give an example, here is the abstract of an article (PubMed ID 11346650):

> *Multiple Mechanisms Regulate Subcellular Localization of Human CDC6.*
>
> *CDC6 is a protein essential for DNA replication, the expression and abundance of which are cell cycle-regulated in Saccharomyces cerevisiae. We have demonstrated previously that the subcellular localization of the human CDC6 homolog, HsCDC6, is cell cycle-dependent: nuclear during G(1) phase and cytoplasmic during S phase. Here we demonstrate that endogenous HsCDC6 is phosphorylated during the G(1)/S transition. The N-terminal region contains putative cyclin-dependent kinase phosphorylation sites adjoining nuclear localization sequences (NLSs) and a cyclin-docking motif, whereas the C-terminal region contains a nuclear export signal (NES). In addition, we show that the observed regulated subcellular localization depends on phosphorylation status, NLS, and NES. When the four putative substrate sites (serines 45, 54, 74, and 106) for cyclin-dependent kinases are mutated to alanines, the resulting HsCDC6A4 protein is localized predominantly to the nucleus. This localization depends upon two functional NLSs, because expression of HsCDC6 containing mutations in the two putative NLSs results in predominantly cytoplasmic distribution. Furthermore, mutation of the four serines to phosphate-mimicking aspartates (HsCDC6D4) results in strictly cytoplasmic localization. This cytoplasmic localization depends upon the C-terminal NES. Together these results demonstrate that HsCDC6 is phosphorylated at the G(1)/S phase of the cell cycle and that the phosphorylation status determines the subcellular localization.*

And here are some citances pointing to it:

> *Much of the soluble Cdc6 protein, however, is translocated from the nucleus to the cytoplasm when CDKs are activated in late G1 phase, thus preventing it from further interaction with replication origins [#C, #C and #TC].*
>
> *To ensure that the pre-RC will not re-form in S or G2, Cdc6p is phosphorylated and degraded in yeast (#C; #C; #C) or exported to the cytoplasm in higher organisms (#TC; #C; #C; #C; #C).*
>
> *It is phosphorylated by cyclin A-cdk2 at the G1-S transition and this modification causes some, but not all, of the Cdc6 to be exported out of the nucleus (#TC; #C; #C and #C).*
>
> *Cdc6CyΔ has a mutation in a cyclin binding motif that is an essential part of the substrate recognition signal for cdks (#TC).*
>
> *After entry into S phase, phosphorylation of HsCdc6, probably by cyclinA/CDK2, leads to its export from nucleus to the cytoplasm via NES [#TC].*
>
> *Once replication begins, Cdc6 is degraded in yeast (#C, #C, #C, #C, #C), whereas for mammals it has been suggested that Cdc6 is translocated out of the nucleus during S phase in a cyclin A-Cdk2- and phosphorylation-dependent manner (#C, #TC, #C,-#C, #C) and then subject to degradation by the anaphase-promoting complex (#C, #C, #C).*

In the above examples, #TC refers to the publication we are comparing against (the target citation: PubMed ID 11346650), whereas #C refers to other publications. Throughout this paper, we will refer to these citation sentences to other publications as *adjoining citations*.

Previous studies have discussed some of the potential of the use of *citances* for literature mining [16, 17]. Similar to anchor text on the web (visible, clickable text in a webpage, clicking on which navigates the user to another webpage), they are votes of confidence about the importance of a research article. Collectively, they also summarize the most important points about the target article, which makes them a potential surrogate for its full text [18] and an important knowledge source when generating a survey of scientific paradigms [19].

While previous work has focused on the *words* in citances, we compare their contents to the contents of the abstracts using coarse-grained biologically meaningful *concepts* such as entities, functions, and experimental methods.

Focusing on the area of molecular interactions, we perform careful *manual* analysis, and we present detailed summary of the differences across information types. We further study the effects that other citations and temporal measures have on the contents of citances. Finally, we verify these manual results with a large-scale automatic analysis.

In the remainder of this paper, we first discuss related work, then we describe our concept annotation scheme, we perform manual and automatic analysis, and we summarize the results, aggregating them over information types. Finally, we discuss the findings and we point to some promising directions for future work.

2. Related Work

In the bioscience literature, several studies focused on comparing the information structure of abstracts to that of full-text. Schuemie et al. [4], building on work by Shaw [3], looked into the density (the number of instances found divided by the number of words) of MeSH terms and gene names in different sections of full text articles. They found that the density was highest in the abstract and lowest in the Methods and the Discussion sections. They further found that nearly twice as many biomedical concepts and nearly four times as many gene names were mentioned in the full text compared to the abstract. In a related study, Yu et al. [2] compared abstracts and full text when retrieving synonyms of gene and protein names and found more synonyms in the former. A more comprehensive study on the structural and content difference of abstracts versus full text can be found in [7].

There has been extensive work on automatically generating an article abstract from full text, which studies the relationship between sentences in full text to those in abstracts [1, 5]. However, this work does not consider citances.

A lot of work on citation analysis has focused on citation links and counts, which have been used to determine the relative importance of publications within a field and to study the interaction between different fields [11–14, 20]. Today, this kind of analysis is at the core of a number of scholarly sites, including CiteSeerX[9], DBLP[10], Google Scholar[11], Microsoft Academic Search[12], ACM Digital Library[13], IEEE Xplore[14], ACL Anthology[15], and ArnetMiner[16], to mention just a few. There have been also specialized research tools for exploring citation networks, for example, [21].

In natural language processing (NLP), research has focused in a different and arguably more interesting direction, using citations as an (additional) information source to solve various text processing problems. The growing interest in the research community on the topic culminated in 2009 in a specialized workshop on Text and Citation Analysis for Scholarly Digital Libraries (collocated with the 2009 Conference on Empirical Methods on Natural Language Processing[17]).

An early overview of this general research direction was presented by White [16], who described three main lines of research.

First, citation sentences can be *categorized*, for example, as conceptual versus operational, organic versus perfunctory, and so forth. For example, Teufel and Moens [22] identified and classified citations in scientific articles and used them as features for classifying noncitance sentences, for the purpose of text summarization.

Second, *context analysis* is concerned with identifying recurring terms in citances and using them to help solve information retrieval tasks. For example, Nanba et al. [23] used citances as features to help classify papers into topics. Similarly, Bradshaw [24] indexed articles with the terms in the citances that cite them. Mercer and Di Marco [25] applied a similar idea to biomedical documents. Tbahriti et al. [26] used paper cocitation as a similarity measure when evaluating a biomedical information retrieval system. Rosario and Hearst [27] demonstrated that using citances to a publication can yield higher accuracy compared to using other sentences for the problem of multiway relation classification, applied to the identification of the interactions between proteins in bioscience text. Similarly, Kolchinsky et al. [28] improved protein-protein interaction extraction using citation network features. Aljaber et al. [29] used citances text as an additional input to improve document clustering, and Aljaber et al. [30] used the text contained in citances as an additional information source to improve the assignment of Medical Subject Headings (MeSH) terms, which are commonly-used in PubMed and other databases administered by the National Library of Medicine.

The third line of research, according to White, is concerned with *citer motivation*, that is, with identifying the reason authors cite earlier work, and why some work is more cited than other. Lehnert et al. [31] created a taxonomy of 18 citation types, such as method, attribution, fact, example, critisism, and built a system to classify citations in these types. Similarly, Teufel et al. [32] annotated citation sentences from computational linguistics papers according to their rhetorical functions (e.g., contrast/comparison in goals or methods, contrast/comparison in results, weakness of cited approach, neutral description, etc.), and Teufel et al. [33] and Teufel and Kan [34] described algorithms to automatically assign such rhetorical functions.

Another informative early overview can be found in Nakov et al. [17], who also proposed the use of *citances* (they coined this neologism to refer to citation sentences) for bioscience papers for various semantic processing tasks, including summarization of target papers, synonym identification and disambiguation, and as a way to generate candidate sentences for manual curation. They further applied text paraphrase techniques to normalize the myriad forms of expression of citances in order to determine which of them express the same subsets of concepts. This last objective was later facilitated by the work of Schwartz et al. [35] using multiple sequence alignment and conditional random fields with posterior decoding.

More importantly, Nakov et al. [17] proposed to use citances as an information source for automatic summarization of the scientific contributions of a research publication, which is somewhat related to the idea of using the information in hyperlinks to summarize the contents of

a web page [36, 37]. This direction has been explored by a number of researchers thereafter.

Schwartz and Hearst [38] hypothesized that in many cases, as time goes by, citances can indicate the most important contributions of a paper more accurately than its original abstract.

Qazvinian and Radev [39] used citation summaries and network analysis techniques to produce a summary of the important contributions of a research paper. A related technique for the same problem was proposed by Mei and Zhai [40], who relied on language modeling techniques. In a subsequent extension, Qazvinian and Radev [41] have proposed a general framework to pick the sentence(s) from a target paper that a citance in another paper is most likely referring to.

More closely related to the present work, Elkiss et al. [18] compared the information contained in the set of all citances citing a given biomedical paper and the abstract for that paper, using a lexical similarity metric called *cohesion*. They found significant overlaps but also many differences since citances focus on different aspects than abstracts.

Mohammad et al. [19] compared and contrasted the usefulness of abstracts and of citances in automatically generating a technical survey on a given topic from multiple research papers from the ACL Anthology. They found that while abstracts are undoubtedly useful, citances contain important additional information. They further noted that abstracts are author-biased and thus complementary to the broader perspective inherent in citances.

There has been also work that goes in the opposite direction: instead of trying to summarize a document using the textual content of multiple citances to it, Wan et al. [42] built a system that summarizes it using its full text in order to provide the reader with a summary relevant to a given citance in another document.

Hoang and Kan [20] introduced another interesting task: automatic related work summarization. Given multiple articles (e.g., conference/journal papers) as input, they created a topic-biased summary of related work that is specific to a given target paper.

Citations, citances, and links between them are similar to hyperlinks and hypertext on the web. Anchor text has been used in most search engines for indexing and retrieval of web pages. Applications of anchor text include identification of home pages of people and companies [43], classification of web pages [44, 45], Web crawlers [46], improved ranking of search results [47], and web page summarization [36]. See [24] for an overview of the uses of anchor text.

Our present work is more general and more quantitative than that in the above publications. First, we do not restrict ourselves to a particular application, while most work above was limited to, for example, summarization. Second, we study the degree of overlap between the information contained in abstracts and citances from a biomedical perspective focusing on molecular interactions and using biomedically meaningful semantic units (rather than words) such as entities, functions, dependencies, characteristics, locations, species, time, experimental methods, chemicals, and disorders. Third, we use and/or map our annotations to MeSH[18], a standardized hierarchical resource, thus allowing for further comparisons and applications. Fourth, we study the effect of time on the way papers are cited. We further investigate the effect of the presence of adjoining citations. Finally, we report the results from both small and focused manual analysis and from large-scale automatic analysis.

3. Methods

We performed small-scale detailed manual analysis and large-scale fully automatic comparison of the information contained in citances and abstracts.

In the manual analysis, we considered 6 abstracts from PubMed in the molecular interaction domain, published during 1996–2002, and 136 citances to them, which we carefully annotated with the mentions of entities, functions, experimental methods, and other biological concepts. More details about the dataset can be found in Table 1. We used this dataset to compare the set of concepts that appear in the abstract of an article to the set of concepts that appear in the citances to that article. We also looked at the concepts mentioned in the citances over a six-year period to study changes over time.

In the automatic comparison, we analyzed 104 journal publications in PubMed (this included the six articles used for the manual analysis), again from the molecular interaction domain, published during 1995–2002, which received a total of 11,199 citances in the period 1995–2005. We annotated the MeSH terms in the abstracts of these publications and in the corresponding citances, and we mapped these terms to broad biomedical concepts; then, we proceeded with the manual analysis. MeSH is a comprehensive controlled vocabulary created for the purpose of indexing journal articles and cataloging books in the life sciences, and it is commonly used for annotations in the biomedical domain. We chose MeSH for our automatic annotations because it is a formal established resource that has a relatively simple structure, allowing for intuitive, pragmatic analysis.

3.1. Data Selection. Our goal was to find articles that are highly cited and are in an area of biology that has attracted a lot of text mining interest. The "Molecular Interaction Maps" NIH website[19] lists a number of annotations and references for each interaction map that the site covers. We selected 104 target articles from the "Replication Interaction Maps" collection and used the ISI citation service[20] to find which articles cite the targets. We downloaded them and used the code developed by Nakov et al. [17] to extract the citances. We further collected the abstracts and the full text as well as the MeSH terms and the substances indexed by PubMed for these articles. Six of the 104 articles were used for manual analysis.

3.2. Manual Annotation. We performed detailed manual analysis of the mentions of various biologically meaningful concepts in the abstracts of the target six articles and in 136 citances to them. For one target article, we considered all 46 available (in our dataset) citances, and for another one, we

Table 1: Summary of the data used for the manual analysis.

PubMed ID Of the target	Year of publication	Number of sentences analyzed		Number of annotations in		Number of papers the citances are derived from
		Abstract	Citances	Abstract	Citances	
8939603	1996	17	51	192	728	27
11346650	2001	11	45	141	761	24
11125146	2000	8	10	91	144	10
11251070	2001	12	10	142	128	10
11298456	2001	9	10	146	178	9
11850621	2002	8	10	132	157	8
All		65	136	844	2096	88

selected a comparable number of 51 citances, whereas for each of the remaining four articles, we analyzed 10 randomly selected citances to ensure some variation. We annotated a total of 844 concepts in the six abstracts and 2,096 in the 136 citances. See Table 1 for more detail.

The goal of the annotation was to represent as much of the important contents of the citances as possible. Table 2 describes the different types of concepts we annotate, and Figure 1 shows an example of an annotated citance.

Table 2 shows the categories for manual annotation. All datasets used in this study were annotated manually following a number of rules. Every unit (word or short phrase) was assigned an ID, and any matching unit within the same set was given the same ID. A few categories of units were decided for each set; they were reflected in the first part of the ID by a capital letter. The IDs, whenever possible, were very simple: composed of a single letter and a number. However, sometimes we tried to capture more complex units, for example, if "*Xenopus*" = "S1", "*orc*" = "E1" and "*antibody*" = "E10", then "*anti-Xorc1*" = "E10.E1.S1.1", and if "*DNA*" = "H1" and "*synthesis*" = "F1", then "*DNA synthesis*" = "H1.F1" so "*DNA*" is given the IDs: "H1, H1.P1" and "*synthesis*" is assigned "P1, H1.P1". The last column shows the corresponding MeSH IDs, which were used for the automatic annotation.

We identified the distinct semantic units, words or phrases, and we assigned them annotation IDs, which had different prefixes (E, H, etc.) for different types of information. We assigned suffixes for subtypes (e.g., E2), and we represented complex concepts by combining IDs (e.g., E2.2). We used the same rules to annotate the citances (given below).

Manual Annotation Rules

(1) Try to identify units (words or phrases) that convey information in one of the annotation categories (Table 2). Use words as annotation units, whenever possible.

(2) Compare units by trying to match them to parts of other citances within the set.

(3) If an entity (category E) is comprised of more than one word, consider the words as one unit and assign the same ID to each word.

(4) Try to group entities together (extending to protein complexes and families) if used in the same context throughout the citances for a target document. Use subtypes when necessary to keep related concepts similarly labeled (.a, .b, .c. . . or .1, .2, .3).

(5) If an entity is complex, use "·" to join IDs, but keep the main entity in the front. For example, if *Xenopus* = S1, *orc* = E1 and antibody = E10, then the annotation for anti−*Xorc1* is E10.E1.S1.1 and for Xorc2 is E1.S1.2.

(6) Annotate individual word units, but also consider complex concepts (e.g., *DNA replication*). Similarly to entities, capture concepts that are made of more than one unit by concatenating their IDs with "·".

(7) When annotating complex concepts, annotate each unit of the concept with the unit's ID followed by a comma, followed by the concept ID.

(8) Consider *opposite* information units (e.g., competent-incompetent, increase-decrease). Capture these in the IDs by adding ".o".

(9) Consider subcategories of IDs by appending .a, .b, . . . or .1, .2,. . . extensions if appropriate for the same citance set, for example, *prevent and inhibit*.

3.3. Data Analysis. The annotations of the citances and abstract sentences shown in Table 1 enabled us to run a number of comparisons between the content of the abstract and the corresponding citances, the outcomes of which are presented in the next section.

In our automatic analysis, we relied on MeSH, the U.S. National Library of Medicine's controlled hierarchical vocabulary. There are 15 main subtrees in MeSH, each corresponding to a major branch of the biomedical terminology, for example, subtree A corresponds to *anatomy*, subtree B to *organisms*, subtree C to *diseases*, and so forth. Down the MeSH hierarchy, concepts are assigned one or more positional codes, for example, A (*anatomy*), A01 (*body regions*), A01.456 (*head*), A01.456.505 (*face*), and A01.456.505.420 (*eye*). Note that MeSH is not a tree, but a lattice, and thus multiple paths are possible for the same concept, for example, *eye* is ambiguous, and it has one additional code: A09.371 (A09 represents *sense organs*).

Table 2: Categories used in the manual annotation.

Categories	Description	Examples	MeSH Tree IDs
E (entities)	Genes and proteins	MCM, protein, ORC, Skp2	D06, D08, D12, and D23.529
F (function)	Biological function or process	Regulation, pathway, and function	G, F01, F02
D (dependency)	Relationship type	Involve, cause	N/A
X (characteristic)	Modifier	Unstable, common, and ionizing	N/A
L (location)	Cellular or molecular part	C-terminal, cytosol, and motif	A
S (species)	Any taxonomic description	Human, mammal, and *S. cerevisiae*	B
T (time)	Temporal information	During, after, and following	N/A
M (exp methods)	Methods and their components	Recombination, transfect	E
H (chemicals)	Not including genes/proteins	DNA, thymidine, and phosphoryl	D (except: D06, D08, D12, and D23.529)
R (disorders)	Names and associated terms	Cancer, tumor, and patient	C, F03
Special Types:			
IDs with subtypes	Subtype of a BASIC type	Retain-change, common-distinct	
IDs with opposite	Opposite of a BASIC type	Cell cycle—G phase, CDK–CDK2	
Complex IDs	Combination of BASIC types	Radio-resistant DNA synthesis	

We used an in-house MeSH term recognizer and normalizer tool, which we originally developed for our participation in the first Genomics Track [48], but which we significantly expanded thereafter. We used a version of the tool developed for the Second BioCreAtIvE Challenge [49]. The tool uses normalization rules in order to allow for the following variations in form: (1) removal of white space, for example, "*BCL 2*"⇒"*BCL2*," (2) substitution of nonalpha-numerical characters with a space, for example, "*BCL-2*"⇒"*BCL2*," and (3) concatenation of numbers to the preceding token, for example, "*BCL 2*"⇒"*BCL2*." All possible normalizations and expansions of all known MeSH terms and their synonyms were generated offline and then matched against a normalized version of the input text using an exact, first-longest-string-matching measure. The matches were then mapped back to the original unnormalized text, and the corresponding MeSH IDs were assigned.

Once the MeSH terms were identified, we considered (1) the whole MeSH tree ID and (2) the MeSH tree tag truncated to maximum 2 levels (xxx.xxx) in abstracts and citances[21]. We performed automatic analysis and mapping to identify different MeSH annotation groups (shown in Figure 2) and their counts in abstracts, corresponding citances, and their overlap. We also looked at annotations in citances with 0 adjoining citations (whose contents must have come from the target article) and how they compare to the annotations in abstracts. Finally, we looked at citances' annotations appearing in the same year as the original publication, as well as at additional/new annotations appearing in the following year, and additional annotations appearing 2, 3, and 4+ years later, and how they compare to annotations from the abstracts.

3.4. Category Mapping. There are a few distinct annotation categories in each manual and automatic schemata. However, for most categories of interest for the area of molecular interactions, the semantic annotations overlap. We provide the mapping in Figure 2.

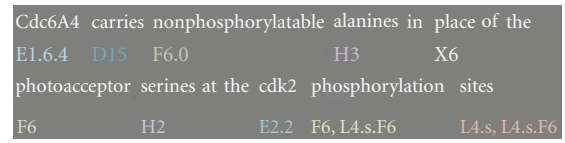


Figure 1: Example of an annotated citance. The citance is for PMID 11346650, demonstrating different categories of annotation (e.g., E, D; F; H. . .), subtypes (e.g., E1.64; L4.s; E2.2. . .), opposite concepts (e.g., F6.o), and complex IDs (e.g., L4.s.F6).

4. Results

Here we describe the results of our manual and automatic analysis, trying to answer the research questions posed in the introduction. We further study the effect of the presence of adjoining citances and of the passage of time.

4.1. Differences between Abstracts and Citances. In order to examine the differences in the contents of abstracts and citances, we compared the distributions of the ten categories of concepts that we considered in the manual analysis (see Table 2). Figure 3 shows these distributions (a) over abstracts and (b) over citances. It further presents these distributions (i) for all six articles, and (ii) for one article only, namely, the one with PubMed ID 11346650.

In Figure 3, we can see that there are generally higher proportions of "entities" and "experimental methods" annotations in citances than in abstracts. The difference for experimental methods was statistically significant for the two

	Annotation group	Mapped MeSH tree IDs	Manual analysis	Automatic analysis
1	Anatomy	[A]	✓	✓
2	Organisms	[B]	✓	✓
3	Diseases	[C]	✓	✓
4	Chemicals-Drugs	[D]: 01-05, 09-10, 13, 20, 23 (not: 23.529)	✓	✓
5	Chemicals-Entities	[D]: 06, 08, 12, 23.529	✓	✓
6	Experimental Methods	[E]	✓	✓
7	Psych. Functions	[F]: 01-02	✓	✓
8	Psych.-Disorders	[F03]	✓	✓
9	Processes-Functions	[G], [F]: 01-02	✓	✓
10	Disciplines/Occupations	[H]		✓
11	Anthr./Educ./Social	[I]		✓
12	Technol./Indus./Agric.	[J]		✓
13	Humanities	[K]		✓
14	Information Science	[L]		✓
15	Named Groups	[M]		✓
16	Health Care	[N]		✓
17	Publication Character.	[V]		✓
18	Geographicals	[Z]		✓
19	Dependency	n/a (involve, cause...)	✓	
20	Characteristic	n/a (unstable, common, ionizing...)	✓	
21	Time	n/a (during, after, following...)	✓	

Figure 2: Semantic annotation groups. This figure depicts all different annotation types associated with abstract sentences and with citances. The overlap and, where possible, the mapping of automatic and manual annotations categories are also shown. See also Table 2 for details on the mapping of MeSH IDs to categories from the manual annotation.

larger sets, corresponding to PubMed IDs 11346650 and 8939603.

The top of Figures 4 and 5 use Venn diagrams to show the overlap of unique (i.e., each ID was counted just once regardless of how many times it actually occurred) semantic annotations between abstracts and citances for the large-scale automatic analysis. Figure 4 shows the overlap over MeSH annotation categories that can be mapped (see Figure 2) to the manually assigned annotations, that is, those categories that were included in both the automatic and the manual analysis, whereas Figure 5 presents the overlap over annotation categories that were studied in the automatic but not in the manual analysis.

We see that indeed the categories in Figure 4, which we considered important for our dataset and used for the manual annotation, have a lot more unique annotations than the categories in Figure 5 that are largely less pertinent for molecular interactions (see Figure 2 for more details on the categories). We do see, however, that across all categories in both figures, citances carry a lot more annotations than abstracts with the overlap between the two being at least 50% of the abstract's unique annotations (with the exception of psychological disorders, representing a very small portion of the annotations). For most categories, the overlap is about 75–80%.

4.2. The Effect of Adjoining Citations and the Differences between Abstracts and Citances. Looking more closely at the data in Figure 3, we found that every annotation in our six manually annotated abstracts could be found in at least one citance. For the four articles for which we only consider 10 citances, we had to look for additional unannotated citances to get complete coverage for some of the concepts.

The contrary, however, was not true: some concepts found in citances were not mentioned in the abstract. Before describing this point in detail, we would like to note that very often in bioscience journal articles, a citation sentence backs up its claims with more than one reference. As we mentioned earlier, we call the references that appear in addition to the target *adjoining citations*. Our analysis has shown that citances containing adjoining citations are the source of most of that extra information. Thus, we decided to have a closer look at the clean cases of citances with zero adjoining citations (referred to as "zero adjoining citations"

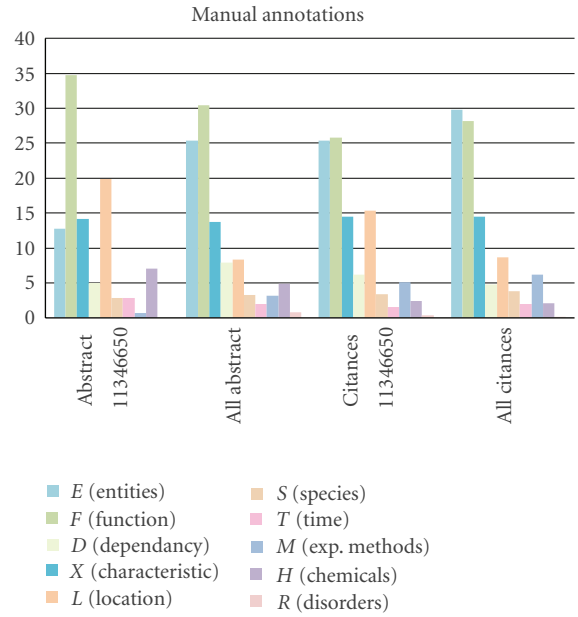


Figure 3: Distribution (in %) of the manually annotated categories for abstracts and citances. Shown are results for all abstracts and for the one with PubMed ID 11346650.

or "cit_0" below), that is, those that cited our target article only. Citances that refer to only one paper should really contain information that can be found in the citing paper.

In the manual analysis, we examined 23 citances with no adjoining citations, which corresponded to five of our target papers, and we found 73 distinct annotation types in the citances that did not appear in the abstracts. First, we checked whether the annotations conveyed biological meaning; if not, they were marked as "n/a." Then we tried to find the extra annotations in the full text of the targets, and we examined the "MeSH/substances" that the target article was indexed with in PubMed. After all these checks, a few annotations were still "not found." The distribution for each of the six articles is shown in Table 3.

Table 3 and Figure 6 (manual evaluation) show that most of the concepts that abstracts do not contain fall under the entities or the experimental methods categories. Two others were mentioned in figures of the full text paper (PMID: 11298456) as part of describing an experimental technique. Two more were actually found in the full text (PMID: 8939603) as restriction enzymes, which are commonly used in experiments to cut *dsDNA*. Some other distinct annotation types missed by abstracts were also related to Methods, for example, *plasmid*, which was annotated as a chemical; in fact, plasmids are commonly-used in genetic engineering as vectors.

Some other entities had subtypes (e.g., *Wee1A*) and although the main type was matched in the full text, the specific subtype was not. In the species category, a sentence from cit_0 for the target PubMed ID 11251070 was referring to the animal category, which was not mentioned in the abstract. The full text mentioned *eukaryotes* and *various* organisms, but it was indexed with the more general MeSH term *animals*.

We further analyzed how adjoining citations affect the number of distinct annotation types by grouping the citances into five groups: cit_0, which cites the target paper only, cit_1/cit_2/cit_3, with one/two/three adjoining citations, and cit_4+, with four or more adjoining citations. In order to compare the effect of the adjoining citation, we took the abstract of each set (representing the minimum number of distinct annotation types), and we added each of the above groups separately as well as together (the abstract and the citances representing the maximum number of distinct annotation types). The results are shown in Table 4. We can see that the more references a citance has, the more distinct the annotation types that are introduced. The effect is most clearly pronounced for the two papers with a larger set of citances, those with PubMed IDs 8939603 and 1346650.

We also studied the effect of the adjoining citations in the larger dataset, which we used for the automatic analysis. Figure 4 shows the effect that adjoining citations have on the semantic annotation content of citances. We can see that "zero adjoining citances" contain much less annotations in comparison to all citances, but the overlap of annotations with the abstracts' annotations are, proportionately, much larger.

4.3. The Effect of Time. Next, we studied how the concepts mentioned in the citances changed over time. For each target article in our large dataset, we grouped the citances per year of citation, from cited in the same year of publication to cited up to 4+ years thereafter.

Our results (see Figure 7) show that with every year passing, new annotations are being assigned to the target paper via its citances. The majority of citances' annotations that overlap with abstracts' annotations appear within the first couple of years, but more are constantly added each following year. This is quite uniform across all categories. It would be of interest to conduct more in-depth analysis to see if these new annotations are representative of the research trends progression across the biomedical literature.

5. Discussion

In this section, we discuss the effect of the internal structure of the sentences on our methodology. We further provide a critical overview of our combination of manual and automatic analysis. Finally, we discuss the significance of our results and how they can be applied in a number of areas aiming at improving literature-mining solutions for life sciences research.

5.1. The Internal Structure of Citances. As we have seen above, the relationship between citances and citations is not always 1 : 1, for example, in some cases, a citance would contain citations to multiple target articles. While we acknowledged and analyzed the issue, we still treated citances as *atomic* from the viewpoint of the target article(s), assuming that the whole citance was commenting on it/them.

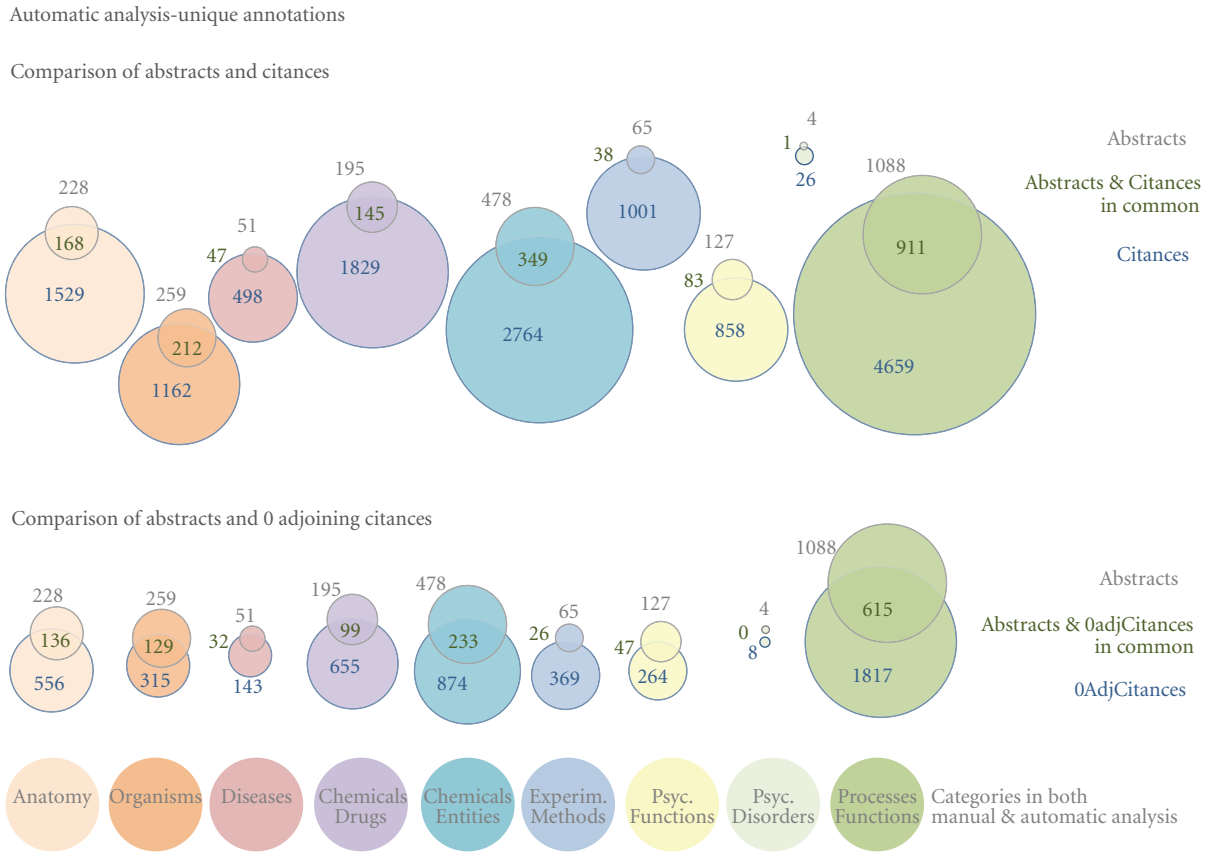


Figure 4: Number of unique concepts found in abstracts, all citances, and citances with 0 adjoining citations. Also shown is the overlap between all citances and abstracts.

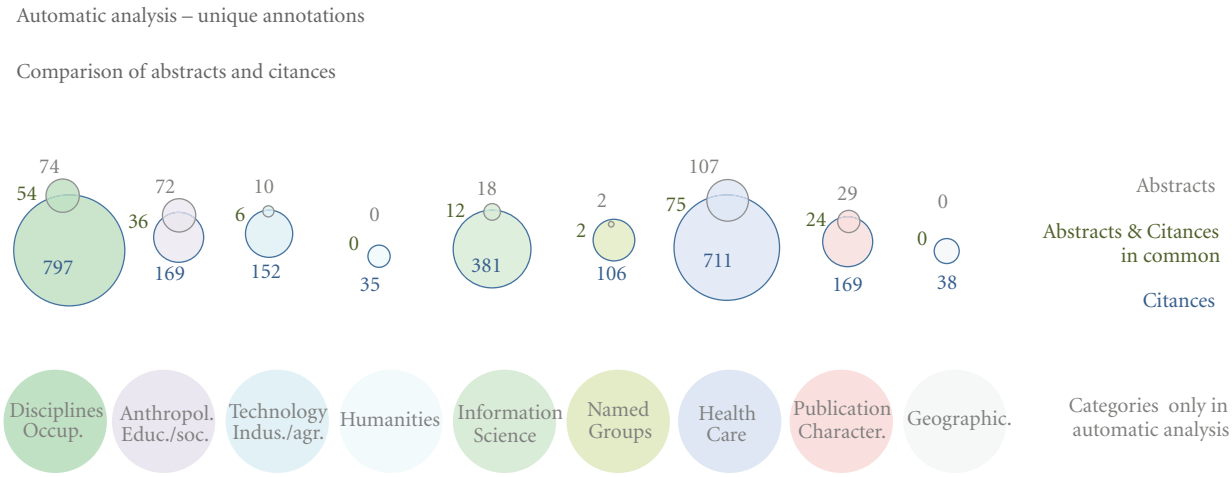


Figure 5: Unique Annotations found in abstracts, citances, and their overlap for the annotation categories defined only in the automatic analysis.

Things are more complicated though: it is often the case that only part of a citance is really relevant. This is similar to HTML pages, where only part of a sentence containing a hyperlink is actually included in the hyperlink. Unfortunately, research publications, unless published in some hyperlink-friendly format, do not use such precise mechanisms for pointing out the relevant part of a citance. Yet, authors of research articles do use citations that refer to part of a citance, which poses interesting challenges to research on citances. See [50] for an overview. Below we list and illustrate three ways in which authors use references:

Type 1. Use separate citations for different parts of the citance.

Table 3: Comparison of the number of distinct annotation types in abstracts and citances with zero adjoining citations. We used all sentences from the 6 abstracts and all 23 citances that were only citing one paper for this analysis.

PubMed ID	Abstract	Abstract and citances_0	Difference	n/a	In full text	In MeSH or substances	Not found
8939603	52	65	13	1	10		2
11346650	52	75	23	3	14		6
11251070	57	73	16	2	3	2	9
11298456	60	71	11		6		5
11850621	61	71	10		9		1
Total	282	355	73	6	42	2	25

Table 4: Number of citances with a different number of adjoining citations in each article and the number of distinct annotation types they contain. These statistics are for the manual analysis. For the automatic analysis, see Figure 4 and the supplementary material.

PMID	Citance number						Distinct annotation types (abstract and citances)					
	All cit.	Cit_0	Cit_1	Cit_2	Cit_3	Cit_4+	All cit.	Cit_0	Cit_1	Cit_2	Cit_3	Cit_4+
8939603	51	3	8	12	10	18	121	65	68	63	87	85
11346650	45	7	3	4	7	24	170	75	66	66	73	144
11125146	10	0	6	3	1	0	80		67	65	43	
11251070	10	7	0	0	0	3	88	73				73
11298456	10	3	3	2	0	2	96	71	72	66		70
11850621	10	3	4	1	0	2	98	71	76	67		71
Total	136	23	24	22	18	49	653	355	349	327	203	443

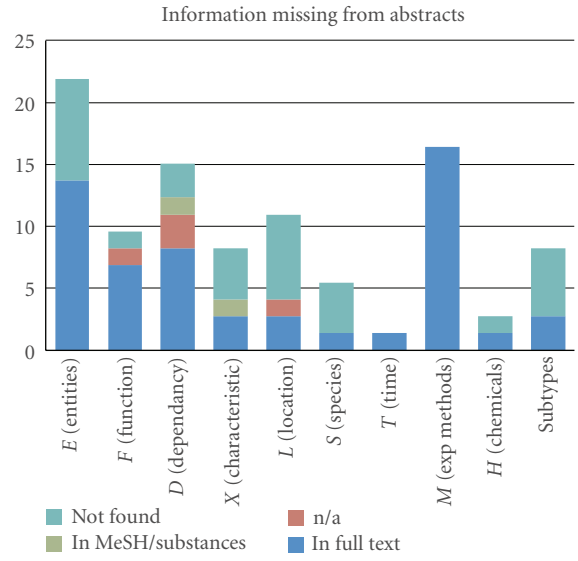


Figure 6: Categories of distinct manual annotation types not found in abstracts.

Example. Subsequently, it has been observed that a similar motif is present also in substrates like Cdc6 [21] and retinoblastoma family of proteins [22] and the activator Cdc25A [13].

Type 2. Use citation(s) for part of the citance only.

Example. The nucleosolic or nonutilized Cdc6 then could either be translocated to the cytoplasm (10, 11, 16, 28, 33) or have its affinity for chromatin reduced but still remain in the nucleus (as our immunohistochemical and biochemical data would suggest); this would prevent inappropriate pre-RC formation and reinitiation of DNA replication.

Type 3. List multiple references together at the end of the citance.

Example. These and other biochemical and genetic studies in Drosophila and Xenopus demonstrate that the ORC functions in chromosomal DNA replication in multicellular eukaryotes, just as it does in yeast (25, 28–30, 48, 49).

Citances of Type 2 might have been the reason that a number of biological concepts mentioned in citances were not found in the full text of the target citations. Additionally, we could have used citances of Type 1 to detect more accurately the origin of the information in citances.

Notwithstanding that having considered this variation in citance structure would had enabled us to determine the source of information more accurately, as we discussed in the related work section, a lot of work has been done on the basis that references that appear together are related. Therefore, any additional information from other references can be used to augment the information from the target citation.

Finally, we should note that even knowing when a sentence contains a citation is a challenging task by itself since citation markers can differ in style. Moreover, even after a citation has been identified in text, resolving its target article is not a trivial task. For a further discussion on these issues, see [51–53].

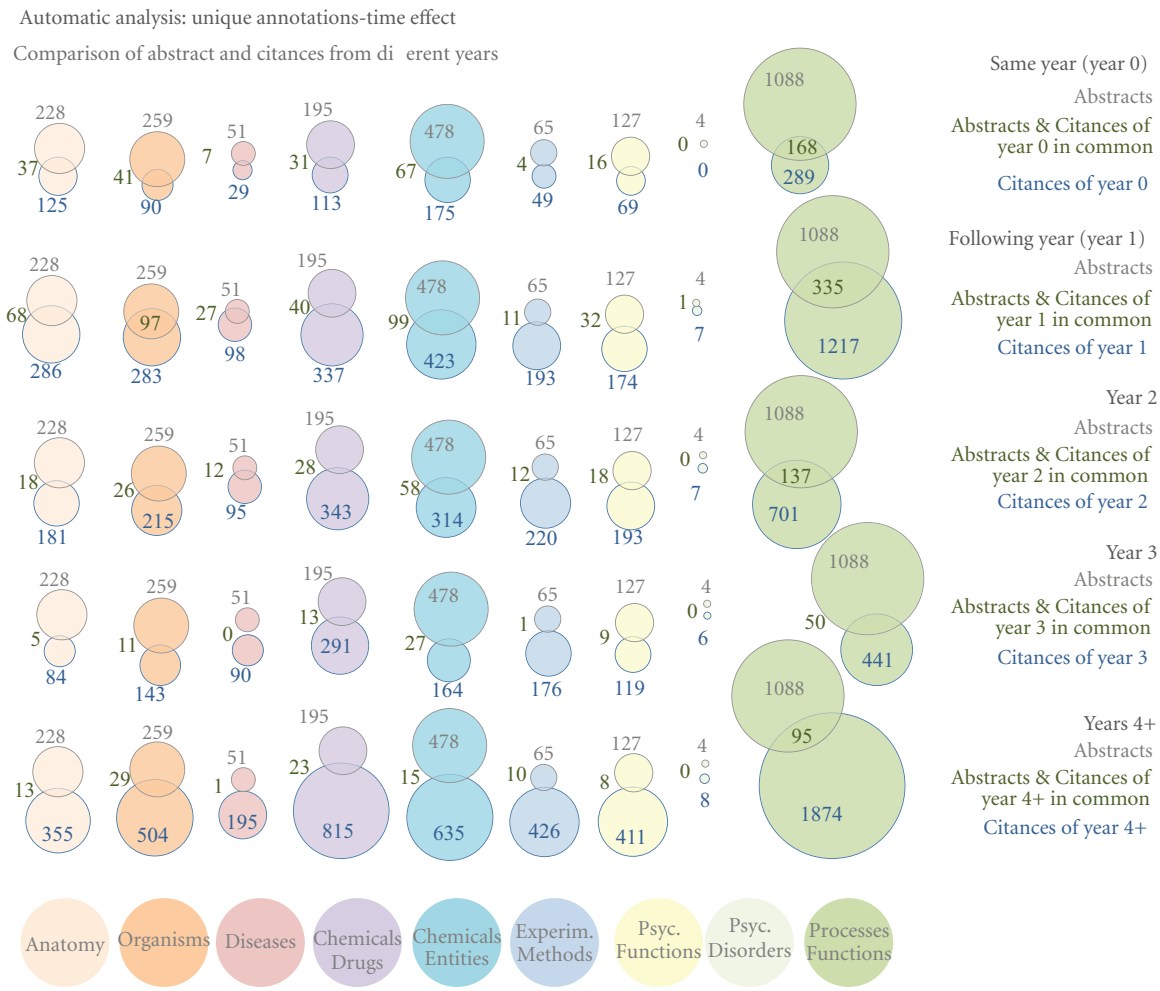


Figure 7: The effect of time. We show the unique semantic categories mentioned in the citances from the same publication year as the original target paper and how they overlap with the semantic categories matched in the target abstracts. Semantic annotations and overlap with the abstract for the following 1, 2, 3, and 4+ years are also shown. Note that only *new* unique semantic annotations are counted, for example, annotations of "citances of year 2" do not include any annotations that already appeared in years 0 or 1.

5.2. Combining Manual and Automatic Analysis. We strived to map the categories of our manual schema to the automatic annotation schema the best possible way, while keeping them pertinent to the area of molecular interactions. Despite the significant overlap between these two schemata, the mapping was not ideal, as Figure 2 shows. For example, we could not use MeSH to automatically generate concepts covering events and relations, which were present in the manual annotation. To compensate for this, we added a number of additional concept categories that were easy to identify in MeSH, for example, disciplines, humanities, healthcare, and so forth (see Table 2, Figure 2).

Another issue with the automatic analysis was that the 1 : 1 mapping to the concept categories for the manual annotation was not possible since MeSH categories did not always align perfectly to our concepts. On the positive side, we relied on MeSH, which is a standard resource that is widely used in biomedical text mining. It provides many variants and synonyms for the concepts it covers, which allows us to handle the variety in expression that is inherent in natural language. Moreover, the MeSH concepts are organized in a hierarchical structure, which allows for a very easy mapping of whole subtrees to predefined categories; in the ideal case, all that is needed to define the mapping is to find the correct level of generalization in MeSH. Table 2 shows how this was done in our case.

5.3. Using Semantic Annotations Found in Citances to Augment Annotations in Abstracts. While studying the effect of adjoining citations, we found that the majority of citances' unique annotation IDs that overlap with unique annotations found in the original target abstract can indeed be found in

citances with 0 adjoining citations. This means that citances that cite multiple papers can be used to complement the abstract of each citing paper with more annotations. Imagine that an abstract with 20 semantic annotations assigned to it has 0adj_citances with 30 annotations and 15 of them overlap with the abstract annotations. Now, we have 15 more annotations that can be mapped to the abstract. The target paper is about to have 1+adj_citances that can be associated with a larger number of annotations, say 60; these new annotations can now also be associated with the original paper.

Much like modern media's social boosting from users assigning tags, these new annotations provided by expert peers can be used to help various NLP tasks. Here we propose utilization of these annotations for document summarization, document ranking, and automatic biological database annotation.

In the case of document summarization where most related work has concentrated on, we observe the following opportunities (1) A way to expand information by combining (union) the citances, which contain the best representative information from the full text (rich peer-produced resource), with the abstract (author-produced resource)—this would offer the best complete, inclusive summary. (2) A way to narrow down the information by using the intersection of the information found in citances and abstracts, especially years later—this would offer the most distilled, concentrated summary. (3) A way to generate a summary for a paper, even when its abstract and/or full text are not available in electronic form—that is, use just the citances.

In the case of document/sentence ranking, the density of these annotations in a sentence (or, alternatively, the category/type of annotation, or the relationship of the annotations to the original source) can be used to boost a weight-based ranking system.

Furthermore, our approach can be extended to other standardized resources (e.g., GO and UMLS) that are often used in biomedical databases to automatically map normalized entities and concepts to each other as well as to articles.

5.4. The Four Questions. Let us now go back to the four original research questions, keeping in mind that our dataset focused on molecular interactions, a very hot area for literature mining, as it is the main resource for constructing molecular networks and thus answering systems biology questions.

(1) *How informative is the abstract compared to the full text?* We have shown that the information contained in the abstract and in the citances overlap to a large extent. Yet, there is information in the full text that is important enough to be referred to in citances, but it is not included in the abstract. Thus, abstracts cannot substitute the full text since peers cite information from the full text that is not always included in the abstract.

(2) *What important information in the full text does not appear in the abstract?* We have shown that citances contain additional information that does not appear in abstracts. Since this information appears in a citance, then (1) it should come from the full text, and (2) it should be seen by peers as important. We studied several categories of biologically meaningful concepts and we found that citances contained more information for each of these categories; still, the differences were most pronounced for biological entities and experimental methods.

(3) *What should a summary of the full text contain that is not already in the abstract?* We believe that a good summary of an article should combine the information from its abstract and from citances. Citances give the viewpoint of multiple peers and are thus a very valuable information source. Our study has found that citances tend to mention more biological entities and to care more about experimental methods than authors do in their abstracts. Thus, we would recommend that summaries pay more attention to molecular entities and even consider including information on methods.

(4) *What are the differences in the way authors and peers see an article?* Authors' viewpoint is summarized in the articles abstract, while peers' viewpoint is reflected in the citances to that article. Thus, articles are author-biased, while the set of citances, which are produced by many peers, is more objective. Moreover, citances are written years after the article was published, which also contributes to a more objective view to the contribution of an article: we have seen that, in the first year peers largely agreed with the authors, while differentiation was observed later when the citances have become arguably more divergent in content than the original target paper. The overlapping information though (found both in abstracts and in citances from years later) can be perceived as the most interesting, as it remains pertinent scientifically years later. Overall, we have found that authors focused in their abstracts on a smaller number of concepts compared to their peers. Moreover, peers tended to pay more attention to experimental methods compared to authors.

5.5. Future Directions. In future work, we would like to do a more careful study that would cover more and finer-grained categories in MeSH; trying resources like UMLS and GO is another attractive option. Looking at facts of larger granularity than just concepts, for example, looking at predicate-argument relations is another interesting direction for future work. We further plan to analyze the internal structure of citances, so that we can identify which part of the citance is relevant to a given citation. It would be also interesting to try similar analysis for other disciplines and areas of science, where the way research publications are written and the number of citations a publication receives may differ a lot from what we observe in life sciences.

Another interesting aspect is the passage of time. We have seen that while early citations tend to agree with the authors, later ones tended to diverge more from the original abstract. It would be interesting to see whether this means that later citations are really more objective. An important tool in this respect would be to look at the repetitiveness of citations, which we ignored in our present study, where we focused on unique concept mentions instead: if many peers stated the same fact, then maybe it should be deemed not only more important, but also more objective. Peer motivation for citing an article is important as well, for example, citations that cite a fact would probably agree with the abstract more than those that criticize it.

Last but not least, we are interested in using citances to help NLP applications. While previous work has already shown a number of such examples including information retrieval [24, 25], document summarization [19, 39, 40], document categorization [23], document clustering [29], MeSH terms assignment [30], relation extraction [27], and automatic paraphrasing [17], we believe that this list can be extended significantly.

6. Conclusion

Citances tell us what peers see as contributions of a given target article, while abstracts reflect the authors viewpoint on what is important about their work. Unlike citances, which typically focus on a small number of important aspects, abstracts serve a more general purpose: they not only state the contributions, but also provide a summary of the main points of the paper; thus, abstracts tend to be generally broader than citances. Yet, our manual and automatic comparison of abstracts and citances for articles describing molecular interactions has shown that, collectively, citances contain more information than abstracts.

We performed manual evaluation, which revealed that while all concepts in an article's abstract could be found in the citances for that article (provided that the article has already accumulated enough citations), the reverse was not true: citances mentioned about 20% more concepts than abstracts. Assuming that any information that is not mentioned in the abstract but is important enough to be referred to in citances should be coming from the full text, we can conclude that full text contains important information that is not mentioned in the abstract. We did not detect any significant changes in concept mentions over time.

The automatic analysis verified the results of the manual analysis on a larger scale, using MeSH terms, which were automatically mapped to the biological concepts from the manual analysis. These experiments confirmed our findings that most concepts mentioned in abstracts can be also found in citances. They further confirmed that citances contained some additional information, which in our case was primarily related to biological entities and experimental methods. The large-scale analysis has shown that the manual analysis could indeed be automated; the approach can be extended to other commonly-used biomedical resources such as GO and UMLS, which allow for uniform representation of concepts, that is, useful information about the semantic relationship between abstracts and citation sentences and among concepts themselves.

Overall, our results show that citances are good surrogates of the information contained in a biomedical journal article. The set of all citances citing a given research publication can be seen as concise summaries of its important contributions and thus using them can be preferable to the full text in a variety of scenarios. For example, they allow text mining applications to concentrate on potentially useful sentences without the need to deal with the full text, which is long, has a complex structure, and often would not be available at all, for example, for older publications. Since our work was based on biologically meaningful semantic concepts, it provides quantitative justification of their usefulness for text mining as it has been observed in previous work [17, 27, 30].

We can conclude that, with the recent growth of free access to journal articles and open access publications, full text should be seriously considered for yet another reason: it contains citances with information on the publications referenced therein. Peers cite (mention and comment) information that they see as important even if it is not mentioned in the original publication's abstract. We would further like to draw special attention to citances, as a good source of concise, verifiable information on molecular interaction networks. To answer the question posed by our title "Do Peers See More in a Paper than its Authors?": yes they do, and we should leverage this information.

Acknowledgments

The authors would like to thank the anonymous reviewers for their constructive comments, which has led to great improvements in this paper. This research was funded in part by NSF Grant DBI-0317510.

Endnotes

1. http://www.ncbi.nlm.nih.gov/pubmed/
2. http://www.wellcome.ac.uk/
3. http://www.ncbi.nlm.nih.gov/pmc/
4. http://grants.nih.gov/grants/guide/notice-files/NOT-OD-08-033.html
5. http://www.biomedcentral.com/
6. http://www.plos.org/
7. Our study also helps answer the question: *what abstract claims are not (strongly) supported by the full text*? We hypothesize that these would be those claims that are cited very infrequently or not cited at all, but a separate study is required to answer this question.
8. Note that here we assume that peers base their citations on full text and not only on the abstract. While this is a strong assumption, we believe that it generally holds in the research community. Our previous studies have shown that biomedical researchers like to verify reported results, for example, by looking at the methods that were used and by exploring the images and the

tables in the full text. This has also motivated us to create a specialized search engine, the BioText Search Engine (http://biosearch.berkeley.edu/), for searching the figures and tables contained in open access journals, which is described in [54, 55].

9. CiteSeerX: http://citeseer.ist.psu.edu/
10. DBLP: http://www.informatik.uni-trier.de/~ley/db/
11. Google Scholar: http://scholar.google.com/
12. Microsoft Academic Search: http://academic.research.microsoft.com/
13. ACM Digital Library: http://dl.acm.org/
14. IEEE Xplore Digital Library: http://ieeexplore.ieee.org/Xplore/
15. ACL Anthology: http://aclweb.org/anthology-new/
16. ArnetMiner: http://arnetminer.org/
17. EMNLP 2009: http://conferences.inf.ed.ac.uk/emnlp09/
18. http://www.nlm.nih.gov/mesh/
19. http://discover.nci.nih.gov/mim/index.jsp
20. http://isiknowledge.com/
21. The data on the analysis considering the extended tree IDs can be found in the supplementary material available online at doi:10.1155/2012/750214. The majority of results discussed in this paper refer to higher MeSH level annotation representing broader entities and concepts.

References

[1] I. Mani and M. Maybury, *Advances in Automatic Text Summarization*, MIT Press, 1999.

[2] H. Yu, V. Hatzivassiloglou, C. Friedman, A. Rzhetsky, and W. Wilbur, "Automatic extraction of gene and protein synonyms from MEDLINE and journal articles," in *Proceedings of the AMIA Symposium (AMIA '02)*, pp. 919–923, 2002.

[3] P. K. Shah, C. Perez-Iratxeta, P. Bork, and M. A. Andrade, "Information extraction from full text scientific articles: where are the keywords?" *BMC Bioinformatics*, vol. 4, article 20, 2003.

[4] M. J. Schuemie, M. Weeber, B. J. A. Schijvenaars et al., "Distribution of information in biomedical abstracts and full-text publications," *Bioinformatics*, vol. 20, no. 16, pp. 2597–2604, 2004.

[5] H. T. Dang, "Overview of DUC 2005," in *Proceedings of the HLT/EMNLP Workshop on Text Summarization DUC*, 2005.

[6] M. Walport and R. Kiley, "Open access, UK PubMed central and the wellcome trust," *Journal of the Royal Society of Medicine*, vol. 99, no. 9, pp. 438–439, 2006.

[7] K. B. Cohen, H. L. Johnson, K. Verspoor, C. Roeder, and L. E. Hunter, "The structural and content aspects of abstracts versus bodies of full text journal articles are different," *BMC Bioinformatics*, vol. 11, article 492, 2010.

[8] E. Garfield, "Can citation indexing be automated," *National Bureau of Standards Miscellaneous Publication*, vol. 269, pp. 189–192, 1965.

[9] M. Liu, "Progress in documentation. the complexities of citation practice: a review of citation studies," *Journal of Documentation*, vol. 49, no. 4, pp. 370–408, 1993.

[10] M. Moravcsik and P. Murugesan, "Some results on the function and quality of citations," *Social Studies of Science*, vol. 5, pp. 86–92, 1975.

[11] E. Garfield, "Citation indexes for science," *Science*, vol. 122, no. 3159, pp. 108–111, 1955.

[12] C. L. Giles, K. D. Bollacker, and S. Lawrence, "CiteSeer: an automatic citation indexing system," in *Proceedings of the 3rd ACM Conference on Digital Libraries*, pp. 89–98, ACM Press, June 1998.

[13] F. Menczer, "Correlated topologies in citation networks and the Web," *European Physical Journal B*, vol. 38, no. 2, pp. 211–221, 2004.

[14] M. E. J. Newman, "The structure of scientific collaboration networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 2, pp. 404–409, 2001.

[15] C. Duy, V. Hoang, and M.-Y. Kan, "Towards automated related work summarization," in *Proceedings of the 23rd International Conference on Computational Linguistics (COLING '10)*, pp. 427–435, Posters, 2010.

[16] H. D. White, "Citation analysis and discourse analysis revisited," *Applied Linguistics*, vol. 25, no. 1, pp. 89–116, 2004.

[17] P. Nakov, A. Schwartz, and M. Hearst, "Citances: citation sentences for semantic analysis of bioscience text," in *Proceedings of the Workshop on Search and Discovery in Bioinformatics (SIGIR '04)*, 2004.

[18] A. Elkiss, S. Shen, A. Fader, G. Erkan, D. States, and D. Radev, "Blind men and elephants: what do citation summaries tell us about a research article?" *Journal of the American Society for Information Science and Technology*, vol. 59, no. 1, pp. 51–62, 2008.

[19] S. Mohammad, B. Dorr, M. Egan et al., "Using citations to generate surveys of scientific paradigms," in *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL '09)*, pp. 584–592, Boulder, Colo, USA, 2009.

[20] H. D. White and B. C. Griffith, "Author cocitation: a literature measure of intellectual structure," *Journal of the American Society for Information Science*, vol. 32, no. 3, pp. 163–171, 1981.

[21] A. Aris, B. Shneiderman, V. Qazvinian, and D. Radev, "Visual overviews for discovering key papers and influences across research fronts," *Journal of the American Society for Information Science and Technology*, vol. 60, no. 11, pp. 2219–2228, 2009.

[22] S. Teufel and M. Moens, "Summarizing scientific articles: experiments with relevance and rhetorical status," *Computational Linguistics*, vol. 28, no. 4, pp. 409–445, 2002.

[23] H. Nanba, N. Kando, and M. Okumura, "Classification of research papers using citation links and citation types: towards automatic review article generation," in *Proceedings of the American Society for Information Science SIG Classification Research Workshop: Classification for User Support and Learning*, pp. 117–134, 2000.

[24] S. Bradshaw, "Reference directed indexing: redeeming relevance for subject search in citation indexes," in *Proceedings of the 7th European Conference on Research and Advanced Technology for Digital Libraries*, 2003.

[25] R. Mercer and C. Di Marco, "A design methodology for a biomedical literature indexing tool using the rhetoric of science," in *Proceedings of the BioLink Workshop in Conjunction with NAACL/HLT*, pp. 77–84, 2004.

[26] I. Tbahriti, C. Chichester, F. Lisacek, and P. Ruch, "Using argumentation to retrieve articles with similar citations: an inquiry into improving related articles search in the MEDLINE digital library," *International Journal of Medical Informatics*, vol. 75, no. 6, pp. 488–495, 2006.

[27] B. Rosario and M. Hearst, "Multi-way relation classification: application to protein-protein interactions," in *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT '05)*, 2005.

[28] A. Kolchinsky, A. Abi-Haidar, J. Kaur, A. A. Hamed, and L. M. Rocha, "Classification of protein-protein interaction full-text documents using text and citation network features," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 7, no. 3, pp. 400–411, 2010.

[29] B. Aljaber, N. Stokes, J. Bailey, and J. Pei, "Document clustering of scientific texts using citation contexts," *Information Retrieval*, vol. 13, no. 2, pp. 101–131, 2010.

[30] B. Aljaber, D. Martinez, N. Stokes, and J. Bailey, "Improving MeSH classification of biomedical articles using citation contexts," *Journal of Biomedical Informatics*, vol. 44, no. 5, pp. 881–896, 2011.

[31] W. Lehnert, C. Cardie, and E. Riloff, "Analyzing research papers using citation sentences," in *Proceedings of the 12th Annual Conference of the Cognitive Science Society*, pp. 511–518, Lawrence Erlbaum Associates, 1990.

[32] S. Teufel, A. Siddharthan, and D. Tidhar, "An annotation scheme for citation function," in *Proceedings of Sigdial-06*, Sydney, Australia, 2006.

[33] S. Teufel, A. Siddharthan, and D. Tidhar, "Automatic classification of citation function," in *Proceedings of EMNLP-06*, Sydney, Australia, 2006.

[34] S. Teufel and M. Y. Kan, "Robust argumentative zoning for sensemaking in scholarly documents," in *Advanced Language Technologies for Digital Libraries*, vol. 6699 of *Lecture Notes in Computer Science*, pp. 154–170, Springer, Berlin, Germany, 2011.

[35] C. Schwartz, A. Divoli, and M. Hearst, "Multiple alignment of citation sentences with conditional random fields and posterior decoding," in *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL '07)*, pp. 847–857, 2007.

[36] E. Amitay and C. Paris, "Automatically summarising web sites: is there a way around it," in *Proceedings of the 9th International Conference on Information and knowledge Management*, pp. 173–179, ACM Press, 2000.

[37] J. Y. Delort, B. Bouchon-Meunier, and M. Rifqi, "Enhanced web document summarization using hyperlinks," in *Proceedings of the 14th ACM Conference on Hypertext and Hypermedia*, pp. 208–215, August 2003.

[38] A. Schwartz and M. Hearst, "Summarizing key concepts using citation sentences," in *Proceedings of the Workshop on Linking Natural Language Processing and Biology: Towards Deeper Biological Literature Analysis (BioNLP '06)*, pp. 134–135, New York, NY, USA, 2006.

[39] V. Qazvinian and D. Radev, "Scientific paper summarization using citation summary networks," in *Proceedings of the 22nd International Conference on Computational Linguistics (COLING '08)*, vol. 1, pp. 689–696, Manchester, UK, 2008.

[40] Q. Mei and C. Zhai, "Generating impact-based summaries for scientific literature," in *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL '08)*, pp. 816–824, Columbus, Ohio, USA, 2008.

[41] V. Qazvinian and D. Radev, "Identifying non-explicit citing sentences for citation-based summarization," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguisticsproceedings of (ACL '10)*, pp. 555–564, 2010.

[42] S. Wan, C. Paris, and R. Dale, "Whetting the appetite of scientists: producing summaries tailored to the citation context," in *Proceedings of theACM/IEEE Joint Conference on Digital Libraries (JCDL '09)*, pp. 59–68, June 2009.

[43] N. Craswell, D. Hawking, and S. Robertson, "Effective site finding using link anchor information," in *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 250–257, ACM Press, 2001.

[44] S. Chakrabarti, B. Dom, P. Raghavan, S. Rajagopalan, D. Gibson, and J. Kleinberg, "Automatic resource compilation by analyzing hyperlink structure and associated text," in *Proceedings of the 7th International Conference on World Wide Web 7*, pp. 65–74, Elsevier Science Publishers B.V., 1998.

[45] J. Fürnkranz, "Exploiting structural information for text classification on the www," in *Proceedings of the 3rd International Symposium on Advances in Intelligent Data Analysis*, pp. 487–498, Springer, 1999.

[46] J. Rennie and A. McCallum, "Using reinforcement learning to spider the web efficiently," in *Proceedings of the 16th International Conference on Machine Learning*, pp. 335–343, Morgan Kaufmann Publishers, 1999.

[47] M. Richardson and P. Domingos, "The intelligent surfer: probabilistic combination of link and content information in pagerank," in *Proceedings of the Advances in Neural Information Processing Systems*, vol. 14, MIT Press, 2002.

[48] G. Bhalotia, P. Nakov, A. Schwartz, and M. Hearst, "BioText team report for the TREC, 2003 Genomics track," in *Proceedings of the 13th Text REtrieval Conference (TREC '04)*, Gaithersburg, Md, USA, 2004.

[49] P. Nakov and A. Divoli, "BioText report for the second BioCreAtIvE challenge," in *Proceedings of BioCreAtIvE II Workshop*, Madrid, Spain, April 2007.

[50] A. Ritchie, S. Teufel, and S. Robertson, "How to find better index terms through citations," in *Proceedings of the Workshop on How Can Computational Linguistics Improve Information Retrieval?* pp. 25–32, Sydney, Australia, 2006.

[51] D. Bergmark, "Automatic extraction of reference linking information from online documents," Technical Report CSTR 2000-1821, Cornell Digital Library Research Group, 2000.

[52] D. Bergmark, P. Phempoonpanich, and S. Zhao, "Scraping the ACM digital library," *SIGIR Forum*, vol. 35, no. 2, pp. 1–7, 2001.

[53] B. Powley and R. Dale, "Evidence-based information extraction for high-accuracy citation extraction and author name recognition," in *Proceedings of the 8th RIAO International Conference on Large-Scale Semantic Access to Content*, 2007.

[54] M. A. Hearst, A. Divoli, H. H. Guturu et al., "BioText Search Engine: beyond abstract search," *Bioinformatics*, vol. 23, no. 16, pp. 2196–2197, 2007.

[55] A. Divoli, M. A. Wooldridge, and M. A. Hearst, "Full text and figure display improves bioscience literature search," *PLoS ONE*, vol. 5, no. 4, Article ID e9619, 2010.

Hindawi Publishing Corporation
Advances in Bioinformatics
Volume 2012, Article ID 573846, 10 pages
doi:10.1155/2012/573846

Review Article

Literature Retrieval and Mining in Bioinformatics: State of the Art and Challenges

Andrea Manconi,[1] Eloisa Vargiu,[2,3] Giuliano Armano,[2] and Luciano Milanesi[1]

[1] *Institute for Biomedical Technologies, National Research Council, Via F.lli Cervi, 93, 20090 Segrate, Italy*
[2] *Department of Electrical and Electronic Engineering, University of Cagliari, Piazza d'Armi, 09123 Cagliari, Italy*
[3] *Barcelona Digital Technological Center, C/Roc Boronat 117, 08018 Barcelona, Spain*

Correspondence should be addressed to Andrea Manconi, andrea.manconi@itb.cnr.it

Received 22 November 2011; Revised 18 May 2012; Accepted 18 May 2012

Academic Editor: Jörg Hakenberg

The world has widely changed in terms of communicating, acquiring, and storing information. Hundreds of millions of people are involved in information retrieval tasks on a daily basis, in particular while using a Web search engine or searching their e-mail, making such field the dominant form of information access, overtaking traditional database-style searching. How to handle this huge amount of information has now become a challenging issue. In this paper, after recalling the main topics concerning information retrieval, we present a survey on the main works on literature retrieval and mining in bioinformatics. While claiming that information retrieval approaches are useful in bioinformatics tasks, we discuss some challenges aimed at showing the effectiveness of these approaches applied therein.

1. Introduction

Nowadays, most of the scientific publications are electronically available on the Web, making the problem of retrieving and mining documents and data a challenging task. To this end, automated document management systems have gained a main role in the field of intelligent information access [1]. Thus, research and development in the area of bioinformatics literature retrieval and mining is aimed at providing intelligent and personalized services to biologists and bioinformaticians while searching for useful information in scientific publications. In particular, the main goal of bioinformatics text analysis is to provide access to unstructured knowledge by improving searches, providing automatically generated summaries, linking publications with structured resources, visualizing contents for better understanding, and guiding researchers to formulate novel hypotheses and to discover knowledge.

In the literature, several methods, systems, and tools to retrieve and mine bioinformatics publications have been proposed and adopted, some of them being currently available on the Web. In this paper, we provide a survey of existing end-user-oriented literature retrieval and/or mining solutions for bioinformatics, together with a short discussion on open challenges. The rest of the paper is organized as follows: Section 2 illustrates the main topics addressed in this paper, that is, information retrieval, text mining, and literature retrieval and mining. In Section 3, the state of the art on literature retrieval and mining in bioinformatics is presented. Section 4 discusses some relevant open problems and challenges. Section 5 ends the paper.

2. Background

Supporting users in handling the huge and widespread amount of Web information is becoming a primary issue. Information retrieval is the task of representing, storing, organizing, and accessing information items. Information retrieval has considerably changed in recent years: initially with the expansion of the World Wide Web and the advent of modern and inexpensive graphical user interfaces and mass storage [2], and then with the advent of modern Internet technologies [3] and of the Web 2.0 [4].

Information retrieval can cover various and heterogeneous kinds of data and information problems beyond that specified in the core definition above. More generally, an information retrieval system does not inform (i.e., does not change the knowledge of) the user on the subject of her/his inquiry. It merely informs on the existence (or nonexistence) and whereabouts of documents relating to her/his request. According to [5], information retrieval is defined as the task of finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need, from large collections (usually stored on computers). Nowadays, information retrieval solutions rely on the adoption of Web services and suitable Semantic Web approaches, such as ontologies. Indeed, Semantic Web inference can improve traditional text search, and text search can be used to facilitate or augment Semantic Web inference [6].

Text Mining is an information retrieval task aimed at discovering new, previously unknown information, by automatically extracting it from different text resources [7]. In fact, the term "text mining" is generally used to denote any system that analyzes large quantities of natural language text and detects lexical or linguistic usage patterns in an attempt to extract probably useful (although only probably correct) information [8]. Automatic extraction of metadata (e.g., subjects, language, authors, key-phrases) is a prime application of text mining techniques. Although contemporary automatic document retrieval techniques bypass the metadata creation stage and work directly on the full-text of the documents, text mining has been largely applied to learn metadata from documents. Language identification is a relatively simple mining task aimed at providing an important piece of metadata for documents in international collections. A simple representation for document categorization is to characterize each document by a profile that consists of "n-grams," that is, sequences of n consecutive words, that appear in it. Occurrence probabilities of common words are then compared with the most frequent words of the text data. Author's metadata is one of the primary attributes of most documents, and it is usually known. However, in some cases, authorship is uncertain and must be guessed from the text. Text mining is also applied to provide summaries of documents or groups of documents. Text summarization is aimed at producing a condensed representation of its input, intended for human consumption [9]. Earliest instances of research on summarization of scientific documents extract salient sentences from text using features like word and phrase frequency [10], positions in the text [11], and key phrases [12]. Various works published since then had concentrated on other domains, mostly on newswire data [13] and contextual advertising [14]. Overall, summarization techniques can be divided in two groups [15]: those that extract text containing the most relevant information from the source documents (*extraction-based approaches*) and those that perform paraphrasing on the source documents (*abstraction-based approaches*).

Document clustering is an unsupervised learning technique in which there is no predefined category or class, but groups of documents that belong together are sought. For example, document clustering may assist in retrieval tasks by creating links between similar documents, which in turn allows related documents to be retrieved once one of the documents has been deemed relevant to a query [16]. Although they do not require training data to be preclassified, clustering techniques are generally often more computation intensive than supervised schemes [17]. Nevertheless, clustering has been largely applied in text mining applications. Trials of unsupervised schemes include the work by Aone et al. [18], who use the conceptual clustering scheme COBWEB [19] to induce natural groupings of close-captioned text associated with video newsfeeds; Liere and Tadepalli [20], who explore the effectiveness of AutoClass [21] in producing a classification model for a portion of the Reuters corpus; Green and Edwards [22], who use AutoClass to cluster news items gathered from several sources into *stories* (i.e., groupings of documents covering similar topics). One of the main subfields of text mining is information extraction, that is, the task of filling templates from natural language input [23]. Typical extraction problems address simple relationships among entities, such as finding the predicate structure of a small set of predetermined propositions. Machine learning has been applied to the information extraction task by seeking pattern-match rules that extract fillers for slots in the template [24–27]. Extracted information can be used in a subsequent step to learn rules that characterize the content of the text itself.

In the academic area, online search engines are used to find out scientific resources, as journals and conference proceedings. However, finding and selecting appropriate information on the Web is still difficult. To simplify this process, several frameworks and systems have been developed to retrieve scientific publications from the Web. Bollacker et al. [28] developed CiteSeer (http://citeseer.ist.psu.edu/), the well-known automatic generator of digital libraries of scientific literature. Being aimed at eliminating most of the manual effort of finding useful publications on the Web, CiteSeer uses sophisticated acquisition, parsing, and presentation methods. In particular, CiteSeer uses a three-stage process: database creation and feature extraction; personalized filtering of new publications; personalized adaptation and discovery of interesting research and trends. These functions are interdependent: information filtering affects what is discovered, whereas useful discoveries tune the information filtering. In [29], the authors study how to recommend research publications using the citation between publications to create a user-item matrix. In particular, they test the ability of collaborative filtering to recommend citations that could be additional references for a target research publication. Janssen and Popat [30] developed UpLib, a personal digital library system that consists of a full-text indexed repository accessed through an active agent via a Web interface. UpLib is mainly concerned with the task of collecting personal collections comprising tens of thousands of documents. In [31], Mahdavi et al. start from the assumption that trend detection in scientific publication retrieval systems helps scholars to find relevant, new and popular special areas. To

this end, they developed a semiautomatic system based on a semantic approach.

3. State of the Art

A great deal of biological information accumulated through years is currently available in online text repositories such as Medline. These resources are essential for biomedical researchers in their everyday activities to plan and perform experiments and verify the results.

Among other kinds of information, let us concentrate on publications and scientific literature, largely available on the Web for any topic. As for bioinformatics, the steady work of researchers, in conjunction with the advances in technology (e.g., high-throughput technologies), has arisen in a growing amount of known sequences. The information related with these sequences is daily made available as scientific publications. Digital archives like BMC Bioinformatics (http://www.biomedcentral.com/bmcbioinformatics/), PubMed Central (http://www.pubmedcentral.gov/) and other online journals and resources are more and more searched for by bioinformaticians and biologists, with the goal of retrieving publications relevant to their scientific interests. For researchers, it is still very hard to find out which publications are of interest without an explicit classification of the relevant topics they describe. Thus, these resources must provide adequate mechanisms for retrieving the required information.

3.1. Literature Retrieval in Bioinformatics. Discovering and accessing the appropriate bioinformatics resource for a specific research task has become increasingly important, as suggested in earlier reports [32]. To address this issue, various significant projects and initiatives have been carried out, leading to several pioneering indexes of bioinformatics resources that are currently available on the Internet. Available search engines can be categorized according to different criteria. In particular, in agreement with [33], search engines can be categorized in three groups, depending on the way a query is performed: (i) those that perform the query only in the fields of citations; (ii) those that perform the query in the full text article; (iii) those that further process the retrieved citations to organize them and/or to retrieve further information.

As for the first category, let us recall here RefMed [34], MedlineRanker [35], and iPubMed [36]. RefMed (http://dm.postech.ac.kr/refmed/) is a search engine for PubMed that provides relevance ranking. It is widely known that ranking according to the global importance often does not meet the user interests. Given a starting keyword-based query, an initial list of results is presented to the user, who analyzes the proposed documents and passes judgment on their relevance. Then, RefMed induces a new ranking according to the user judgment by exploiting a machine learning algorithm (i.e., RankSVM [37]). MedlineRanker (http://cbdm.mdc-berlin.de/tools/medlineranker/) and iPubMed (http://ipubmed.ics.uci.edu/) are search engines for Medline. For a given topic, the former learns the most discriminative words by comparing a set of abstracts provided by the user with the whole Medline (or a subset). Then, it ranks abstracts according to the learned discriminative words. The latter, which implements the *search-as-you-type* paradigm, has the main advantage to provide results on the fly, which allows users to dynamically modify their query.

eTBLAST [38] and QUERTLE [39] belong to the second category. eTBLAST (http://etest.vbi.vt.edu/etblast3/) allows searching for both citations (i.e., Medline) and full-text articles (i.e., PubMed Central). To retrieve useful documents, it performs a text-similarity search by comparing documents in a target database with an input text. In doing so, it finds the documents that best match the keywords extracted from the query by analyzing the word alignment.

QUERTLE (http://www.quertle.info/) is a new semantic search engine able to perform queries on PubMed, Toxline, National Institutes of Health Re-PORTER, PubMed Central and BioMed Central. Unlike the above-mentioned systems, QUERTLE is able to perform queries based on the meaning and the context of documents. It exploits a meta-database of subject-verb-object relationships asserted by the authors and automatically extracted using semantic-based linguistics. The search engine matches the user query against these relationships.

Finally, GoPubMed [40], XploreMed [41], EBIMed [42], and iHOP [43] are search engines that belong to the third category. GoPubMed allows (http://www.gopubmed.com/web/gopubmed/) submitting keywords to PubMed, extracts Gene Ontology (GO) terms from the retrieved abstracts (GO is becoming a standard for gene/protein function annotation), and supplies the user with the relevant ontology for browsing. It indexes PubMed search results with ontological background knowledge, such as GO and MeSH. The approach to search can also help to answer questions. In particular, the summary of important terms in "top five & more" is a most helpful feature for answering questions or reducing the big initial result to a smaller set of relevant publications in one click. XploreMed (http://www.ogic.ca/projects/xplormed/) filters PubMed results according to the eight main MeSH (http://www.nlm.nih.gov/mesh/) categories and then extracts topic keywords and their cooccurrences, with the goal of extracting abstracts. EBIMed (http://www.ebi.ac.uk/Rebholz-srv/ebimed/) combines information retrieval and extraction from Medline. It analyzes retrieved Medline abstracts to highlight associations among UniProtKB/Swiss-Prot proteins, GO terms, drugs, and species. All identified terms, sentences, and abstracts are displayed in tables, and all terms are linked to their entries in biomedical databases. iHOP (http://www.ihop-net.org/UniPub/iHOP/) uses genes and proteins as hyperlinks among sentences and abstracts. It converts the information in PubMed into navigable resources. The navigation along gene network allows a stepwise and controlled exploration of the information space. Each step through the network produces information about one single gene and its interactions.

3.2. Literature Mining in Bioinformatics. Given the growth of biomedical information on the Internet, Web-based tools

capable of mining the public databases and of highlighting their relevant information in a well-organized and coherent manner are more and more required. Tanabe et al. [44] have proposed MedMiner (MedMiner is no longer available), an Internet-based hypertext program able to filter and organize large amounts of textual and structured information returned from public search engines—like GeneCards (http://www.genecards.org/) and PubMed. MedMiner offered a potentially significant new aid for coping with the torrent of molecular biology data confronting today researchers. By filtering and organizing material retrieved from high-quality Internet sites, it makes complex database searches much easier to execute and digest. MedMiner successfully integrated public and local databases, using a local database as a "proxy" to the (much larger) public ones. Additional databases could be merged into the system, integrating a wider variety of filters with a consistent user interface. PubCrawler (http://pubcrawler.gen.tcd.ie/) is a free alerting service that scores daily updates to the NCBI Medline [45] and GenBank databases. PubCrawler can keep scientists informed of the current contents of Medline and GenBank by listing new database entries that match their research interests.

To facilitate retrieval and analysis of the huge amount of data contained in documents on biological and medical data, several researchers developed dedicated information extraction systems that attempt to simplify the underlying tasks. Most of the corresponding works use the abstract only, owing to the convenience of access and the quality of data.

As abstracts provide a concise summarization of a publication, very useful to categorize it. On the other hand, analyzing full text is essential to detect all detailed information (e.g., methods, tables, and figures). Abstracts are generally available through central collections with easy direct access (e.g., PubMed). Full texts are distributed across many locations (e.g., publishers websites, journal websites, and local repositories), making their access more difficult [46].

Interactions between proteins are very important to understand their functions and biological processes. Several approaches and tools have been defined to deal with this challenge. Thomas et al. [47] present a system aimed at extracting occurrences of protein interactions from Medline abstracts, producing a database of protein pairs characterized by a type of interaction. To this end, the authors customized the Highlight system, a general purpose information extraction engine for commercial applications [47]. The main customizations of highlight consist of (i) adapting the natural language component to make it able to correctly recognize the relevant entities and events (ii) developing a set of templates or outlines of the kinds of relevant information, and (iii) developing patterns aimed at deciding how to slot items and events into templates. PPI Finder (liweilab.genetics.ac.cn/tm/) [48] is a web application aimed at mining human protein-protein interactions from PubMed abstracts. It is able to (i) find the genes related to the gene of interest based on their cooccurrence frequencies and (ii) extract the semantic descriptions of interactions from the co-occurring literature by computational linguistic methods. Moreover, PPI Finder maps the known interactions from the widely used PPI databases, with the aim to distinguish between novel and known interactions. PIE (http://pie.snu.ac.kr/) (Protein Interaction information Extraction) is a web application to extract protein-protein interaction sentences from PubMed abstracts as well as user-provided articles. To extract hidden interactions, PIE exploits natural language processing and machine learning techniques.

Another important challenge is to automatically translate biomedical literature text into a structured form. Due the huge increase of biomedical literature, manual annotation databases are often incomplete and inconsistent with the literature [49]. In this perspective, Craven and Kumlien [50] applied machine learning techniques to automatically map information from text sources to structured representations. In particular, the goal of their research is to develop methods that can accurately map information from scientific text sources to structured representations, such as knowledge bases or databases. To this end, they developed a system to automatically extract key facts from scientific texts. Their system could be used as a support to construct and update databases and knowledge bases. The authors used the system in the development of an ontology of localization patterns and to populate the corresponding database with text-extracted facts describing localization patterns of individual proteins. Another application of this system is to provide structured summaries of what is known about biological objects. Moreover, according to Swanson and Smalheiser [51], the system can be used to extract relationships among entities by automatically eliciting information from the literature. PreBIND (http://bind.ca) [52] is a system developed to solve a very specific problem. It has been devised to curate the BIND database. BIND is a database aimed at curating and archiving protein-protein interaction from the literature using a standard data representation. In doing so, PreBind exploits both statistical and rule-based methods. Statistical methods are used to retrieve relevant documents, whereas rule-based methods are used for biomolecule name recognition, with the aim to find statements about protein interactions. Wiegers et al. [53] proposed another tailored solutions. The authors presented a text-mining prototype to curate the Comparative Toxicogenomics Database (CTD), a publicly available resource that promotes understanding about the etiology of environmental diseases. It provides manually curated chemical-gene/protein interactions and chemical- and gene-disease relationships from the peer-reviewed published literature. The goals of the research reported here were to establish a baseline analysis of current CTD curation, develop a text-mining prototype from readily available open-source components, and evaluate its potential value in augmenting curation efficiency and increasing data coverage. PathText [54] is an integrated environment for combining standards compliant biological pathway models and original publications regarding selected parts of the pathway, through the use of text mining technology and tools, to facilitate the creation of manual annotations. PathText integrates three knowledge sources indispensable

for systems biology: (i) external databases, such as SwissProt, EntreGene, Flybase, HUGO; (ii) text databases such as MEDLINE and full publications; (iii) pathways as organized interpretations of biological facts. PathText successfully provides integration of text to pathways and has been used by three groups that make research on biological topics at the Systems Biology Institute, the University of Tokyo [55], and the Manchester Centre for Integrative Systems Biology in the UK [56]. Karamanis et al. [57] apply natural language processing techniques to develop a generic tool aimed at assisting FlyBase curators. Kiritchenko et al. [58] proposed a tool aimed at retrieving Medline publications that mention genes. After being retrieved, publications are categorized according to the GO codes. The purpose of their work is to retrieve the known functionality of a group of genes from the literature and translate it into a controlled vocabulary. The categorization process can be used for automatic or semiautomatic database curation and maintenance. At the same time, it can be used as a stage in gene expression analysis. After that microarray experiments have been performed and gene expression data have been preprocessed and clustered, the information on gene functions can be added as background knowledge. Literature-Based Discovery (LBD for short) is another relevant research area that applies text-mining with the goal of finding new relationships from knowledge typically available on the Web, in terms of scientific documents, books, and papers. The technique was pioneered by Don R. Swanson in the 1980s and has been widely studied afterwards. It is worth pointing out that LBD techniques do not generate knowledge by means of experiments. Rather, they seek to connect existing knowledge from empirical results by searching and highlighting relationships not yet put into evidence. The pioneering work of Swanson [59] hypothesized the role of fish oil in clinical treatment of Raynaud's disease, combining different pieces of information from the literature, and the hypothesis was later confirmed with experimental evidence. Swanson was using the so-called ABC model of discovery, which asserts that, in the event A and B are related and B and C are related, then A and C might be (indirectly) related. Swanson's ABC model can be implemented in accordance with two different discovery processes: closed and open. The former tries to identify existing links between a hypothesis and the existing literature, whereas the latter generalizes the closed approach by rendering the hypothesis a "free variable" in the discovering task. Hence, a closed discovery process is characterized by the elaboration of a hypothesis, whereas an open discovery process is also concerned with hypothesis generation. LBD has been extensively investigated and applied to many areas of biomedicine, mainly using textual information derived from MedLine (typically in terms of titles, abstracts, and MeSH headings). Among relevant tools and systems proposed and/or experimented in this research field (for a review, see, e.g., [60]), let us recall the work of Hristovski et al. [61]. The authors use semantic predications to enhance cooccurrence-based open discovery systems. Predications are produced by using two natural language processing systems in combination that is, BioMedLEE [62] and SemRep [63], together with the BITOLA system. BITOLA is an open discovery system, compliant with the Swanson's approach, which uses MeSH terms instead of title words and employs association rules instead of word frequencies to relate medical concepts. The authors include also domain-specific knowledge, as they use information in the form of chromosome location and gene expression localization.

Corney et al. [64] propose BioRAT (http://bioinf.cs.ucl.ac.uk/software_downloads/biorat/) (Biological Research Assistant for Text mining), an information extraction tool specifically tailored for biomedical tasks. Able to access and analyze both abstracts and full-length publications, it incorporates a domain specific document search ability. BioRAT uses natural language processing techniques and domain-specific knowledge to search for patterns in documents, with the aim of identifying interesting facts. These facts can then be extracted to produce a database of information, which has a higher *information density* than a pile of publications. PolySearch (http://wishart.biology.ualberta.ca/polysearch/) [65] is a web application aimed at extracting and analyzing text-derived relationships between human diseases, genes, proteins, mutations (SNPs), drugs, metabolites, pathways, tissues, organs, and subcellular localizations. To this end, it analyzes documents and data from several sources, including PubMed, OMIM, DrugBank, SwissProt, HMDB, HGMD, and Entrez SNP. The system has been designed to address queries of the form "*Given a single X, find all Y's*," where X and Y are biomedical terms (e.g., diseases, tissues, cell compartments, and gene/protein names). Metabolic and signaling pathways are an increasingly important part of organizing knowledge in systems biology. They serve to integrate collective interpretations of facts scattered throughout the literature. Biologists construct a pathway by reading a large number of publications and interpret them as a consistent network, but most of the models currently lack direct links to those publications. Biologists who want to check the original publications have to spend substantial amounts of time to collect relevant publications and to identify the sections relevant to the pathway [66]. PathwayAssist [67] is a software application developed to navigate and analyze biological pathways, gene regulation networks, and protein interaction maps. PathwayAssist enables researchers to create their own pathways and produces pathway diagrams. For visualization purposes, pathways are represented as a graph with two types of nodes: those reserved for proteins, small molecules, and cellular processes and those that represent events of functional regulation, chemical reactions, and protein-protein interactions. PathwayAssist comes with a database of molecular networks automatically assembled from scientific abstracts. The database has been populated by using the text-mining tool MedScan on the whole PubMed. MedScan preprocesses input text to extract relevant sentences, which undergo natural language processing. The preprocessing step uses a manually curated dictionary of synonyms to recognize biological terms. Sentences that do not contain at least one matched term are filtered out. The natural language processing kernel deduces the syntactic structure of a sentence and establishes logical relationships between concepts. Finally, results are matched against the functional

ontology to produce the biological interpretation. SciMiner (http://jdrf.neurology.med.umich.edu/SciMiner/) [68] is a web-based literature mining and functional analysis tool aimed at analyzing Medline abstracts and full-text articles to identify genes and proteins. Gene and proteins are extracted and ranked by the number of documents in which they appear. Moreover, they are further analyzed for their enrichments in GO terms, pathways, Medical Subject Heading (MeSH) terms, and protein-protein interaction networks based on external annotation resources. Anni 2.0 (http://biosemantics.org/anni) [69] retrieves documents and associations for several classes of biomedical concepts. It exploits an ontology-based interface to MEDLINE that defines concepts and their relations. Anni finds related concepts based on their associated sets of texts. Peregrine [70] is a concept recognition software, that has been used in Anni to identify references to concepts in text. Texts can be also related to a concept by using manually curated annotation databases. Texts related with a concept are characterized by a concept profile, which consists of a list of concepts used to infer functional associations between genes, between genes and GO codes, to infer novel genes associated with the nucleolus, and to identify new uses for drugs and other substances in the treatment of diseases. FACTA (http://text0.mib.man.ac.uk/software/facta/) [71] is a text search engine aimed at browsing biomedical concepts that are potentially relevant to a query. FACTA differs from other similar tools for its ability to deliver real-time responses and to accept flexible queries.

4. Open Problems and Challenges

As already pointed out, the steady work of researchers has brought a huge increase of publications in life sciences. This amount of scientific literature requires an extra work by researchers, typically involved in keeping up-to-date all information related to their favorite research topics. This effort mainly depends on two aspects: the continuous increase of the scientific production and the poor amount of communication among life science disciplines [72]. In this scenario, devising suitable strategies, techniques, and tools aimed at supporting researchers in the task of automatically retrieving relevant information on the Web (in particular, from text documents), has become an issue of paramount importance.

The research field of literature retrieval and mining in bioinformatics is intrinsically manifold, which makes more complex the task of identifying open problems and challenges. However, in our view, some specific issues deserve the attention of researchers more than others, along the way that leads to significant improvements. Without claiming exhaustiveness, let us briefly point out some of them: (i) encoding/preprocessing techniques; (ii) intrinsic complexity of literature retrieval and mining problems; (iii) standards and requirement for further standardization; (iv) assessment of existing tools.

Encoding/Preprocessing Techniques. Roughly speaking, preprocessing techniques can be divided according to the following dimensions: (i) natural language processing (NLP), (ii) lexical techniques, and (iii) semantic techniques. Currently, NLP does not guarantee to come up with effective solutions able to account for the virtually infinite set of variations concerning the way relevant information is "deployed" in text documents. However, this field may become of primary importance in the next future, due to its great potential. Lexical techniques, focused on finding relevant terms able to characterize documents, are usually simpler to implement, no matter whether they are actually framed in a perspective based on frequencies or information theory. As a matter of fact, they should be considered only a starting point, as preprocessing made using purely lexical techniques (e.g., TFIDF [2]) appears not suitable for typical literature retrieval and mining problems. To some extent, semantic techniques lie in the middle between NLP and lexical techniques. A usual schema adopted while applying semantic techniques is to enrich lexical information with additional knowledge, which can be obtained in several ways. Just to cite few: (i) any given text document can be mapped to an existing taxonomy/ontology, with the goal of identifying relevant concepts and attach them to the document itself, to facilitate further processing; (ii) specific term disambiguation techniques (e.g., latent semantic indexing, synset analysis, or NER analysis) may be applied, with the goal of improving the significance of candidate terms, to be used for representing (together with other terms) a given document; (iii) space transformation techniques (e.g., feature extraction) may be applied, with the goal of limiting the amount of information required for disambiguating text documents. Based on singular value decomposition, Latent Semantic Indexing (LSI) [73] is a technique used to compute document and term similarities according to a "soft" term matching rule. In doing so, terms and documents can be expressed as vectors of statistically independent factors, so that the similarity of any two terms can be better estimated by the cosine of their vector expressions. Synsets have become popular with the advent of WordNet [74], a lexical database for the English language. In Wordnet, English words are grouped into sets of synonyms called synsets, each containing all synonyms related to a specific concept. Named Entity Recognition (NER) [75] is aimed at detecting all instances of a biological term (e.g., protein name, gene name, drug name) in an article or a in a collection.

Intrinsic Complexity of Literature Retrieval and Mining Problems. Beyond the difficulties related to the task of identifying the "right" encoding/preprocessing technique to be adopted, some tasks are in fact inherently complex. For instance, let us consider a generic open discovery processes, framed in the subfield of LBD, which requires to select the hypothesis to be investigated. Even under the assumption that the corresponding task is guided by suitable heuristics aimed at restricting the set of candidate hypotheses, the complexity in time of an open discovery process remains very high and requires specific AI techniques and algorithms. Besides,

at least in principle, complexity issues hold also for closed discovery processes, as they can be framed in the general context of abduction.

Standards and Requirements for Further Standardization. Life sciences are evolving very quickly. To this end, a wide agreement by the scientific community on describing biological concepts is more and more required. On one hand, resolving names, abbreviations, and acronyms is very difficult, due to the fact that different entities could be referenced through the same (or similar) names, abbreviations, and acronyms. On the other hand, it is difficult to detect when a composite name begins and ends in a text. In our opinion, these problems strictly depend on the lack of standard nomenclature and software tools. Fortunately, a good initiative aimed at promoting standardization has been the Unified Medical Language System (UMLS), which brings together many health and biomedical vocabularies and standards—with the goal of enabling interoperability between computer systems. The UMLS has three tools: Metathesaurus (which contains terms and codes from many vocabularies), Semantic Network (able to navigate throughout relevant categories and their relationships), and Specialist (equipped with language processing tools). However, several other problems are still open—due to a lack of standardization. In our opinion, one of the most challenging problems is the need for automatically fusing literature and biological databases. Indeed, the activities of bioinformaticians are unrelated from those of database curators. In this scenario, standard tools able to facilitate the tasks of extracting text and relationships from the literature and to facilitate database curators in the task of identifying relevant literature for annotation would greatly contribute to make the problem less severe or even absent. Other challenging problems are strongly related to the structure of scientific publications. Indeed, although it is quite easy to detect relationships between sentences by analyzing an abstract, the same it is not true while analyzing a full-text publication. This happens because the ability of a software system to detect relationships within a publication is closely related to the structure therein. In particular, each section may be in charge of addressing a specific topic. For instance, the *Introduction* is devoted to describe and analyze the problem; the *Methods* section is aimed at illustrating and explaining the methodological approach, whereas *Results* and *Discussion* are devoted to report experimental results and to discuss whether the initial goals have been achieved. This implies that different concepts (e.g., entity names, experimental conditions, and results) might be located at different sections of a publication [76]. As a consequence, a term could be related to different concepts, depending of the section(s) in which it appears. For example, the name of a gene in the *Introduction* can be related to results published in previous works rather than to novel discoveries presented in the document under analysis. The same might happen for sentences belonging to different sections of the same publication. To solve these problems, recent advances in literature retrieval and mining, together with the increase of open-access journals, are propelling publishers to provide structured version of full-text publications (usually as XML files). We completely agree that the adoption of suitable standards able to represent documents in a structured way would greatly improve the effectiveness of text mining procedures.

Assessment of Existing Tools. Nowadays, the scientific community is strongly concerned with finding how the proposed techniques can provide better access to the existing literature. Some competitions have been organized with the aim of assessing to which extent new approaches allow to navigate and mine the literature. A good example in this direction is given by the critical assessment of text mining methods in molecular biology (BioCreAtIvE) [77, 78]. This competition, which gets together every two years many researchers, is aimed at comparing methods devised to detect: (i) biologically significant entities and their association to existing database entries and (ii) entity-fact associations (e.g., protein-functional term associations). In our view, further initiatives in this direction could promote the sharing of relevant knowledge and skills, while pushing researchers to make a step forward in their specific topics of interest.

5. Conclusions

Research and development in the analysis of bioinformatics literature aims to provide bioinformaticians with effective means to access and exploit the knowledge contained in scientific publications. Although the majority of scientific publications are nowadays electronically available, keeping up to date with recent findings remains a tedious task hampered by the difficulty of accessing the relevant literature. Bioinformatics text analysis aims to improve the access to unstructured knowledge by alleviating searches, providing auto-generated summaries, linking publications with structured resources, visualizing content for better understanding, and supporting researchers in the task of formulating novel hypotheses and of discovering knowledge. Research over recent years has improved fundamental methods in bioinformatics text mining, ranging from document retrieval to the extraction of relationships. Consequently, more and more integrative literature analysis tools have been put forward, targeting a broad audience of life scientists. In this paper, after briefly introducing information retrieval, text mining, and literature retrieval and mining, we first recalled the state of the art on literature retrieval and mining in bioinformatics. In the second part of the paper, we discussed some challenges deemed worth of further investigation, with the goal of improving bioinformatics literature-retrieval-and-mining tools and systems. Summarizing, the scientific community is strongly involved in addressing different problems in literature retrieving and mining, and several solutions have been currently proposed and adopted. Nevertheless, they will remain largely ineffective until the scientific community will make further significant steps towards common standards concerning the way existing

knowledge is published and shared among researchers—with particular emphasis on the structure of the scientific publications.

Acknowledgments

This work has been supported by the Italian Ministry Education and Research through the Flagship "InterOmics," ITALBIONET (RBPR05ZK2Z), Bioinformatics analysis applied to Populations Genetics (RBIN064YAT 003), and the European "SHIWA" projects.

References

[1] G. Armano, M. de Gemmis, G. Semeraro, and E. Vargiu, *Intelligent Information Access*, vol. SCI 301 of *Studies in Computational Intelligence*, Springer, Heidelberg, Germany, 2010.

[2] R. A. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*, Addison-Wesley Longman, Boston, Mass, USA, 1999.

[3] M. Kobayashi and K. Takeda, "Information retrieval on the web," *ACM Computing Surveys*, vol. 32, no. 2, pp. 165–173, 2000.

[4] S. Bao, G. Xue, X. Wu, Y. Yu, B. Fei, and Z. Su, "Optimizing web search using social annotations," in *16th International World Wide Web Conference (WWW '07)*, pp. 501–510, New York, NY, USA, May 2007.

[5] C. D. Manning, P. Raghavan, and H. Schtze, *Introduction to Information Retrieval*, Cambridge University Press, New York, NY, USA, 2008.

[6] J. Mayfield and T. Finin, "Information retrieval on the semantic web: Integrating inference and retrieval," in *Proceedings of the SIGIR Workshop on the Semantic Web*, August 2003.

[7] M. W. Berry, *Survey of Text Mining*, Springer, New York, NY, USA, 2003.

[8] F. Sebastiani, "Machine learning in automated text categorization," *ACM Computing Surveys*, vol. 34, no. 1, pp. 1–47, 2002.

[9] I. Mani, *Automatic Summarization*, John Benjamins, Amsterdam, The Netherlands, 2001.

[10] H. Luhn, "The automatic creation of literature abstracts," *IBM Journal of Research and Development*, vol. 2, pp. 159–165, 1958.

[11] P. Baxendale, "Machine-made index for technical literature—an experiment," *IBM Journal of Research and Development*, vol. 2, pp. 354–361, 1958.

[12] H. P. Edmundson, "New methods in automatic extracting," *Journal of ACM*, vol. 16, pp. 264–285, 1969.

[13] A. Nenkova, "Automatic text summarization of newswire: lessons learned from the document understanding conference," in *Proceedings of the 20th National Conference on Artificial Intelligence*, vol. 3, pp. 1436–1441, AAAI Press, 2005.

[14] G. Armano, A. Giuliani, and E. Vargiu, "Studying the impact of text summarization on contextual advertising," in *Proceedings of the 8th International Workshop on Text-based Information Retrieval*, 2011.

[15] A. Kołcz, V. Prabakarmurthi, and J. Kalita, "Summarization as feature selection for text categorization," in *Proceedings of the 10th ACM International Conference on Information and Knowledge Management (CIKM '01)*, pp. 365–370, New York, NY, USA, November 2001.

[16] J. Martin, "Clustering full text documents," in *Proceedings of the Workshop on Data Engineering for Inductive Learning at (IJCAI '95)*, 1995.

[17] P. Willett, "Recent trends in hierarchic document clustering: a critical review," *Information Processing and Management*, vol. 24, no. 5, pp. 577–597, 1988.

[18] C. Aone, S. W. Bennett, and J. Gorlinsky, "Multi-media fusion through application of machine learning and nlp," in *AAAI Spring Symposium Working Notes on Machine Learning in Information Access*, 1996.

[19] D. H. Fisher, "Knowledge acquisition via incremental conceptual clustering," *Machine Learning*, vol. 2, no. 2, pp. 139–172, 1987.

[20] R. Liere and P. Tadepalli, "Active learning with committees for text categorization," in *Proceedings of the 14th National Conference on Artificial Intelligence (AAAI '97)*, pp. 591–596, July 1997.

[21] P. Cheeseman, J. Kelly, M. Self, J. Stutz, W. Taylor, and D. Freeman, *Readings in Knowledge Acquisition and Learning, chap. AutoClass: A Bayesian Classification System*, Morgan Kaufmann, San Francisco, Calif, USA, 1993.

[22] C. Green and P. Edwards, "Using machine learning to enhance software tools for internet information management," in *Proceedings of the AAAI Workshop on Internetbased Information Systems*, pp. 48–55, 1996.

[23] D. E. Appelt, "Introduction to information extraction," *AI Communications*, vol. 12, no. 3, pp. 161–172, 1999.

[24] S. Soderland, D. Fisher, J. Aseltine, and W. Lehnert, "Crystal inducing a conceptual dictionary," in *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, vol. 2, pp. 1314–1319, Morgan Kaufmann, San Francisco, Calif, USA, 1995.

[25] S. B. Huffman, "Learning information extraction patterns from examples," in *Connectionist, Statistical, and Symbolic Approaches to Learning for Natural Language Processing*, pp. 246–260, Springer, London, UK, 1996.

[26] M. E. Califf and R. J. Mooney, "Relational learning of pattern-match rules for information extraction," in *Proceedings of the 16th National Conference on Artificial Intelligence (AAAI '99), 11th Innovative Applications of Artificial Intelligence Conference (IAAI '99)*, pp. 328–334, July 1999.

[27] D. Freitag, "Machine learning for information extraction in informal domains," *Machine Learning*, vol. 39, pp. 169–202, 2000.

[28] K. D. Bollacker, S. Lawrence, and C. L. Giles, "Discovering relevant scientific literature on the Web," *IEEE Intelligent Systems and Their Applications*, vol. 15, no. 2, pp. 42–47, 2000.

[29] S. M. McNee, I. Albert, D. Cosley et al., "On the recommending of citations for research papers," in *Proceedings of the 8th Conference on Computer Supported Cooperative Work (CSCW '02)*, pp. 116–125, New York, NY, USA, November 2002.

[30] W. C. Janssen and K. Popat, "UpLib: a universal personal digital library system," in *Proceedings of the 2003 ACM Symposium on Document Engineering*, pp. 234–242, fra, November 2003.

[31] F. Mahdavi, M. A. Ismail, and N. Abdullah, "Semi-automatic trend detection in scholarly repository using semantic approach," in *Proceedings of the World Academy of Science, Engineering and Technology*, pp. 224–226, Amsterdam, The Netherlands, 2009.

[32] N. Cannata, E. Merelli, and R. B. Altman, "Erratum: time to organize the bioinformatics resourceome," *PLoS Computational Biology*, vol. 2, no. 2, p. 112, 2006.

[33] A. K. Bajpai, S. Davuluri, H. Haridas et al., "In search of the right literature search engine(s)," *Nature Preceding*, 2011.

[34] H. Yu, T. Kim, J. Oh, I. Ko, S. Kim, and W. S. Han, "Enabling multi-level relevance feedback on PubMed by integrating

rank learning into DBMS," *BMC Bioinformatics*, vol. 11, supplement 2, p. S6, 2010.

[35] J. F. Fontaine, A. Barbosa-Silva, M. Schaefer, M. R. Huska, E. M. Muro, and M. A. Andrade-Navarro, "MedlineRanker: flexible ranking of biomedical literature," *Nucleic Acids Research*, vol. 37, no. 2, pp. W141–W146, 2009.

[36] J. Wang, I. Cetindil, S. Ji et al., "Interactive and fuzzy search: a dynamic way to explore MEDLINE," *Bioinformatics*, vol. 26, no. 18, Article ID btq414, pp. 2321–2327, 2010.

[37] R. Herbrich, T. Graepel, and K. Obermayer, "Large margin rank boundaries for ordinal regression," in *Advances in Large Margin Classifiers*, Smola B. and Schoelkopf S., Eds., MIT Press, Cambridge, Mass, USA, 2000.

[38] J. Lewis, S. Ossowski, J. Hicks, M. Errami, and H. R. Garner, "Text similarity: an alternative way to search MEDLINE," *Bioinformatics*, vol. 22, no. 18, pp. 2298–2304, 2006.

[39] P. Coppernoll-Blach, "Quertle: the conceptual relationships alternative search engine for pubmed," *Journal of Medical Library Association*, vol. 99, no. 2, pp. 176–177, 2011.

[40] A. Doms and M. Schroeder, "GoPubMed: exploring PubMed with the gene ontology," *Nucleic Acids Research*, vol. 33, no. 2, pp. W783–W786, 2005.

[41] C. Perez-Iratxeta, A. J. Pérez, P. Bork, and M. A. Andrade, "Update on XplorMed: a web server for exploring scientific literature," *Nucleic Acids Research*, vol. 31, no. 13, pp. 3866–3868, 2003.

[42] D. Rebholz-Schuhmann, H. Kirsch, M. Arregui, S. Gaudan, M. Riethoven, and P. Stoehr, "EBIMed—text crunching to gather facts for proteins from Medline," *Bioinformatics*, vol. 23, no. 2, pp. e237–e244, 2007.

[43] R. Hoffmann and A. Valencia, "A gene network for navigating the literature," *Nature Genetics*, vol. 36, no. 7, p. 664, 2004.

[44] L. Tanabe, U. Scherf, L. H. Smith, J. K. Lee, L. Hunter, and J. N. Weinstein, "MedMiner: an internet text-mining tool for biomedical information, with application to gene expression profiling," *BioTechniques*, vol. 27, no. 6, pp. 1210–1217, 1999.

[45] T. Greenhalgh, "How to read a paper. The medline database," *BMJ*, vol. 315, no. 7101, pp. 180–183, 1997.

[46] A. S. Yeh, L. Hirschman, and A. A. Morgan, "Evaluation of text data mining for database curation: lessons learned from the KDD Challenge Cup," *Bioinformatics*, vol. 19, pp. i331–339, 2003.

[47] J. Thomas, D. Milward, C. Ouzounis, S. Pulman, and M. Carroll, "Automatic extraction of protein interactions from scientific abstracts," *Pacific Symposium on Biocomputing*, pp. 541–552, 2000.

[48] M. He, Y. Wang, and W. Li, "PPI finder: a mining tool for human protein-protein interactions," *PLoS ONE*, vol. 4, no. 2, Article ID e4554, 2009.

[49] M. Berardi, D. Malerba, R. Piredda, M. Attimonelli, G. Scioscia, and P. Leo, *16 Biomedical Literature Mining for Biological Databases Annotation*, 2008.

[50] M. Craven and J. Kumlien, "Constructing biological knowledge bases by extracting information from text sources," in *Proceedings of the 7th International Conference on Intelligent Systems for Molecular Biology*, pp. 77–86, 1999.

[51] D. R. Swanson and N. R. Smalheiser, "An interactive system for finding complementary literatures: a stimulus to scientific discovery," *Artificial Intelligence*, vol. 91, no. 2, pp. 183–203, 1997.

[52] I. Donaldson, J. Martin, B. de Bruijn et al., "PreBIND and Textomy—mining the biomedical literature for protein-protein interactions using a support vector machine," *BMC Bioinformatics*, vol. 4, no. 1, p. 11, 2003.

[53] T. C. Wiegers, A. P. Davis, K. B. Cohen, L. Hirschman, and C. J. Mattingly, "Text mining and manual curation of chemical-gene-disease networks for the comparative toxicogenomics Database (CTD)," *BMC Bioinformatics*, vol. 10, article 326, 2009.

[54] B. Kemper, T. Matsuzaki, Y. Matsuoka et al., "PathText: a text mining integrator for biological pathway visualizations," *Bioinformatics*, vol. 26, no. 12, Article ID btq221, pp. i374–i381, 2010.

[55] K. Oda, J. D. Kim, T. Ohta et al., "New challenges for text mining: mapping between text and manually curated pathways," *BMC Bioinformatics*, vol. 9, supplement 3, p. S5, 2008.

[56] M. J. Herrgård, N. Swainston, P. Dobson et al., "A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology," *Nature Biotechnology*, vol. 26, no. 10, pp. 1155–1160, 2008.

[57] I. Karamanis, R. Lewi, R. D. Seal, and B. E, "Integrating natural language processing with flybase curation," in *Proceedings of the Pacific Symposium on Biocomputing*, pp. 245–256, Maui, Hawaii, USA, 2007.

[58] S. Kiritchenko, S. Matwin, and A. F. Famili, "Hierarchical text categorization as a tool of associating genes with gene ontology codes," in *Proceedings of the 2nd European Workshop on Data Mining and Text Mining for Bioinformatics*, pp. 26–30, 2004.

[59] D. R. Swanson, "Fish oil, Raynaud's syndrome, and undiscovered public knowledge," *Perspectives in Biology and Medicine*, vol. 30, no. 1, pp. 7–18, 1986.

[60] P. Bruza and M. Weeber, *Literature-based Discovery*, vol. 15, Springer, Heidelberg, Germany, 2008.

[61] D. Hristovski, B. Peterlin, J. A. Mitchell, and S. M. Humphrey, "Using literature-based discovery to identify disease candidate genes," *International Journal of Medical Informatics*, vol. 74, no. 2–4, pp. 289–298, 2005.

[62] L. Chen and C. Friedman, "Extracting phenotypic information from the literature via natural language processing," *Medinfo*, vol. 11, no. 2, pp. 758–762, 2004.

[63] P. Srinivasan and T. Rindflesch, "Exploring text mining from MEDLINE," *Proceedings of the AMIA Symposium*, pp. 722–726, 2002.

[64] D. P. A. Corney, B. F. Buxton, W. B. Langdon, and D. T. Jones, "BioRAT: extracting biological information from full-length papers," *Bioinformatics*, vol. 20, no. 17, pp. 3206–3213, 2004.

[65] D. Cheng, C. Knox, N. Young, P. Stothard, S. Damaraju, and D. S. Wishart, "PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites," *Nucleic Acids Research*, vol. 36, pp. W399–W405, 2008.

[66] S. Ananiadou, D. B. Kell, and J. I. Tsujii, "Text mining and its potential applications in systems biology," *Trends in Biotechnology*, vol. 24, no. 12, pp. 571–579, 2006.

[67] A. Nikitin, S. Egorov, N. Daraselia, and I. Mazo, "Pathway studio—the analysis and navigation of molecular networks," *Bioinformatics*, vol. 19, no. 16, pp. 2155–2157, 2003.

[68] J. Hur, A. D. Schuyler, D. J. States, and E. L. Feldman, "SciMiner: web-based literature mining tool for target identification and functional enrichment analysis," *Bioinformatics*, vol. 25, no. 6, pp. 838–840, 2009.

[69] R. Jelier, M. J. Schuemie, A. Veldhoven, L. C. J. Dorssers, G. Jenster, and J. A. Kors, "Anni 2.0: a multipurpose text-mining tool for the life sciences," *Genome Biology*, vol. 9, no. 6, article R96, 2008.

[70] M. Schuemie, R. Jelier, and J. K. J, "Peregrine: lightweight gene name normalization by dictionary lookup," in *Proceedings of*

the 2nd BioCreative Challenge Evaluation Workshop, pp. 131–140, 2007.

[71] Y. Tsuruoka, J. Tsujii, and S. Ananiadou, "Facta: a text search engine for finding associated biomedical concepts," *Bioinformatics*, vol. 24, no. 21, pp. 2559–2560, 2008.

[72] M. Weeber, H. Klein, A. R. Aronson, J. G. Mork, L. T. de Jong-van den Berg, and R. Vos, "Msc: Text-based discovery in biomedicine: the architecture of the DAD-system," *Proceedings of the AMIA, the Annual Conference of the American Medical Informatics Association*, pp. 903–907, 2000.

[73] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman, "Indexing by latent semantic analysis," *JASIS*, vol. 41, no. 6, pp. 391–407, 1990.

[74] G. A. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[75] M. Krauthammer and G. Nenadic, "Term identification in the biomedical literature," *Journal of Biomedical Informatics*, vol. 37, no. 6, pp. 512–526, 2004.

[76] P. K. Shah, C. Perez-Iratxeta, P. Bork, and M. A. Andrade, "Information extraction from full text scientific articles: where are the keywords?" *BMC Bioinformatics*, vol. 4, article no. 20, 2003.

[77] L. Hirschman, A. Yeh, C. Blaschke, and A. Valencia, "Overview of BioCreAtIvE: critical assessment of information extraction for biology," *BMC Bioinformatics*, vol. 6, no. 1, article S1, 2005.

[78] M. Krallinger, A. Morgan, L. Smith et al., "Evaluation of text-mining systems for biology: overview of the second biocreative community challenge," *Genome Biology*, vol. 9, no. 2, article no. S1, 2008.

Hindawi Publishing Corporation
Advances in Bioinformatics
Volume 2012, Article ID 582765, 12 pages
doi:10.1155/2012/582765

Research Article

Exploring Biomolecular Literature with EVEX: Connecting Genes through Events, Homology, and Indirect Associations

Sofie Van Landeghem,[1,2] Kai Hakala,[3] Samuel Rönnqvist,[3] Tapio Salakoski,[3,4] Yves Van de Peer,[1,2] and Filip Ginter[3]

[1] *Department of Plant Systems Biology, VIB, Technologiepark 927, 9052 Gent, Belgium*
[2] *Department of Plant Biotechnology and Bioinformatics, Ghent University, Technologiepark 927, 9052 Gent, Belgium*
[3] *Department of Information Technology, University of Turku, Joukahaisenkatu 3-5, 20520 Turku, Finland*
[4] *Turku BioNLP Group, Turku Centre for Computer Science (TUCS), Joukahaisenkatu 3-5, 20520 Turku, Finland*

Correspondence should be addressed to Filip Ginter, ginter@cs.utu.fi

Received 22 November 2011; Revised 16 March 2012; Accepted 28 March 2012

Academic Editor: Jin-Dong Kim

Technological advancements in the field of genetics have led not only to an abundance of experimental data, but also caused an exponential increase of the number of published biomolecular studies. Text mining is widely accepted as a promising technique to help researchers in the life sciences deal with the amount of available literature. This paper presents a freely available web application built on top of 21.3 million detailed biomolecular events extracted from all PubMed abstracts. These text mining results were generated by a state-of-the-art event extraction system and enriched with gene family associations and abstract generalizations, accounting for lexical variants and synonymy. The EVEX resource locates relevant literature on phosphorylation, regulation targets, binding partners, and several other biomolecular events and assigns confidence values to these events. The search function accepts official gene/protein symbols as well as common names from all species. Finally, the web application is a powerful tool for generating homology-based hypotheses as well as novel, indirect associations between genes and proteins such as coregulators.

1. Introduction

The field of natural language processing for biomolecular texts (BioNLP) aims at large-scale text mining in support of life science research. Its primary motivation is the enormous amount of available scientific literature, which makes it essentially impossible to rapidly gain an overview of prior research results other than in a very narrow domain of interest. Among the typical use cases for BioNLP applications are support for database curation, linking experimental data with relevant literature, content visualization, and hypothesis generation—all of these tasks require processing and summarizing large amounts of individual research articles. Among the most heavily studied tasks in BioNLP is the extraction of information about known associations between biomolecular entities, primarily genes, and gene products, and this task has recently seen much progress in two general directions.

First, relationships between biomolecular entities are now being extracted in much greater detail. Until recently, the focus was on extracting untyped and undirected binary relations which, while stating that there is *some* relationship between two objects, gave little additional information about the nature of the relationship. Recognizing that extracting such relations may not provide sufficient detail for wider adoption of text mining in the biomedical community, the focus is currently shifting towards a more detailed analysis of the text, providing additional vital information about the detected relationships. Such information includes the type of the relationship, the specific roles of the arguments (e.g., affector or affectee), the polarity of the relationship (positive versus negative statement), and whether it was stated in a speculative or affirmative context. This more detailed text mining target was formalized as an *event extraction* task and greatly popularized in the series of BioNLP Shared Tasks on Event Extraction [1, 2]. These shared tasks mark

a truly community-wide effort to develop efficient systems to extract sufficiently detailed information for real-world, practical applications, with the highest possible accuracy.

Second, text mining systems are now being applied on a large scale, recognizing the fact that, in order for a text mining service to be adopted by its target audience, that is, researchers in the life sciences, it must cover as much of the available literature as possible. While small-scale studies on well-defined and carefully constructed corpora comprising several hundred abstracts are of great utility to BioNLP research, actual applications of the resulting methods require the processing of considerably larger volumes of text, ideally including all available literature. Numerous studies have been published demonstrating that even complex and computationally intensive methods can be successfully applied on a large scale, typically processing all available abstracts in PubMed and/or all full-text articles in the open-access section of PubMed Central. For instance, the *iHOP* [3] and *Medie* [4] systems allow users to directly mine literature relevant to given genes or proteins of interest, allowing for structured queries far beyond the usual keyword search. *EBIMed* [5] offers a broad scope by also including gene ontology terms such as biological processes, as well as drugs and species names. Other systems, such as the *BioText search engine* [6] and *Yale Image Finder* [7] allow for a comprehensive search in full-text articles, including also figures and tables. Finally, the *BioNOT* system [8] focuses specifically on extracting negative evidence from scientific articles.

The first large-scale application that specifically targets the extraction of detailed events according to their definition in the BioNLP Shared Tasks is the dataset of Björne et al. [9], comprising 19 million events among 36 million gene and protein mentions. This data was obtained by processing all 18 million titles and abstracts in the 2009 PubMed distribution using the winning system of the BioNLP'09 Shared Task. In a subsequent study of Van Landeghem et al. [10], the dataset was refined, generalized, and released as a relational (SQL) database referred to as *EVEX*. Among the main contributions of this subsequent study was the generalization of the events, using publicly available gene family definitions. Although a major step forward from the original text-bound events produced by the event extraction system, the main audience for the EVEX database was still the BioNLP community. Consequently, the dataset is not easily accessible for researchers in the life sciences who are not familiar with the intricacies of the event representation. Further, as the massive relational database contains millions of events, manual querying is not an acceptable way to access the data for daily use in life science research.

In this study, we introduce a publicly available web application based on the EVEX dataset, presenting the first application that brings large-scale event-based text mining results to a broad audience of end-users including biologists, geneticists, and other researchers in the life sciences. The web application is available at http://www.evexdb.org/. The primary purpose of the application is to provide the EVEX dataset with an intuitive interface that does not presuppose familiarity with the underlying event representation. The application presents a comprehensive and thoroughly interlinked overview of all events for a given gene or protein, or a gene/protein pair. The main novel feature of this application, as compared to other available large-scale text mining applications, is that it covers highly detailed event structures that are enriched with homology-based information and additionally extracts indirect associations by applying cross-document aggregation and combination of events.

In the following section, we provide more details on the EVEX text mining dataset, its text-bound extraction results, and the gene family-based generalizations. Further, we present several novel algorithms for event ranking, event refinement, and retrieval of indirect associations. Section 3 presents an evaluation of the EVEX dataset and the described algorithms. The features of the web application are illustrated in Section 4, presenting a real-world use case on the budding yeast gene *Mec1*, which has known mammalian and plant homologs. We conclude by summarizing the main contributions of this work and highlighting several interesting opportunities for future work.

2. Data and Methods

This section describes the original event data, as well as a ranking procedure that sorts events according to their reliability. Further, two abstract layers are defined on top of the complex event structures, enabling coarse grouping of similar events, and providing an intuitive pairwise point of view that allows fast retrieval of interesting gene/protein pairs. Finally, we describe a hypothesis generation module that finds missing links between two entities, allowing the user to retrieve proteins with common binding partners or genes that act as coregulators of a group of common target genes.

2.1. EVEX Dataset

2.1.1. Core Events. The core set of text mining results accessible through the EVEX resource has been generated by the Turku Event Extraction System, the winning system of the BioNLP'09 Shared Task (ST) on Event Extraction [1]. This extraction system was combined with the BANNER named entity recognizer [11], forming a complete event extraction pipeline that had the highest reported accuracy on the task in 2009, and still remains state-of-the-art, as shown in the recent ST'11 [12]. This event extraction pipeline was applied to all citations in the 2009 distribution of PubMed [9]. As part of the current study, citations from the period 2009–2011 have been processed, using essentially the same pipeline with several minor improvements, resulting in 40.3 million tagged gene symbols and 21.3 million extracted events. The underlying event dataset has thus been brought up to date and will be regularly updated in the future.

The dataset contains events as defined in the context of the ST'09, that is, predicates with a variable number of arguments which can be gene/protein symbols or, recursively, other events. Each argument is defined as having the role of *Cause* or *Theme* in the event. There are nine distinct event types: binding, phosphorylation, regulation

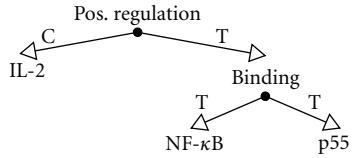


Figure 1: Event representation of the statement *IL-2 acts by enhancing binding activity of NF-κB to p55*, illustrating recursive nesting of events where the (T)heme of the *positive regulation* event is the *binding* event. The (C)ause argument is the gene symbol *IL-2* (figure adapted from [10]).

(positive, negative, and unspecified), protein catabolism, transcription, localization, and gene expression. Further, each event refers to a specific *trigger word* in text. For example, the word *increases* typically triggers a positive regulation event and *degradation* typically refers to protein catabolism. An example event structure is illustrated in Figure 1.

Event definitions impose several restrictions on event arguments: (1) events of the type phosphorylation, protein catabolism, transcription, localization, and gene expression must only have a single argument, a Theme, which must be a gene or a protein, (2) events of the binding type may have any number of gene/protein Theme arguments and cannot have a Cause argument, and finally (3) regulation events must have exactly one Theme argument and may have one Cause argument, with no restrictions as to whether these arguments are genes/proteins or recursively other events. In the following text, we will state events using a simple bracketed notation, where the event type is stated first, followed by a comma-separated list of arguments enclosed in parentheses. For instance, the event in Figure 1 would be stated as *Positive-Regulation(C:IL-2, T:Binding(T:NF-κB, T:p55))*, where *C:* and *T:* denote the role of the argument as (C)ause or (T)heme. For brevity, we will further refer to all biochemical entities, even proteins and mRNA, as *genes*.

2.1.2. Event Generalizations. One of the major limitations of the original core set of events is that they are strictly text-bound and provide no facility for a more general treatment, such as being able to abstract from different name spelling variants and symbol synonymy. Further, biochemical entities were originally treated as merely text strings with no database identity referring to external resources such as UniProt [13] or Entrez Gene [14]. The EVEX dataset addresses these issues by providing event generalizations [10].

First, the identified gene symbols in the EVEX dataset are canonicalized by removing superfluous affixes (prefixes and suffixes) to obtain the core gene symbol, followed by discarding nonalphanumeric characters and lowercasing. For instance, the full string *human Esr-1 subunit* is canonicalized into *esr1*. The purpose of this canonicalization is to abstract away from minor spelling variants and to deal with the fact that the BANNER named entity recognizer often includes a wider context around the core gene symbol. The canonicalization algorithm itself cannot, however, deal with the ambiguity prevalent among the symbols. EVEX thus further resolves these canonical gene symbols, whenever possible, into their most likely families, using two distinct resources for defining homologous genes and gene families: *HomoloGene* (eukaryots, [14]) and *Ensembl* (vertebrates, [15]). As part of this study, we extended EVEX to also include families from *Ensembl Genomes*, which provides coverage for metazoa, plants, protists, fungi, and bacteria [16]. Building on top of these definitions, the EVEX dataset now defines four *event generalizations*, whereby all events whose arguments have the same canonical form, or resolve to the same gene family, are aggregated. As a result, it becomes straightforward to retrieve all information on a specific gene symbol, abstracting away from lexical variants through the canonicalization algorithm, or to additionally apply the synonym-expansion module through the family-based generalizations. These different generalizations are all implemented on the web application (Section 4.2).

2.2. Event Ranking. To rank the extracted events according to their reliability, we have implemented an event scoring algorithm based on the output of the Turku Event Extraction System. This machine learning system uses linear Support Vector Machines (SVMs) as the underlying classifier [17]. Every classification is given a confidence score, the distance to the decision hyperplane of the linear classifier, where higher scores are associated with more confident decisions. There is not a single master classifier to predict the events in their entirety. Rather, individual classifications are made to predict the event trigger and each of its arguments. In order to assign a single confidence score to a specific event occurrence, the predictions from these two separate classifiers must be aggregated.

The confidence scores of the two different classifiers are not directly mutually comparable, and we therefore first normalize all scores in the dataset to zero mean and unit standard deviation, separately for triggers and arguments. Subsequently, the score of a specific event occurrence is assigned to be the *minimum* of the normalized scores of its event trigger and its arguments, that is, the lowest normalized confidence among all classification decisions involved in extracting that specific event. Using minimum as the aggregation function roughly corresponds to the *fuzzy and* operator in that it requires all components of an event to be confident for it to be ranked high. Finally, the score of a generalized event is the average of the scores of all its occurrences.

To assign a meaningful interpretation to the normalized and aggregated confidence values, events within the top 20% of the confidence range are classified as "very high confidence." The other 4 categories, each representing the next 20% of all events, are respectively labeled as "high confidence," "average confidence," "low confidence" and "very low confidence." When presenting multiple possible hits for a certain query, the web application uses the original scores to rank the events from high to low reliability.

2.3. Event Refinement. The extraction of event structures is highly dependent on the lexical and syntactic constructs

used in the sentence and may therefore contain unnecessary complexity. This is because the event extraction system is trained to closely follow the actual statements in the sentence and thus, for instance, will mark both of the words *increase* and *induces* as triggers for positive regulation events in the sentence *Ang II induces a rapid increase in MAPK activity*. Consequently, the final event structure is extracted as *Positive-Regulation(C: Ang II, T: Positive-Regulation(T: MAPK))*, that is, *Ang II* is a Cause argument of a positive regulation event, which has another positive regulation event as its Theme.

While correctly extracted, such nested single-argument regulatory events (i.e., regulations with a Theme but no Cause argument), often forming chains that are several events long, are unnecessarily complex. Clearly, the event above can be restated as *Positive-Regulation(C: Ang II, T: MAPK)*, removing the nested single-argument positive regulation event. This refinement helps to establish the event as equivalent with all other events that can be refined to the same elementary structure, enhancing the event aggregation possibilities in EVEX. However, when presenting the details of the extracted event to the user, the original structure of the event is preserved.

Table 1 lists the set of refinement rules. In this context, positive and negative regulation refer to having a general positive or negative effect, while an unspecified regulation could not be resolved to either category due to missing information in the sentence.

To simplify the single-argument regulatory events, we proceed iteratively, removing intermediary single-argument regulatory events as long as any rule matches. A particular consideration is given to the polarity of the regulations. While a nested chain of single-argument positive regulations can be safely reduced to a single positive regulation, the outcome of reducing chains of single-argument regulations of mixed polarity is less obvious. As illustrated in Table 1, application of the rules may result in a change of polarity of the outer event. For instance, a regulation of a negative regulation is interpreted as a negative regulation, changing the polarity of the outer event from unspecified to negative. To avoid excessive inferences not licensed by the text, the algorithm only allows one such change of polarity. Any subsequent removal of a nested single-argument regulatory event that results in a type change forces the new type of the outer event to be of the unspecified regulation type.

2.4. Pairwise Abstraction. The most basic query issued on the EVEX web application involves a single gene, which triggers the generation of a structured overview page, listing associated genes grouped by their type of connection with the query gene (Section 4.1). The most important underlying functionality implemented by the web application is thus the ability to identify and categorize pairs of related genes. This pairwise point of view comes natural in the life sciences and can be implemented on top of the events with ease by analyzing common event structures and defining argument pairs within. The refinements discussed in Section 2.3 substantially decrease the number of unique event structures present in the data, restricting the required analysis to a comparatively small number of event structures. Furthermore, we only need to consider those events that involve more than one gene or that are a recursive argument in such an event, limiting the set of event occurrences from 21 M to 12 M events.

As an example, let us consider the event *Positive-Regulation(C:Thrombin, T:Positive-Regulatio(C:EGF, Phosphorylation(T:Akt)))*, extracted from the sentence *Thrombin augmented EGF-stimulated Akt phosphorylation*. The pairs of interest here are *Thrombin—Akt* and *EGF—Akt*, both associations coarsely categorized as *regulation*. Therefore, whenever a user queries for *Thrombin*, the *Akt* gene will be listed among the regulation targets, and, whenever a user queries for *Akt*, both *Thrombin* and *EGF* will be listed as regulators. Note, however, that the categorization of the association as *regulation* is only for the purpose of coarse grouping of the results on the overview page. The user will additionally be presented with the details of the original event, which is translated from the bracketed notation into the English statement *Upregulation of AKT phosphorylation by EGF is upregulated by Thrombin*.

There is a limited number of prevalent event structures which account for the vast majority of event occurrences. Table 2 lists the most common structures, together with the gene pairs extracted from them. The algorithm to extract the gene pairs from the event structures proceeds as follows.

(1) All argument pairs are considered a candidate and classified as *binding* if both participants are a Theme of one specific binding event, and *regulation* otherwise. (Note that due to the restrictions of event arguments as described in Section 2.1, only binding and regulation events can have more than one argument.)

(2) If one of the genes is a Theme argument of an event which itself is a Cause argument, for example, *G2* in *Regulation(C:Regulation(C:G1, T:G2), T:G3)*, the association type of the candidate pair *G2-G3* is reclassified as *indirect regulation*, since the direct regulator of *G3* is the Cause argument of the nested regulation (*G1*).

(3) If one of the genes is a Cause argument of an event which itself is a Theme argument, for example, *G2* in *Regulation(C:G1, T:Regulation(C:G2, T:G3))*, the candidate pair (*G1-G2*) is discarded.

While the association between *G1* and *G2* is discarded in step (3) since it in many cases cannot convincingly be classified as a regulation, it is represented as a *coregulation* when indirect associations, described in the following section, are sought.

2.5. Indirect Associations. A cell's activity is often organized into regulatory modules, that is, sets of coregulated genes that share a common function. Such modules can be found by automated analysis and clustering of genome-wide expression profiles [18]. Individual events, as defined by the BioNLP Shared Tasks, do not explicitly express

Table 1: Listing of the refinement rules, involving any nested combination of the three types of regulation: positive regulation (Pos), negative regulation (Neg) and unspecified regulation (Reg). Each parent event has a regulatory (T)heme argument and an optional (C)ause. The nested regulations are all regulations without a Cause and their detailed structure is omitted for brevity. In full, the first structure would read *Pos(C:geneA, T:Pos(T:geneB))* which is rewritten to *Pos(C:geneA, T:geneB)* with *geneA* and *geneB* being any two genes.

Original	Result	Example
Pos(C, T:Pos)	*Pos(C, T)*	BRs induce accumulation of BZR1 protein
Pos(C, T:Reg)	*Pos(C, T)*	PKS5 mediates PM H +- ATPase regulation
Reg(C, T:Pos)	*Pos(C, T)*	CaM regulates activation of HSFs
Neg(C, T:Neg)	*Pos(C, T)*	E2 prevented downregulation of p21
Reg(C, T:Reg)	*Reg(C, T)*	PDK1 is involved in the regulation of S6K
Neg(C, T:Reg)	*Neg(C, T)*	GW5074 prevents this effect on ENT1 mRNA
Neg(C, T:Pos)	*Neg(C, T)*	BIN2 negatively regulates BZR1 accumulation
Reg(C, T:Neg)	*Neg(C, T)*	The effect of hCG in downregulating ER beta
Pos(C, T:Neg)	*Neg(C, T)*	DtRE is required for repression of CAB2

Table 2: The most prevalent (refined) event patterns in the EVEX data, considering only events with more than one gene or protein symbol, and their recursively nested events. These aggregated patterns refer to any type of regulation (**Reg*), to binding events between two genes (*Bind*), and to any physical event (*Phy*) concerning a single gene such as protein-DNA binding, protein catabolism, transcription, localization, phosphorylation, and gene expression. The first two columns refer to the percentage of event occurrences covered by the given pattern and the cumulative percentage of event occurrences up to and including the pattern. The right-most column depicts the extracted gene pair and a coarse classification of its association type. *A* and *B* refer to gene symbols, and bindings are represented with $\times$. Further, $A > B$ means *A regulates B,* while $A \gg B$ expresses an indirect regulation.

Occ. [%]	Cum. occ. [%]	Event pattern	Gene pair
58.6	58.6	*Phy(T:A)*	—
15.0	73.6	**Reg(T:A)*	—
8.4	82.0	**Reg(T:Phy(T:A))*	—
8.0	90.0	*Bind(T:A, T:B)*	$A \times B$
4.7	94.7	**Reg(C:A, T:B)*	$A > B$
3.8	98.5	**Reg(C:A, T:Phy(T:B))*	$A > B$
0.2	98.7	**Reg(C:*Reg(T:Phy(T:A)), T:Phy(T:B))*	$A \gg B$
0.2	98.9	**Reg(C:Phy(T:A), T:B)*	$A \gg B$
0.2	99.1	**Reg(C:Phy(T:A), T:Phy(T:B))*	$A \gg B$

such associations. However, indirect regulatory associations can be identified by combining the information expressed in various events retrieved across different articles. For instance, the events *Regulation(C:geneA, T:geneZ)* and *Regulation(C:geneB, T:geneZ)* can be aggregated to present the hypothesis that *geneA* and *geneB* coregulate *geneZ*. Such hypothesis generation is greatly simplified by the fact that the events have been refined using the procedure described in Section 2.3 and the usage of a relational database, which allows efficient querying across events.

The indirect associations as implemented for the web application include coregulation and common binding partners (Table 3). These links have been precalculated and stored in the database, enabling fast retrieval of, for example, coregulators or genes that are targeted by a common regulator, facilitating the discovery of functional modules through text mining information. However, it needs to be stated that these associations are mainly hypothetical, as, for example, coregulators additionally require coexpression. Details on gene expression events can be found by browsing the sentences of specific genes as described in Section 4.1.

Table 3: Indirect associations between gene *A* and gene *B*, established by combining binding and regulatory events through a common interaction partner gene *Z*. Bindings are represented with $\times$ and for regulations $A > B$ means *A* regulates *B*.

Association	Interpretation
$A > Z < B$	*A* and *B* coregulate *Z*
$A < Z > B$	*A* and *B* are being regulated by *Z*
$A \times Z \times B$	*A* and *B* share a common binding partner *Z*

3. Results and Performance Evaluation

In this section, we present the evaluation of the EVEX resource from several points of view. First, we discuss the performance of the event extraction system used to produce the core set of events in EVEX, reviewing a number of published evaluations both within the BioNLP Shared Task and in other domains. Second, we present several evaluations of the methods and data employed specifically in the EVEX

resource in addition to the core event predictions: we review existing results as well as present new evaluations of the confidence scores and their correlation with event precision, the family-based generalization algorithms, and the novel event refinement algorithms introduced above. Finally, we discuss two biologically motivated applications of EVEX, demonstrating the usability of EVEX in real-world use cases.

3.1. Core Event Predictions. The Turku Event Extraction System (TEES), the source of the core set of EVEX events, was extensively evaluated on the BioNLP Shared Tasks. It was the winning system of the ST'09, achieving 46.73% recall, 58.48% precision, and 51.95%F-score [9]. In the current study, the original set of event predictions extracted from the PubMed 2009 distribution has been brought up to date using an improved version of TEES. This updated system was recently shown to achieve state-of-the-art results in the ST'11, obtaining 50.06% recall, 59.48% precision, and 54.37%F-score on the corresponding abstract-only GENIA subchallenge [12].

To assess the generalizability of the text mining results from domain-specific datasets to the whole of PubMed, a precision rate of 64% was previously obtained by manual evaluation of 100 random events [19]. In the same study, the named entities (i.e., gene and protein symbols) as extracted by BANNER were estimated to achieve a precision of 87%. These figures indicate that the performance of the various text mining components generalize well from domain-specific training data to the entire PubMed.

3.2. Confidence Values. To investigate the correlation of the confidence values (Section 2.2) to the correctness of the extracted events, we have measured the precision and recall rates of binding events between two genes, simulating a use case that involves finding related binding partners for a certain query gene (Section 4.1). This experiment was conducted on the ST'09 development set, consisting of 150 PubMed abstracts with 94 gold-standard binding pairs. For this dataset, 67 interacting pairs were found in EVEX, with confidence values ranging between -1.7 and 1.3. When evaluated against the gold-standard data, the whole set of predictions achieves 59.7% precision and 42.6% recall.

Using the confidence values for ranking, we have subsequently applied a cut-off threshold on the results, only keeping predictions with confidence values above the threshold. A systematic screening was performed between the interval of -1.7 and 1.3, using a step-size of 0.05 (60 evaluations). The results have been aggregated and summarized in Figure 2, depicting the average precision and recall values for each aggregated interval of 0.6 length. For example, a cut-off value between 0.10 and 0.70 (fourth interval) would result in an average precision rate of 70.0% and recall of 14.4%. Only taking the top ranked predictions, with a threshold above 0.7 (fifth interval), results in extremely high precision (91.9%) but only 4.8% recall. On the scale of EVEX, however, 4.8% recall would still translate to more than a million high-precision events.

3.3. EVEX Generalizations. As described in Section 2.1, the EVEX resource provides several algorithms to generalize gene symbols and their events, providing the opportunity to identify and aggregate equivalent events across various articles, accounting for lexical variants and synonymy. In a first step, a canonical form of the gene symbols is produced, increasing the proportion of symbols that can be matched to gene databases. This algorithm has previously been evaluated on the ST'09 training set, which specifically aims at identifying entities that are likely to match gene and protein symbol databases. By canonicalizing the symbols as predicted by BANNER, an increase of 11 percentage points in F-score was obtained [10].

The family-based generalizations have also been previously evaluated for both HomoloGene and Ensembl definitions. To expand the coverage of these generalizations, in this study, we have added definitions from Ensembl Genomes. The statistics on coverage of gene symbols, brought up to date by including the 2009–2011 abstracts, are depicted in Table 4. While only a small fraction of all unique canonical symbols matches the gene families from HomoloGene or Ensembl (Genomes) (between 3 and 6%), this small fraction accounts for more than half of all occurrences (between 51 and 61%). The family disambiguation algorithm thus discards a long tail of very infrequent canonical symbols. These findings are similar to the previous statistics presented by Van Landeghem et al. [10]. Additionally, the newly introduced families of Ensembl Genomes clearly provide a higher coverage: 8-9 percentage points higher than HomoloGene or Ensembl.

3.4. Event Refinement. By removing the chains of single-argument regulatory events, the refinement process simplifies and greatly reduces the heterogeneity in event structures, facilitating semantic interpretation and search for similar events. This process reduces the number of distinct event structures by more than 60%.

The main purpose of the event refinement algorithm, in combination with the pairwise view of the events, is to increase the coverage of finding related genes for a certain input query gene. When applying the algorithm as detailed in Section 2.3, the number of events with more than one gene symbol as direct argument increases from 1471 K to 1588 K, successfully generating more than a hundred thousand simplified events that can straightforwardly be parsed for pairwise relations.

It has to be noted, however, that the results of the refinement algorithm are merely used as an abstract layer to group similar events together and to offer quick access to relevant information. The original event structures as extracted by TEES are always presented to the user when detailed information is requested, allowing the user to reject or accept the inferences made by the refinement algorithm.

3.5. Biological Applications. The EVEX dataset and the associated web application have recently been applied in a focused study targeting the regulation of NADP(H) expression in *E. coli*, demonstrating the resource in a

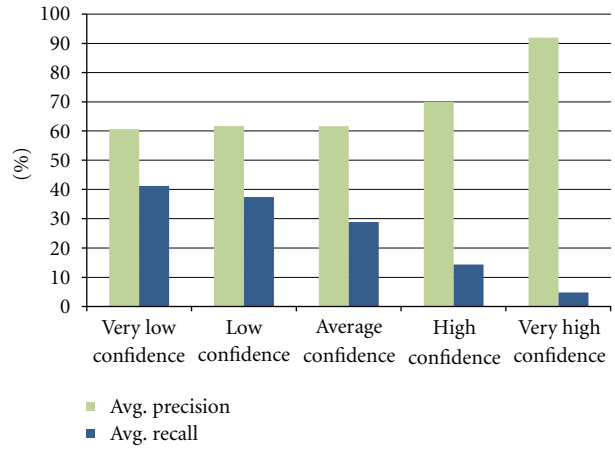


Figure 2: Evaluation of predicted binding events, measured against the gold-standard data of the ST'09 development set. By sorting the events according to their confidence values, a tradeoff between precision and recall is obtained.

Table 4: Gene symbol coverage comparison, showing the number of distinct canonical symbols as well as the number of different occurrences covered, out of the total number of 40.3 M extracted gene symbols.

	Distinct symbols		Occurrences	
Canonical	1833.1 K	100.0%	40.3 M	100.0%
HomoloGene	68.2 K	3.7%	21.1 M	52.3%
Ensembl	60.0 K	3.2%	20.9 M	51.8%
Ensembl Genomes	100.6 K	5.5%	24.3 M	60.1%

real-life biological use case, with encouraging results [20]. The Ensembl Genomes generalization was used to allow for homology-based inference, and the regulatory network extracted from EVEX was integrated with microarray coexpression data. As part of this study, 461 occurrences of two-argument events in the NADP(H) regulatory network were manually evaluated, with precision of 53%. This figure compares favorably with the BioNLP'09 Shared Task official evaluation results of 50% for binding events and 46% for regulation events, the only event types that allow more than one argument. The event occurrences that were judged to be correctly extracted were further evaluated for the correctness of the assignment of their arguments to Ensembl Genomes families: 72% of event occurrences had both of their arguments assigned to the correct family.

In a separate study, the suitability of the EVEX dataset and web application to the task of pathway curation was analyzed with a particular focus on recall [21]. When analysing three high-quality pathway models, TLR, mTOR and yeast cell cycle, 60% of all interactions could be retrieved from EVEX using the canonical generalization. A thorough manual evaluation further suggested that, surprisingly, the most common reason for a pathway interaction not being extracted is not a failure of the event extraction pipeline, but rather a lack of semantic coverage. In these cases, the interaction corresponds to an event type not defined in the ST'09 task and thus out of scope for the event extraction system. Only 11% of interactions in the evaluated pathways were not recovered due to a failure of the event extraction system. This result shows that the recall in EVEX, at least in the pathways under evaluation by Ohta et al., is clearly above the recall value published for the event extraction system in isolation. This increase can very likely be attributed to the volume of the event data in EVEX and the ability to aggregate several event occurrences into a single generalized event, where the failure to extract an individual event occurrence does not automatically mean the failure to extract the generalized event.

4. Web Application

To illustrate the functionality and features of the web application, we present a use case on a specific budding yeast gene, *Mec1*, which is conserved in *S. pombe*, *S. cerevisiae*, *K. lactis*, *E. gossypii*, *M. grisea,* and *N. crassa*. *Mec1* is required for meiosis and plays a critical role in the maintenance of genome stability. Furthermore, it is considered to be a homolog of the mammalian *ATR*/*ATM*, a signal transduction protein [22].

4.1. Gene Overview. The main functionality of the EVEX resource is providing fast access to relevant information and related biomolecular entities of a gene or pair of genes of interest. (Analysis of large gene lists is currently not supported, as such a bioinformatics use case is already covered by the publicly available MySQL database.) The most straightforward way to achieve this is through the canonical generalization, searching for a gene symbol or a pair of genes separated by a comma.

When typing the first characters of a gene symbol, a list of candidate matches is proposed, guiding the user to likely gene symbols found in text. The search page then automatically generates a listing of relevant biomolecular events, grouped by event type. At the top of the page, an overview of all regulators, regulated genes, and binding partners is provided, each accompanied with an example sentence. Further, coregulators are listed together with the number of coregulated genes (Section 2.5). Figure 3 shows the results when searching for *Mec1*. This overview lists 21 regulation targets, 11 regulators, 27 binding partners, and 263 coregulators. Within each category, the events are ranked by confidence, ranging from (very) high to average and (very) low (Section 2.2). Further, example sentences are always chosen to be those associated with the highest confidence score.

Selecting the target *RAD9*, the web application visualises all event structures expressing regulation of *RAD9* by *Mec1* (Figure 4). This enables a quick overview of the mechanisms through which the regulation is established, which can have a certain polarity (positive/negative) and may involve physical events such as phosphorylation or protein-DNA binding. The different types of event structures are coarsely grouped into categories of similar events and presented from most to least reliable using the confidence scores.

Exploring the relationship between *RAD9* and *Mec1* further, EVEX enables a search of all events linking these two genes through any direct or indirect association (Figure 5). This page provides conclusive evidence for a binding event between *RAD9* and *Mec1*. Further, both a *Mec1 regulates RAD9* and a *RAD9 regulates Mec1* event are presented. However, inspecting the sentences, the first one is obviously the only correct one. This illustrates the opportunity to use the large-scale event extraction results for pruning false positives of the text mining algorithm, as the false result only has 1 piece of evidence, and with a "very low" confidence, while the correct regulation is supported by 3 different evidence excerpts, two of which are of "high" confidence, and is thus displayed first.

Apart from the regulatory and binding mechanisms, the overview page also lists potential coregulations, enumerating targets that are regulated by both genes, such as *Rad53*. When accessing the details for this hypothesis, all evidence excerpts supporting both regulations are presented. Other indirect associations, such as common regulators and binding partners, can be retrieved equally fast.

Finally, the overview page of *Mec1* (Figure 3) contains additional relevant information including links to sentences stating events of *Mec1* without a second argument, grouped by event type. While these events incorporate only a single gene or protein and may not be very informative by themselves, they are highly relevant for information retrieval purposes, finding interesting sentences and articles describing specific processes such as protein catabolism or phosphorylation.

At the bottom of the overview page, a similar and even more general set of sentences can be found, providing pointers to relevant literature while still requiring manual analysis to determine the exact type of information. Such sentences, even though they contain no extracted events, may include useful background information on the gene such as relevant experimental studies, related diseases, or general functions and pathways.

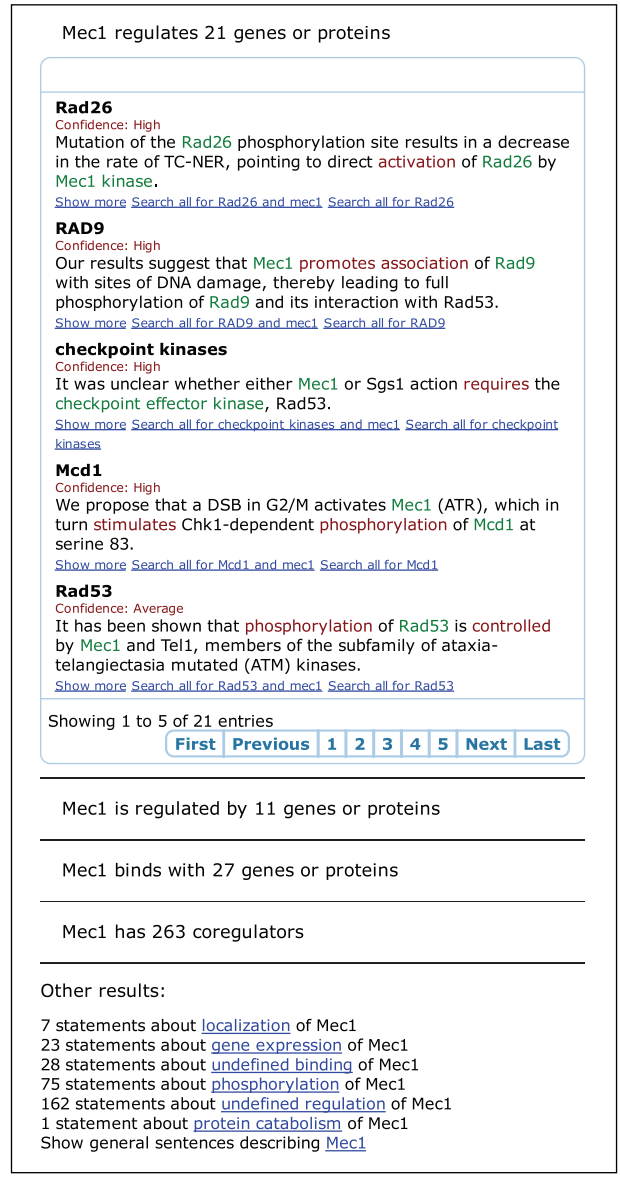


Figure 3: Search results for *Mec1* on the canonical generalization. An overview of directly associated genes is presented, grouped by event type. In the screenshot, only the box with regulation targets is shown, but the other event types may also be expanded. At the bottom, relevant links to additional sentences and articles are provided.

4.2. Homology-Based Inference. In comparative genomics, it is common practice to transfer functional annotations between related organisms for genes sharing sequence similarity [23, 24]. The EVEX resource provides such functionality for inferring interactions and other biomolecular events based on homology, by summarizing all events pertaining

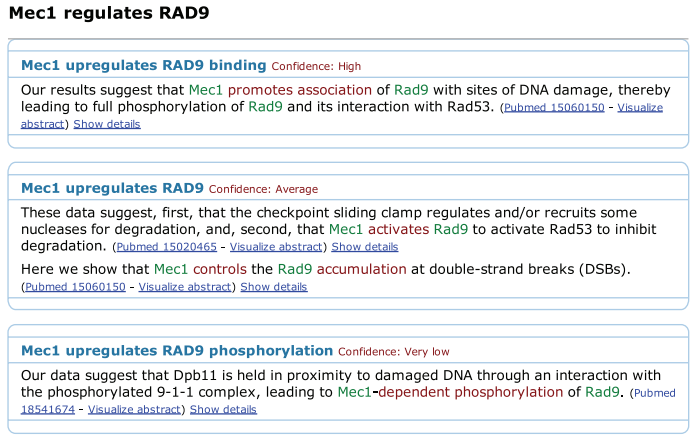


Figure 4: Detailed representation of all evidence supporting the regulation of RAD9 by Mec1. Regulatory mechanisms can have a certain polarity (positive/negative) and may involve physical events such as phosphorylation or protein-DNA binding.

to a certain family when searching for one of its members (Section 2.1).

For example, instead of only looking at the information for one particular gene symbol as described previously, we can extend the search through Ensembl Genomes and retrieve information on homologous genes and their synonyms. The generated listings of regulators and binding partners are structured in exactly the same way as before, but this time each symbol refers to a whole gene family rather than just one gene name.

Conducting such a generalized search for *Mec1*, EVEX retrieves interaction information for *Mec1* and its homologs. The resulting page presents not only results for the symbol *Mec1*, but also for common symbols which are considered synonyms on the gene-family level, such as *ATR*. This type of synonym expansion goes well beyond a simple keyword query.

For each gene family present in the text mining data, a family profile lists all genes and synonyms for a specific family, linking to the authoritative resources such as Entrez Gene and the Taxonomy database at NCBI. While *ESR1* is a known but deprecated synonym of *Mec1* [25], it is not considered as a viable synonym of *Mec1*, considering *Esr1* generally refers to the family of estrogen receptors. The synonym disambiguation algorithm of Van Landeghem et al. [10], which is the basis of the gene family generalizations, will thus prevent *Esr1* from being used as a synonym for *Mec1*. Reliable synonyms found in text do however include *ATR* and *SCKL*.

The EVEX web application includes several distinct methods of defining gene families (Section 2.1), each accommodating for specific organisms and use cases. For example, Ensembl Genomes defines rather coarse grained families resulting in a family of 19 evolutionarily conserved genes, including the budding yeast gene *Mec1*, its mammalian *ATR* orthologs, and genes from green algae and Arabidopsis. In contrast, the corresponding family defined by HomoloGene only includes the 6 conserved *Mec1* genes in the Ascomycota.

4.3. Manual Inspection of Text Mining Results. An important aspect of the EVEX web application is the ability to retrieve the original sentences and articles for all claims extracted from literature. In the previous sections, we have described how EVEX can assist in the retrieval of directly and indirectly associated genes and proteins by generating summary overviews. However, to be applicable in real-life use cases and to be valuable to a domain expert, it is necessary to distinguish trustworthy predictions from unreliable hypotheses. For this reason, automatically generated confidence values are displayed for each extracted interaction, ranging from (very) high to average and (very) low. On top of those, the site always provides the opportunity to inspect the textual evidence in detail.

Consider, for example, the phosphorylation of *RAD9*, regulated by *Mec1* (Figure 4). To allow a detailed inspection of this event, the web application integrates the *stav* visualiser [26], which was developed as a supporting resource for the ST'11 [2] (Figure 6). This open-source tool provides a detailed and easily graspable presentation of the event structures and the associated textual spans. To any user interested in the text mining details, this visualization provides valuable insights into the automated event extraction process. Additionally, the web application provides the opportunity to visualise whole PubMed abstracts with the *stav* visualiser, allowing a fast overview of event information contained within an abstract.

4.4. Site Navigation. To easily trace back previously found results, a session-based search history at the righthand side of the screen provides links to the latest searches issued on the site. Further, a box with related searches suggests relevant queries related to the current page. Finally, the web application provides a powerful method to browse indirectly associated information, by allowing the retrieval of nested and parent interactions of a specific event. For example, when accessing the details of *Mec1*'s regulation of *RAD9* phosphorylation and selecting the phosphorylation event,

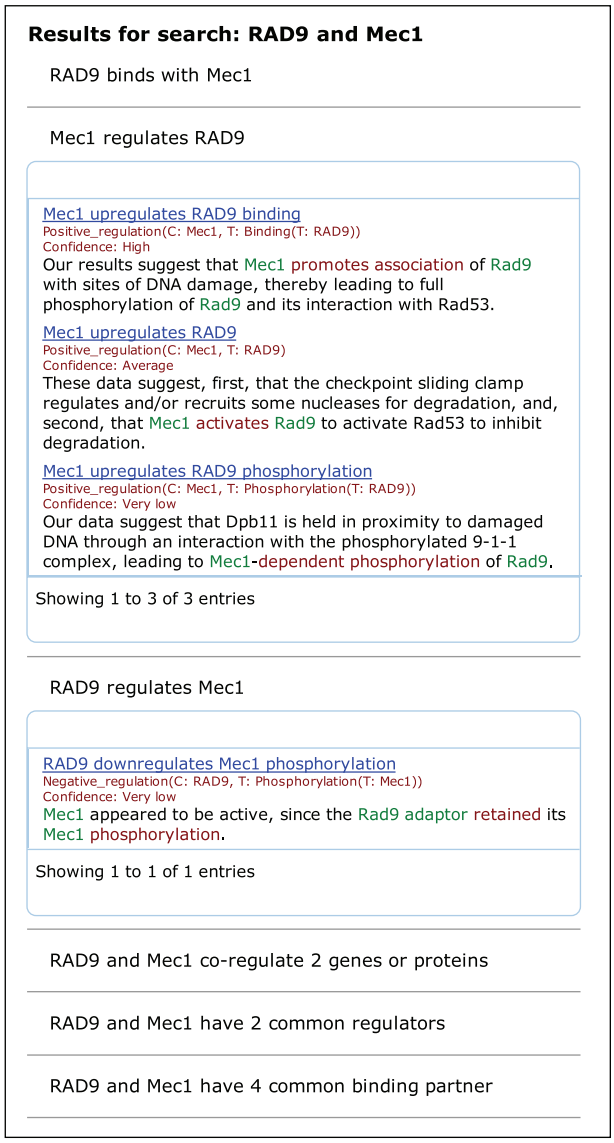


Figure 5: All events linking *Mec1* and *RAD9* through either direct or indirect associations. In the screenshot, only the regulation boxes are shown in detail, but the other event types may also be expanded. This page enables a quick overview of the mechanisms through which two genes interact, while at the same time highlighting false positive text mining results which can be identified by comparing confidence values and the evidence found in the sentences.

evidence is shown for many parent events involving different regulation polarities and various genes causing this specific phosphorylation. As such, we quickly learn that *RAD9* phosphorylation has many different potential regulators, such as *Ad5*, *Ad12,* and *C-Abl*. This sort of explorative information retrieval and cross-article discovery is exactly the type of usage aimed at by the EVEX resource.

5. Conclusions and Future Work

This paper presents a publicly available web application providing access to over 21 million detailed events among more than 40 million identified gene/protein symbols in nearly 6 million PubMed titles and abstracts. This dataset is the result of processing the entire collection of PubMed titles and abstracts through a state-of-the-art event extraction system and is regularly updated as new citations are added to PubMed. The extracted events provide a detailed representation of the textual statements, allowing for recursively nested events and different event types ranging from phosphorylation to catabolism and regulation. The EVEX web application is the first publicly released resource that provides intuitive access to these detailed event-based text mining results.

As the application mainly targets manual explorative browsing for supporting research in the life sciences, several steps are taken to allow for efficient querying of the large-scale event dataset. First, events are assigned confidence scores and ranked according to their reliability. Further, the events are refined to unify different event structures that have a nearly identical interpretation. Additionally, the events are aggregated across articles, accounting for lexical variation and generalizing gene symbols with respect to their gene family. This aggregation allows for efficient access to relevant information across articles and species. Finally, the EVEX web application groups events with respect to the involvement of pairs of genes, providing the users with a familiar gene-centric point of view, without sacrificing the expressiveness of the events. This interpretation is extended also to combinations of events, identifying indirect associations such as common coregulators and common binding partners, as a form of literature-based hypothesis generation.

There are a number of future directions that can be followed in order to extend and further improve the EVEX web application. The core set of events can be expanded by also processing all full-text articles from the open-access section of PubMed Central. Further, as BioNLP methods keep evolving towards more detailed and accurate predictions, the dataset can be enriched with new information, for example, by including epigenetics data as recently introduced by the BioNLP'11 Shared Task [2, 27] and integrating noncausal entity relations [28, 29]. Additionally, gene normalization data can be incorporated, enabling queries using specific gene or protein identifiers [30]. Finally, a web service may be developed to allow programmatic access to the EVEX web application, allowing bulk queries and result export for further postprocessing in various bioinformatics applications.

Acknowledgments

S. Van Landeghem would like to thank the Research Foundation Flanders (FWO) for funding her research and a travel grant to Turku. Y. Van de Peer wants to acknowledge support from Ghent University (Multidisciplinary Research Partnership Bioinformatics: from nucleotides to networks) and the Interuniversity Attraction Poles Programme (IUAP P6/25), initiated by the Belgian State, Science Policy Office (BioMaGNet). This work was partly funded by the Academy

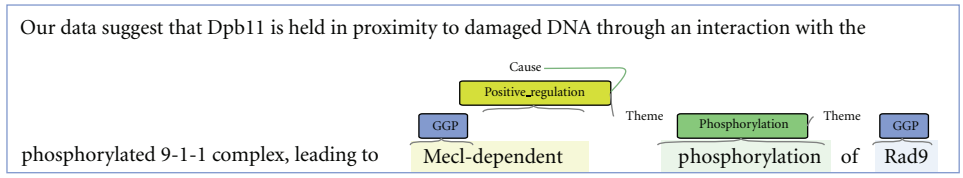


Figure 6: Visualization of a specific event occurrence by the stav text annotation visualiser. Genes and gene products ("GGPs") are marked, as well as the trigger words that refer to specific event types. Finally, arrows denote the roles of each argument in the event (e.g. Theme or Cause). This visualization corresponds to the formal bracketed format of the event: *Positive-regulation(C: Mec1, T:Phosphorylation(T:RAD9))*.

of Finland, and the computational resources were provided by CSC-IT Center for Science Ltd., Espoo, Finland and the Department of IT, University of Turku, Finland.

References

[1] J.-D. Kim, T. Ohta, S. Pyysalo, Y. Kano, and J. Tsujii, "Overview of BioNLP'09 shared task on event extraction," in *Proceedings of the BioNLP Workshop Companion Volume for Shared Task*, pp. 1–9, Association for Computational Linguistics, 2009.

[2] J.-D. Kim, S. Pyysalo, T. Ohta, R. Bossy, N. Nguyen, N. Nguyen, and J. Tsujii, "Overview of BioNLP shared task 2011," in *Proceedings of the BioNLP Workshop Companion Volume for Shared Task*, pp. 1–6, Association for Computational Linguistics, 2011.

[3] R. Homann and A. Valencia, "A gene network for navigating the literature," *Nature Genetics*, vol. 36, no. 7, aricle 664, 2004.

[4] T. Ohta, Y. Miyao, T. Ninomiya et al., "An intelligent search engine and GUI-based efficient MEDLINE search tool based on deep syntactic parsing," in *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pp. 17–20, Association for Computational Linguistics, 2006.

[5] D. Rebholz-Schuhmann, H. Kirsch, M. Arregui, S. Gaudan, M. Riethoven, and P. Stoehr, "EBIMed—text crunching to gather facts for proteins from Medline," *Bioinformatics*, vol. 23, no. 2, pp. e237–e244, 2007.

[6] M. A. Hearst, A. Divoli, H. H. Guturu et al., "BioText search engine: beyond abstract search," *Bioinformatics*, vol. 23, no. 16, pp. 2196–2197, 2007.

[7] S. Xu, J. McCusker, and M. Krauthammer, "Yale Image Finder (YIF): a new search engine for retrieving biomedical images," *Bioinformatics*, vol. 24, no. 17, pp. 1968–1970, 2008.

[8] S. Agarwal, H. Yu, and I. Kohane, "BioNOT: a searchable database of biomedical negated sentences," *BMC Bioinformatics*, vol. 12, Article ID 420, 2011.

[9] J. Björne, F. Ginter, S. Pyysalo, J. Tsujii, and T. Salakoski, "Scaling up biomedical event extraction to the entire PubMed," in *Proceedings of the BioNLP Workshop Companion Volume for Shared Task*, pp. 28–36, Association for Computational Linguistics, 2010.

[10] S. Van Landeghem, F. Ginter, Y. Van de Peer, and T. Salakoski, "EVEX: a PubMed-scale resource for homology-based generalization of text mining predictions," in *Proceedings of the BioNLP Workshop Companion Volume for Shared Task*, pp. 28–37, Association for Computational Linguistics, 2011.

[11] R. Leaman and G. Gonzalez, "BANNER: an executable survey of advances in biomedical named entity recognition," *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pp. 652–663, 2011.

[12] J. Björne, F. Ginter, and T. Salakoski, "Generalizing biomedical event extraction," *BMC Bioinformatics*, vol. 13, supplement 8, article S4, 2012.

[13] The UniProt Consortium, "Ongoing and future developments at the universal protein resource," *Nucleic Acids Research*, vol. 39, supplement 1, pp. D214–D219, 2011.

[14] E. W. Sayers, T. Barrett, D. A. Benson et al., "Database resources of the National Center for Biotechnology Information," *Nucleic Acids Research*, vol. 38, supplement 1, pp. D5–D16, 2009.

[15] P. Flicek, M. R. Amode, D. Barrell et al., "Ensembl 2011," *Nucleic Acids Research*, vol. 39, no. 1, pp. D800–D806, 2011.

[16] P. J. Kersey, D. Lawson, E. Birney et al., "Ensembl genomes: extending ensembl across the taxonomic space," *Nucleic Acids Research*, vol. 38, supplement 1, pp. D563–D569, 2009.

[17] K. Crammer and Y. Singer, "Ultraconservative online algorithms for multiclass problems," *Journal of Machine Learning Research*, vol. 3, no. 4-5, pp. 951–991, 2003.

[18] E. Segal, M. Shapira, A. Regev et al., "Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data," *Nature Genetics*, vol. 34, no. 2, pp. 166–176, 2003.

[19] J. Björne, F. Ginter, S. Pyysalo, J. Tsujii, and T. Salakoski, "Complex event extraction at PubMed scale," *Bioinformatics*, vol. 26, no. 12, Article ID btq180, pp. i382–i390, 2010.

[20] S. Kaewphan, S. Kreula, S. Van Landeghem, Y. Van de Peer, P. Jones, and F. Ginter, "Integrating large-scale text mining and co-expression networks: targeting NADP(H) metabolism in *E. coli* with event extraction," in *Proceedings of the 3rd Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM '12)*, 2012.

[21] T. Ohta, S. Pyysalo, and J. Tsujii, "From pathways to biomolecular events: opportunities and challenges," in *Proceedings of the BioNLP Workshop Companion Volume for Shared Task*, pp. 105–113, Association for Computational Linguistics, 2011.

[22] J. A. Carballo and R. S. Cha, "Meiotic roles of Mec1, a budding yeast homolog of mammalian ATR/ATM," *Chromosome Research*, vol. 15, no. 5, pp. 539–550, 2007.

[23] Y. Loewenstein, D. Raimondo, O. C. Redfern et al., "Protein function annotation by homology-based inference," *Genome Biology*, vol. 10, no. 2, article 207, 2009.

[24] S. Proost, M. Van Bel, L. Sterck et al., "PLAZA: a comparative genomics resource to study gene and genome evolution in plants," *Plant Cell*, vol. 21, no. 12, pp. 3718–3731, 2009.

[25] R. Kato and H. Ogawa, "An essential gene, ESR1, is required for mitotic cell growth, DNA repair and meiotic recombination in Saccharomyces cerevisiae," *Nucleic Acids Research*, vol. 22, no. 15, pp. 3104–3112, 1994.

[26] P. Stenetorp, G. Topić, S. Pyysalo, T. Ohta, J.-D. Kim, and J. Tsujii, "BioNLP Shared Task 2011: supporting resources," in

Proceedings of the BioNLP Workshop Companion Volume for Shared Task, pp. 112–120, Portland, Oregon, USA, 2011.

[27] J. Björne and T. Salakoski, "Generalizing biomedical event extraction," in *Proceedings of the BioNLP Workshop Companion Volume for Shared Task*, pp. 183–191, Association for Computational Linguistics, 2011.

[28] S. Pyysalo, T. Ohta, and J. Tsujii, "Overview of the entity relations (REL) supporting task of BioNLP Shared Task 2011," in *Proceedings of the BioNLP Workshop Companion Volume for Shared Task*, pp. 83–88, Association for Computational Linguistics, 2011.

[29] S. Van Landeghem, J. Björne, T. Abeel, B. De Baets, T. Salakoski, and Y. Van de Peer, "Semantically linking molecular entities in literature through entity relationships," *BMC Bioinformatics*, vol. 13, supplement 8, article S6, 2012.

[30] Z. Lu, H. Y. Kao, C. H. Wei et al., "The gene normalization task in BioCreative III," *BMC Bioinformatics*, vol. 12, supplement 8, article S2, 2011.

Hindawi Publishing Corporation
Advances in Bioinformatics
Volume 2012, Article ID 509126, 12 pages
doi:10.1155/2012/509126

Research Article

BioEve Search: A Novel Framework to Facilitate Interactive Literature Search

Syed Toufeeq Ahmed,[1] Hasan Davulcu,[2] Sukru Tikves,[2] Radhika Nair,[2] and Zhongming Zhao[1,3]

[1] *Department of Biomedical Informatics, Vanderbilt University, Nashville, TN 37232, USA*
[2] *Department of Computer Science and Engineering, Arizona State University, Tempe, AZ 85281, USA*
[3] *Department of Cancer Biology, Vanderbilt University School of Medicine, Nashville, TN 37232, USA*

Correspondence should be addressed to Syed Toufeeq Ahmed, syed.t.ahmed@vanderbilt.edu

Received 15 November 2011; Revised 7 March 2012; Accepted 28 March 2012

Academic Editor: Jin-Dong Kim

Background. Recent advances in computational and biological methods in last two decades have remarkably changed the scale of biomedical research and with it began the unprecedented growth in both the production of biomedical data and amount of published literature discussing it. An automated extraction system coupled with a cognitive search and navigation service over these document collections would not only save time and effort, but also pave the way to discover hitherto unknown information implicitly conveyed in the texts. *Results*. We developed a novel framework (named "BioEve") that seamlessly integrates Faceted Search (Information Retrieval) with Information Extraction module to provide an interactive search experience for the researchers in life sciences. It enables guided step-by-step search query refinement, by suggesting concepts and entities (like genes, drugs, and diseases) to quickly filter and modify search direction, and thereby facilitating an enriched paradigm where user can discover related concepts and keywords to search while information seeking. *Conclusions*. The BioEve Search framework makes it easier to enable scalable interactive search over large collection of textual articles and to discover knowledge hidden in thousands of biomedical literature articles with ease.

1. Background

Human genome sequencing marked the beginning of the era of large-scale genomics and proteomics, leading to large quantities of information on sequences, genes, interactions, and their annotations. In the same way that the capability to analyze data increases, the output by high-throughput techniques generates more information available for testing hypotheses and stimulating novel ones. Many experimental findings are reported in the -omics literature, where researchers have access to more than 20 million publications, with up to 4,500 new ones per day, available through to the widely used PubMed citation index and Google Scholar. This vast increase in available information demands novel strategies to help researchers to keep up to date with recent developments, as *ad hoc* querying with Boolean queries is tedious and often misses important information.

Even though PubMed provides an advanced keyword search and offers useful query expansion, it returns hundreds or thousands of articles as result; these are sorted by publication date, without providing much help in selecting or drilling down to those few articles that are most relevant regarding the user's actual question. As an example of both the amount of available information and the insufficiency of naïve keyword search, the name of the protein *p53* occurs in 53,528 PubMed articles, and while a researcher interested specifically in its role in *cancer* and its interacting partners might try the search "*p53 cancer interaction*" to narrow down the results, this query still yields 1,777 publications, enough for months of full-time reading [1]. Nonetheless, PubMed is a very widely used free service and is providing an invaluable service to the researchers around the world. In March 2007, PubMed served 82 million (statistics of Medline searches: http://www.nlm.nih.gov/bsd/medline_growth.html) query

searches and the usage is ever increasing. A few commercial products are currently available that provide additional services, but they also rely on basic keyword search, with no real discovery or dynamic faceted search. Examples are OvidSP and Ingenuity Answers, both of which support bookmarking as one means of keeping track of visited citations. Research tools such as EBIMed (EBIMed: http://www.ebi.ac.uk/Rebholz-srv/ebimed/index.jsp) [2] and AliBaba (AliBaba: http://alibaba.informatik.hu-berlin.de) [3] provide additional cross-referencing of entities to databases such as UniProt or to the GeneOntology. They also try to identify relations between entities, such as protein-protein interactions, functional protein annotations, or gene-disease associations.

Search tools should provide dedicated and intuitive strategies that help to find relevant literature, starting with initial keyword searches and drilling down results via overviews enriched with autogenerated suggestions to refine queries. One of the first steps in biomedical text mining is to recognize named entities occurring in a text, such as genes and diseases. Named entity recognition (NER) is helpful to identify relevant documents, index a document collection, and facilitate information retrieval (IR) and semantic searches [4]. A step on top of NER is to normalize each entity to a base form (also called grounding and identification); the base form often is an identifier from an existing, relevant database; for instance, protein names could be mapped to UniProt IDs [5, 6]. Entity normalization (EN) is required to get rid of ambiguities such as homonyms, and map synonyms to one and the same concept. This further alleviates the tasks of indexing, IR, and search. Once named entities have been identified, systems aim to extract relationships between them from textual evidences; in the biomedical domain, these include gene-disease associations and protein-protein interactions. Such relations can then be made available for subsequent search in relational databases or used for constructing particular pathways and entire networks [7].

Information extraction (IE) [8–11] is the extraction of salient facts about prespecified types of events, entities [12], or relationships from free text. Information extraction from free text utilizes shallow-parsing techniques [13], part-of-speech tagging [14], noun and verb phrase chunking [15], predicate-subject and object relationships [13], and learned [8, 16, 17] or hand-build patterns [18] to automate the creation of specialized databases. Manual pattern engineering approaches employ shallow parsing with patterns to extract the interactions. In the system presented in [19], sentences are first tagged using a dictionary-based protein name identifier and then processed by a module which extracts interactions directly from complex and compound sentences using regular expressions based on part of speech tags. IE systems look for entities, relationships among those entities, or other specific facts within text documents. The success of information extraction depends on the performance of the various subtasks involved.

The Suiseki system of Blaschke et al. [20] also uses regular expressions, with probabilities that reflect the experimental accuracy of each pattern to extract interactions into predefined frame structures. Genies [21] utilizes a grammar-based natural language processing (NLP) engine for information extraction. Recently, it has been extended as GeneWays [22], which also provides a Web interface that allows users to search and submit papers of interest for analysis. The BioRAT system [23] uses manually engineered templates that combine lexical and semantic information to identify protein interactions. The GeneScene system [24] extracts interactions using frequent preposition-based templates.

Over the last years, a focus has been on the extraction of protein-protein interactions in general, recently including extraction from full text articles, relevance ranking of extracted information, and other related aspects (see, for instance, the BioCreative community challenge [25]). The BioNLP'09 Shared Task concentrated on recognition of more fine-grained molecular events involving proteins and genes [26]. Both papers give overviews over the specific tasks and reference articles by participants.

One of the first efforts to extract information on biomolecular events was proposed by Yakushiji et al. [27]. They implemented an argument structure extractor based on full sentence parses. A list of target verbs have specific argument structures assigned to each. Frame-based extraction then searches for filler of each slot required according to the particular arguments. On an small in-house corpus, they found that 75% of the errors can be attributed to erroneous parsing and another 7% to insufficient memory; both causes might have less impact on recent systems due to more accurate parsers and larger memory.

Ding et al. [28] studied the extraction of protein-protein interactions using the Link Grammar parser. After some manual sentence simplification to increase parsing efficiency, their system assumed an interaction whenever two proteins were connected via a link path; an adjustable threshold allowed to cut off too long paths. As they used the original version of Link Grammar, Ding et al. [28] argued that adaptations to the biomedical domain would enhance the performance.

An information extraction application analyzes texts and presents only the specific information from them that the user is interested in [29]. IE systems are knowledge intensive to build and are to varying degrees tied to particular domains and scenarios such as target schema. Almost all IE applications start with fixed target schema as a goal and are tuned to extract information from unstructured text that will fit the schema. In scenarios where target schema is unknown, open information extraction systems [30] like KnowItNow [31] and TextRunner [32] allow rules to be defined easily based on the extraction need. An hybrid application (IR + IE) that leverages the best of information retrieval (ability to relevant texts) and information extraction (analyze text and present only specific information user is interested in) would be ideal in cases when the target extraction schema is unknown. An iterative loop of IR and IE with user's feedback will be potentially useful. For this application, we will need main components of IE system (like parts-of-speech tagger, named entity taggers, shallow parsers) preprocesses the text before being indexed by a custom-built augmented index that helps retrieve queries of the type "Cities such as ProperNoun(Head(NounPhrase))." Cafarella and Etzioni

[33] have done work in this direction to build a search engine for natural language and information extraction applications.

Exploratory search [34] is a topic that has grown from the fields of information retrieval and information seeking but has become more concerned with alternatives to the kind of search that has received the majority of focus (returning the most relevant documents to a Google-like keyword search). The research is motivated by questions like "what if the user does not know which keywords to use?" or "what if the user is not looking for a single answer?". Consequently, research began to focus on defining the broader set of information behaviors in order to learn about situations when a user is—or feels—limited by having only the ability to perform a keyword search (source: http://en.wikipedia.org/wiki/Exploratory_search). Exploratory search can be defined as specialization of information exploration which represents the activities carried out by searchers who are either [35]:

(1) unfamiliar with the domain of their goal (i.e., need to learn about the topic in order to understand how to achieve their goal);

(2) unsure about the ways to achieve their goals (either the technology or the process); or even

(3) unsure about their goals in the first place.

A faceted search system (or parametric search system) presents users with key value metadata that is used for query refinement [36]. By using facets (which are metadata or class labels for entities such as genes or diseases), users can easily combine the hierarchies in various ways to refine and drill down the results for a given query; they do not have to learn custom query syntax or to restart their search from scratch after each refinement. Studies have shown that users prefer faceted search interfaces because of their intuitiveness and ease of use [37]. Hearst [38] shares her experience, best practices, and design guidelines for faceted search interfaces, focusing on supporting flexible navigation, seamless integration with directed search, fluid alternation between refining and expanding, avoidance of empty results sets, and most importantly making users at ease by retaining a feeling of control and understanding of the entire search and navigation process. To improve web search for queries containing named entities [39], automatically identify the subject classes to which a named entity might refer to and select a set of appropriate facets for denoting the query.

Faceted search interfaces have made online shopping experiences richer and increased the accessibility of products by allowing users to search with general keywords and browse and refine the results until the desired sub-set is obtained (SIGIR'2006 Workshop on Faceted Search (CFP): http://sites.google.com/site/facetedsearch/). Faceted navigation delivers an experience of progressive query refinement or elaboration. Furthermore, it allows users to see the impact of each incremental choice in one facet on the choices in other facets. Faceted search combines faceted navigation with text search, allowing users to access (semi) structured content, thereby providing support for discovery and exploratory search, areas where conventional search falls short [40].

2. Approach

In an age of ever increasing published research documents (available in search-able textual form) containing amounts of valuable information and knowledge that are vital to further research and understanding, it becomes imperative to build tools and systems that enable easier and quick access to right information the user is seeking for, and this has already become an information overload problem in different domains. Information Extraction (IE) systems provide an structured output by extracting nuggets of information from these text document collections, for a defined schema. The output schema can vary from simple pairwise relations to a complex, nested multiple events.

Faceted search and navigation is an efficient way to browse and search over a structured data/document collection, where the user is concerned about the completeness of the search, not just top ranked results. Faceted search system needs structured input documents, and IE systems extract structured information from text documents. By combining these two paradigms, we are able to provide faceted search and navigation over unstructured text documents, and, with this fusion, we are also able to leverage real utility of information extraction, that is, finding hidden relationships as the user goes through a search process, and to help refine the query to more satisfying and relevant level, all while keeping user feel incontrol of the whole search process.

We developed BioEve Search (http://www.bioeve.org/) framework to provide fast and scalable search service, where users can quickly refine their queries and drill down to the articles they are looking for in a matter of seconds, corresponding to a few number of clicks. The system helps identify hidden relationships between entities (like drugs, diseases, and genes), by highlighting them using a tag cloud to give a quick visualization for efficient navigation. In order to have sufficient abstraction between various modules (and technologies used) in this system, we have divided this framework into four different layers (refer to Figure 1) and they are (a) Data Store layer, (b) Information Extraction layer, (c) Faceting layer, and (d) Web Interface layer. Next sections explain each layer of this framework in more details.

2.1. Data Store Layer. The Data Store layer preprocesses and stores the documents in an indexed data store to make them efficiently accessible to the modules of upper layer (information extraction layer). Format conversion is needed sometimes (from ASCII to UTF-8 or vice versa), or XML documents need to be converted to text documents before being passed to next module. After the documents are in the required format and cleansed, they are stored in a indexed data store for efficient and fast access to either individual documents or the whole collections. The data store can be implemented using an Indexer service like (Apache Lucene (Lucene: http://lucene.apache.org/) or any database like MySQL). The Medline dataset is available as zipped XML

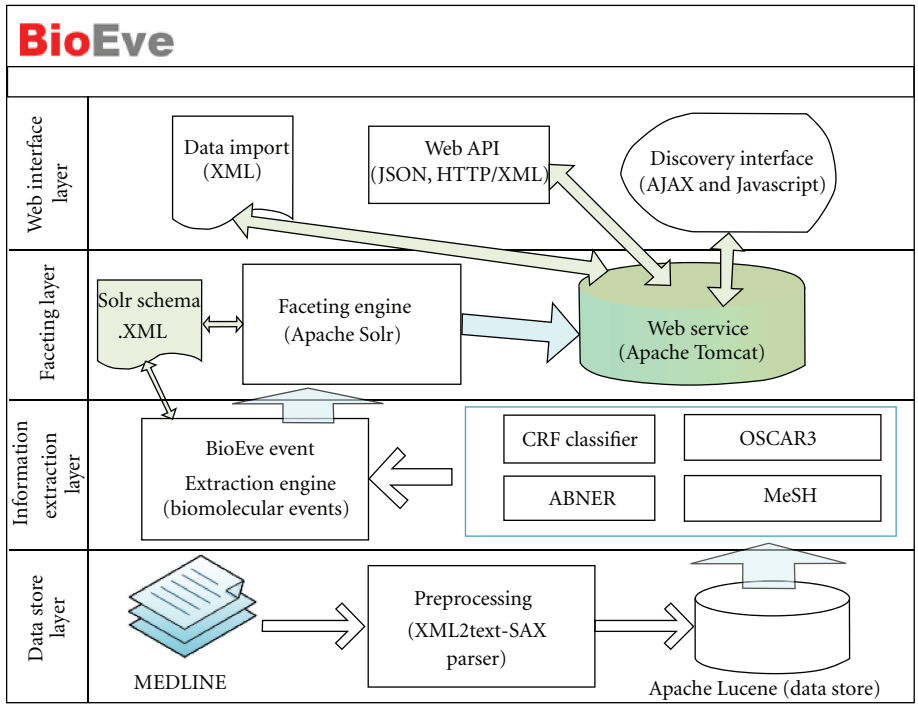


Figure 1: BioEve search framework architecture.

files that needed XML2 text conversion, after which we could ingest them into an indexer, Apache Lucene in our case. Such an indexer allows for faster access and keyword-based text search to select a particular subset of abstracts for further processing.

2.2. Information Extraction Layer. For recognizing different gene/protein names, DNA, RNA, cell line, and cell types, we leveraged ABNER [41], A Biomedical Named Entity Recognizer. We used OSCAR3 (Oscar3: http://sourceforge.net/projects/oscar3-chem/) (Open Source Chemistry Analysis Routines) to identify chemical names and chemical structures. To annotate disease names, symptoms, and causes, we used a subset of the Medical Subject Heading (MeSH) dataset (MeSH: http://www.nlm.nih.gov/mesh/).

2.2.1. Annotating Biomolecular Events in the Text. A first step towards bio-event extraction is to identify phrases in biomedical text which indicate the presence of an event. The labeled phrases are classified further into nine event types (based on the Genia corpus (BioNLP'09 Shared Task 1: http://www.nactem.ac.uk/tsujii/GENIA/SharedTask/)). The aim of marking such interesting phrases is to avoid looking at the entire text to find participants, as deep parsing of sentences can be a computationally expensive process, especially for the large volumes of text. We intend to mark phrases in biomedical text, which could contain a potential event, to serve as a starting point for extraction of event participants. Section 6.1 gives more details about our experimentations with classification and annotation of biomedical entities.

All the classification and annotation were done *offline* before the annotated articles are indexed for the search as once an article is classified and annotated with different entity types, it does not need to be processed again for each search query. This step can be done preindexing and as a batch process.

2.3. Faceting Layer

2.3.1. Faceting Engine. To provide faceted classification and navigation over these categories (facets), many off-the-shelf systems are available such as in academia; Flamenco project (Flamenco: http://flamenco.berkeley.edu/) (from University of California Berkeley) and mspace (mspace: http://mspace.fm/) (University of Southampton) and in enterprise area; Apache Solr (Apache Solr: http://lucene.apache.org/solr/) and Endeca (Endeca: http://www.endeca.com/). We used the Apache Solr library for faceted search, which also provides an enterprise quality full-text search.

2.3.2. Shared Schema between IE Layer and Faceting Layer. In order to facilitate indexing and faceting over the extracted semi-structured text articles, both IE layer and faceting layer needs to share a common schema. A sample of shared schema used for enabling interaction between these layers is shown in Scheme 1.

2.4. Web Interface Layer. With the advent of Web 2.0 technologies, web-based interfaces have undergone delightful improvements and now provide rich dynamic experiences. Key component in this layer is a user interface that connects the user with the web service from the faceting layer and provides features that allow search, selection of facet/values, refinement, query restart, and dynamic display of a result

```
<field name="pmid" type="string" indexed="true" stored="true" required="true"/>
<field name="text" type="text" indexed="true" stored="true" multiValued="true"/>
<field name="title" type="text" indexed="true" stored="true"/>
<field name="gene" type="string" indexed="true" stored="true" multiValued="true"/>
<field name="drug" type="string" indexed="true" stored="true" multiValued="true"/>
<field name="disease" type="string" indexed="true" stored="true" multiValued="true"/>
```

Scheme 1

set as user interacts and navigates. It also provides the bulk import of data for further analysis of the faceting/extraction.

The web interface provides following features for interactive search and navigation. The interface presents a number of entities types (on the left panel) along with the specific instances/values, from previous search results, and the current query. Users can choose any of the highlighted values of these entity types to interactively refine the query (add new values/remove any value from the list with just one click) and thereby drill down to the relevant articles quickly without actually reading the entire abstracts. Users can easily remove any of the previous search terms, thus widening the current search. We implemented the BioEve user interface using AJAX (AJAX: http://evolvingweb.github.com/ajax-solr/), Javascript, and JSON to provide rich dynamic experience. The web interface runs on an Apache Tomcat server. Next section explains about navigation aspect of the user interface.

3. User Interface: A Navigation Guide

Search interface is divided into left and right panels, see Figure 2, basically displaying enriched keywords and results, respectively.

Left panel: it offers suggestions and insights (based on cooccurrence frequency with the query terms) for different entities types, such as genes, diseases, and drugs/chemicals.

(i) Left panel shows navigation/refinement categories (genes, diseases, and drugs); users can click on any of the entity names (in light blue) to refine the search. By clicking on an entity, the user adds that entity to the search and the results on the right panel are refreshed on the y to reflect the refined results.

(ii) Users can add or remove any number of refinements to the current search query until they reach the desired results set (shown in the right panel).

Right panel: it shows the user's current search results and is automatically refreshed based on user's refinement and navigation choices on the left panel.

(i) The top of the panel shows users current query terms and navigation so far. Here, users can also deselect any of the previously selected entities or even all of them by single click on "remove all." By deselecting any entities, user is essentially expanding the search and the results in the right panel are refreshed *on the fly* to remaining query entities to offer a dynamic navigation experience.

(ii) Abstracts results on this panel show "title" of the abstract (in light red), full abstract text (in black, if abstract text is available).

(iii) Below the full abstract text, the list of entities mentioned in that abstracts (in light blue) is shown. These entities names are clickable and will start a new search for that entity name, with a single click.

(iv) A direct URL is also provided to the abstract page on http://pubmed.gov in case the user wants to access additional information such as authors, publication type, or links to a full-text article.

4. Interactive Search and Navigation: A Walk through and behind the Scenes

Let us start an example search process, say with the query "cholesterol" and the paragraph titled "behind-the-scenes" gives details of the computational process behind the action.

(1) The autocomplete feature helps in completion of the name while typing if the word is previously mentioned in the literature, which is the case here with "cholesterol."

Behind-the-scenes: as user starts typing, the query is tokenized (in case of multiple words) and search is made to retrieve word matches (and not the result rows yet) using the beginning with the characters user has already typed, and this loop continues. Technologies at play are jQuery, AJAX, and faceting feature of Apache Solr. Once the query is submitted by the user, the results rows also contain the annotated entity names and these are used to generate tag clouds, using the faceting classification entity frequency count.

The search results in 27177 articles hits (Figure 3). Those are a lot of articles to read. How about narrowing down these results with some insights given by BioEve Search?

(2) In left panel, "hepatic lipase" is highlighted; let us click on that as it shows some important relationship between "cholesterol" and "hepatic lipase." The search results are now narrowed down to 195 articles from 27177 (Figure 4). That is still a lot of articles to read this afternoon, how about some insights on diseases.

Behind-the-scenes: once user click on a highlighted entity name in tag cloud, this term (*gene: "hepatic lipase"*) is added to the search filter and the whole search process and tag could be generated again for the new query.

You can see disease "hyperthyroidism" highlighted in Figure 5.

(3)Selecting "hyperthyroidism" drills results down to 3, as can be seen in Figure 6.

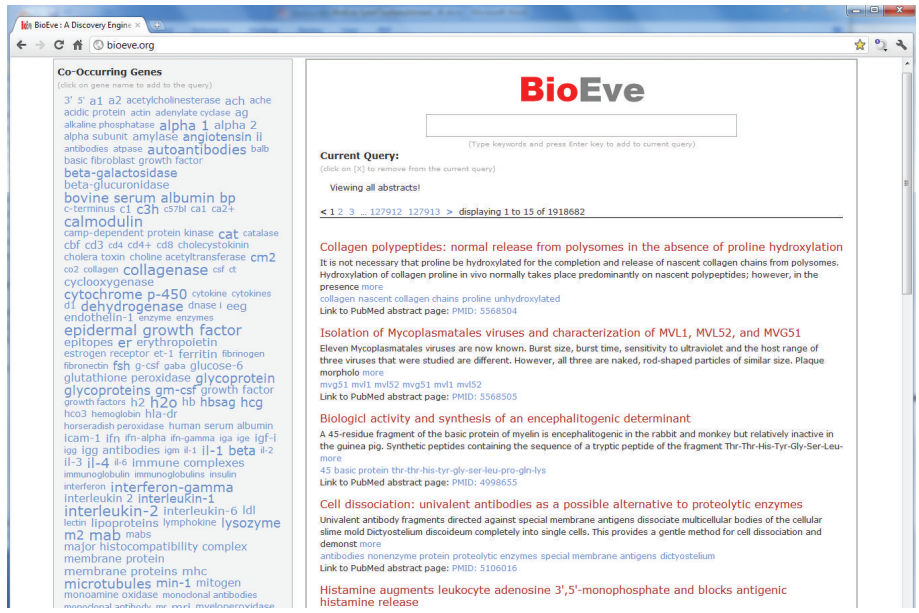


Figure 2: A sample screen shot of the main search screen. Left panel shows clickable top relevant entities, which if selected refines the query and results dynamically. User can deselect any of the previously selected entities to refine query more, and the results are updated dynamically to reflect the current selected list of entities.

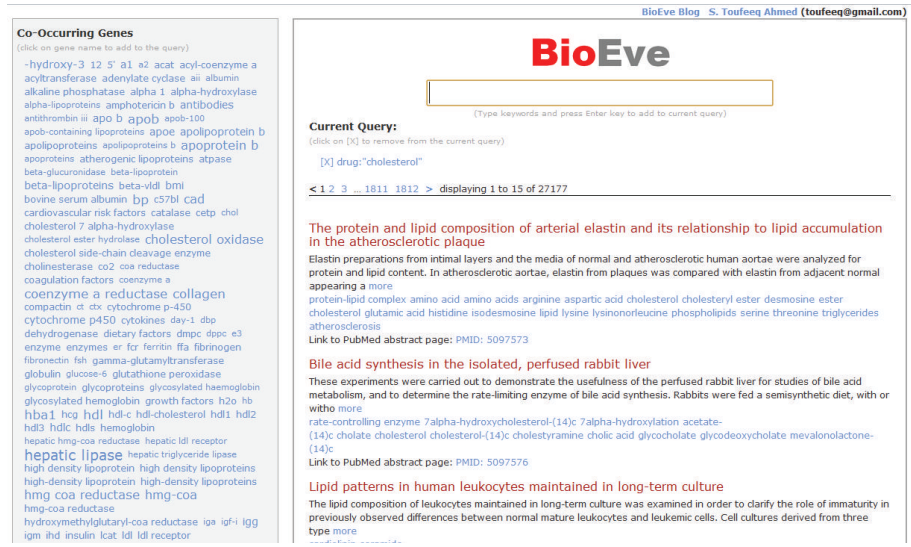


Figure 3: A sample result set with the query "cholesterol."

The top result is about "Treatment of hyperthyroidism: effects on hepatic lipase, lipoprotein lipase, LCAT and plasma lipoproteins". With few clicks user can refine search results to more relevant articles.

5. Initial User Reviews and Feedback

We asked three life science researchers to review and provide feedback on ease of search and novelty of the system, and shown below is their feedback (paraphrased). Their names and other details are removed for privacy purposes.

5.1. *Researcher One, P.h.D, Research Fellow, Microbiology, University of California, Berkeley*

> " I am impressed by ease of its use." "When I have the confidence that BioEve is indexing all the data without missing any critical article, *I*

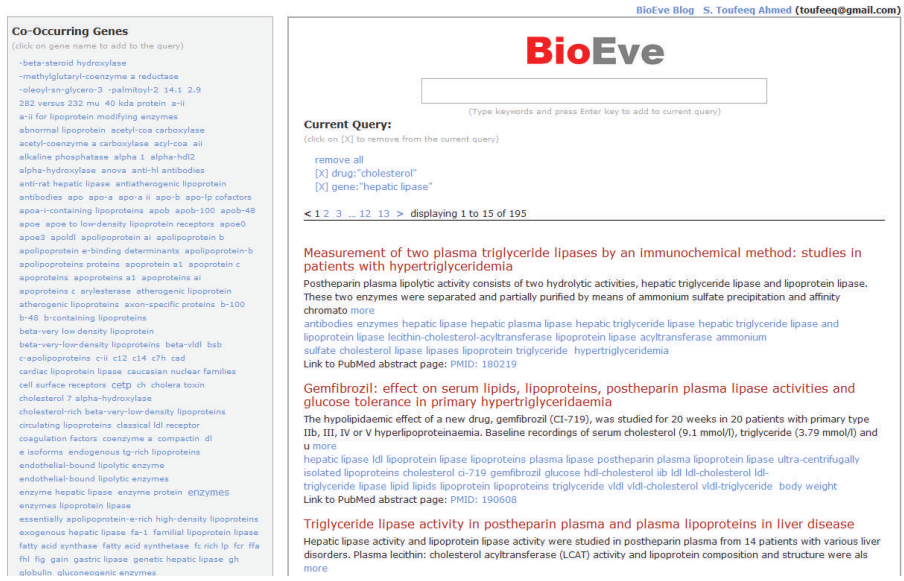


Figure 4: "Hepatic-lipase" selected.

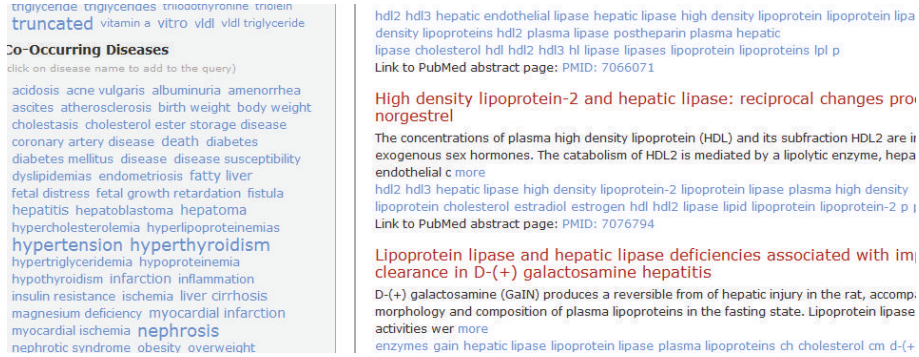


Figure 5: "Hyperthyroidism" highlighted.

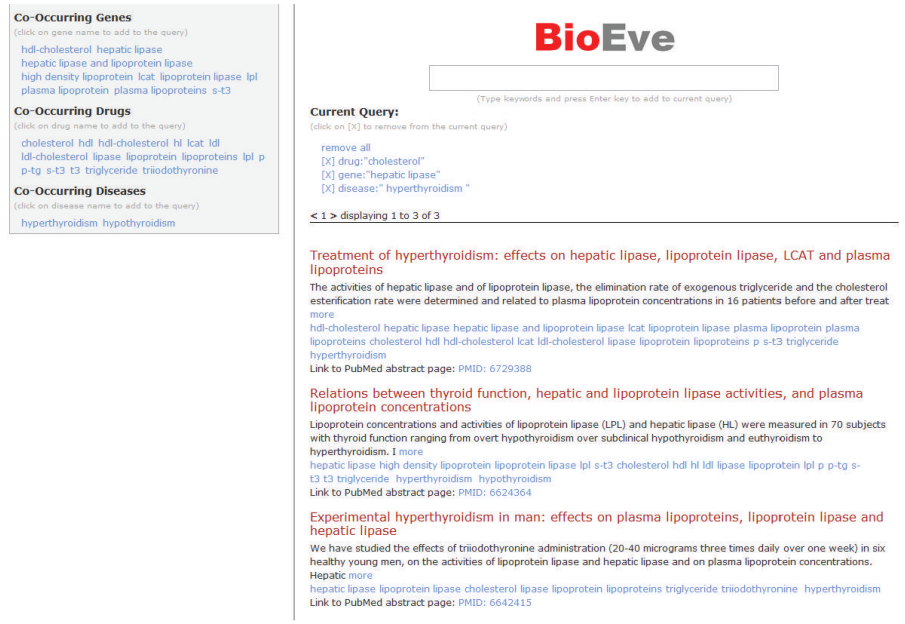


Figure 6: Final refined search results.

> *will be compelled to use this search tool.* I believe a finished product will be immensely useful and could become a popular tool for life science researchers."

5.2. Researcher Two, P.h.D, Investigator and Head, Molecular Genetics Laboratory

> "*You have a powerful search.* Synchronize this with MEDLINE. Connect with more databases, OMIM, Entrez Gene …. You can get cell line database from ATCC.org."

5.3. Researcher Three, P.h.D, Postdoc Researcher, Faculty of Kinesiology, University of Calgary

> "I particularly like the idea of having larger fonts for the more relevant terms highlighting what is researched more often."

6. Methods

6.1. Information Extraction: Annotating Sentences with Biomolecular Event Types. The first step towards bioevent extraction is to identify phrases in biomedical text which indicate the presence of an event. The aim of marking such interesting phrases is to avoid looking at the entire text to find participants. We intend to mark phrases in biomedical text, which could contain a potential event, to serve as a starting point for extraction of event participants. We experimented with well-known classification approaches, from a naïve Bayes classifier to the more sophisticated machine classification algorithms Expectation Maximization, Maximum Entropy, and Conditional Random Fields. Overview of different classifiers applied at different levels of granularity and the features used by these classifiers is shown in Table 1.

For naïve Bayes classifier implementation, we utilized WEKA (WEKA: http://www.cs.waikato.ac.nz/ml/weka/) library, a collection of machine learning algorithms for data mining tasks, for identifying single label per sentence approach. WEKA does not support multiple labels for the same instance. Hence, we had to include a tradeoff here by including the first encountered label in the case where the instance had multiple labels. For Expectation Maximization (EM) and Maximum Entropy (MaxEnt) algorithms, we used classification algorithms from MALLET library (MALLET: http://mallet.cs.umass.edu/index.php). Biomedical abstracts are split into sentences. For training purposes, plain text sentences are transformed into training instances as required by MALLET.

6.1.1. Feature Selection for Naïve Bayes, EM, and MaxEnt Classifiers. For the feature sets mentioned below, we used the TF-IDF representation. Each vector was normalized based on vector length. Also, to avoid variations, words/phrases were converted to lowercase. Based on WEKA library token delimiters, features were filtered to include those which had an alphabet as a prefix, using regular expressions. For example, features like −300 bp were filtered out, but features like *p*55, which is a protein name, were retained. We experimented with the list of features described below, to understand how well each feature suits the corpus under consideration.

Table 1: Classification approaches used: Naïve Bayes classifier (NBC), NBC + Expectation Maximization (EM), Maximum Entropy (MaxEnt), Conditional Random Fields (CRFs).

Granularity	Features	Classifier
Single label, Sentence level	Bag-of-words (BOW)	NBC
	BOW + gene names boosted	
	BOW + trigger words boosted	
	BOW + gene names and trigger words boosted	
Multiple labels	BOW	NBC + EM
Sentence level		MaxEnt
Event trigger phrase labeling	BOW + 3-gram and 4-gram prefixes and suffixes + orthographic features + trigger phrase dictionary	CRFs

(i) Bag-of-words model: this model classified sentences based on word distribution.

(ii) Bag-of-words with gene names boosted: the idea was to give more importance to words, which clearly demarcate event types. To start with, we included gene names provided in the training data. Next, we used the ABNER (ABNER: http://pages.cs.wisc.edu/~bsettles/abner/), a gene name tagger, to tag gene names, apart from the ones already provided to us. We boosted weights for renamed feature "protein", by 2.0.

(iii) Bag-of-words with event trigger words boosted: we separately tried boosting event trigger words. The list of trigger words was obtained from training data. This list was cleaned to remove stop words. Trigger words were ordered in terms of their frequency of occurrence with respect to an event type, to capture trigger words which are most discriminative.

(iv) Bag-of-words with gene names and event trigger words boosted: the final approach was to boost both gene names and trigger words together. Theoretically, this approach was expected to do better than previous two feature sets discussed. Combination of discriminative approach of trigger words and gene name boosting was expected to train the classifier better.

6.1.2. Evaluation of Sentence Level Classification Using Naïve Bayes Classifier. This approach assigns a single label to

Table 2: Single label, sentence level results.

Classifier	Feature set	Precision
NBC	Bag-of-words	62.39%
	Bag-of-words + gene name boosting	50.00%
	Bag-of-words + trigger word boosting	49.92%
	Bag-of-words + trigger word boosting + Gene name boosting	49.77%
	Bag-of-POS tagged words	43.30%

each sentence. For evaluation purposes, the classifier is tested against GENIA development data. For every sentence, evaluator process checks if the event type predicted is the most likely event in that sentence. In case a sentence has more than one event with equal occurrence frequency, classifier predicted label is compared with all these candidate event types. The intent of this approach was to just understand the features suitable for this corpus. Classifier evaluated was NaiveBayesMultinomial classifier from Weka (http://www.cs.waikato.ac.nz/ml/weka/) library, which is a collection of machine learning algorithms for data mining tasks. Table 2 shows precision results for NBC classifier with different feature sets for single label per sentence classification.

6.2. Conditional Random Fields Based Classifier. Conditional Random fields (CRFs) are undirected statistical graphical models, a special case of which is a linear chain that corresponds to a conditionally trained finite-state machine [41]. CRFs in particular have been shown to be useful in parts-of-speech tagging [42] and shallow parsing [42]. We customized ABNER which is based on MALLET, to suit our needs. ABNER employs a set of orthographic and semantic features.

6.2.1. Feature Selection for CRF Classifier. The default model included the training vocabulary (provided as part of the BIONLP-NLPBA 2004 shared task) in the form of 17 orthographic features based on regular expressions [41]. These include upper case letters (initial upper case letter, all upper case letters, mix of upper and lower case letters), digits (special expressions for single and double digits, natural numbers, and real numbers), hyphen (special expressions for hyphens appearing at the beginning and end of a phrase), other punctuation marks, Roman and Greek words, and 3-gram and 4-gram suffixes and prefixes. ABNER uses semantic features that are provided in the form of hand-prepared (Greek letters, amino acids, chemical elements, known viruses, abbreviations of all these) and database-referenced lexicons (genes, chromosome locations, proteins, and cell lines).

6.3. Evaluation of Sentence Classification Approaches. The framework is designed for large-scale extraction of molecular events from biomedical texts. To assess its performance, we evaluated the underlying components on the GENIA event dataset made available as part of BioNLP'09 Shared Task [26]. This data consists of three different sets: the training set consists of 800 PubMed abstracts (with 7,499 sentences), the development set has 150 abstracts (1,450 sentences), and the test set has 260 abstracts (2,447 sentences). We used the development set for parameter optimization and fine tuning and evaluated the final system on the test set. Employed classifiers were evaluated based on precision and recall. Precision indicates the correctness of the system, by measuring number of samples correctly classified in comparison to the total number of classified sentences. Recall indicates the completeness of the system, by calculating the number of results which actually belong to the expected set of results. Sentence level single label classification and sentence level multilabel classification approaches were evaluated based on how well the classifier labels a given sentence from a test set with one of the nine class labels. Phrase level classification using CRF model was evaluated based on how well the model tags trigger phrases. Evaluating this approach involved measuring the extent to which the model identifies that a phrase is a trigger phrase and how well it classifies a tagged trigger phrase under one of the nine predefined event types. Retrieved trigger phrases refer to the ones which are identified and classified by the CRF sequence tagger. Relevant trigger phrases are the ones which are expected to be tagged by the model. Retrieved and relevant trigger words refer to the tags which are expected to be classified and which are actually classified by the CRF model. All the classifiers are trained using BioNLP shared task training data and tested against BioNLP shared task development abstracts.

We compare the above three approaches for classification in Table 3. CRF has a good tradeoff as compared to Maximum Entropy classifier results. As compared to multiple labels, sentence level classifiers, it performs better in terms of having a considerably good accuracy for most of the event types with a good recall. It not only predicts the event types present in the sentence, but also localizes the trigger phrases. There are some entries where ME seems to perform better than CRF; for example, in case of *positive regulation*, where the precision is as high as 75%. However, in this case, the recall is very low (25%). The reason noticed (in training examples) was that, most of the true example sentences of positive regulation or negative regulation class type were misclassified as either phosphorylation or gene expression. The F1-score for CRF indicates that, as compared to the other approaches, CRF predicts 80% of the relevant tags, and, among these predicted tags, 68% of them are correct.

6.3.1. Evaluation of Phrase Level Labeling. Evaluation of this approach was focused more on the overlap of phrases between the GENIA annotated development and CRF tagged labels. The reason being for each abstract in the GENIA corpus, there is generally a set of biomedical entities present in it. For the shared task, only a subset of these entities was considered in the annotations, and accordingly only events concerning these annotated entities were extracted. However, based on the observation of the corpus, there was a probable chance of other events involving entities not selected for the annotations. So we focused on the coverage, where both the GENIA annotations and CRF annotations agree upon. CRF

Table 3: Summary of classification approaches: test instances (marked events) for each class type in test dataset. Precision, recall, and F1-score in percentage. Compared to NB + EM and CRF, Maximum Entropy based classifier had better average precision, but CRF has best recall and good precision, giving it best F-Measure of the three well-known classifiers.

Event type	Test instances	NB + EM			MaxEnt			CRF		
	Total: 942	P	R	F1	P	R	F1	P	R	F1
Phosphorylation	38	62	42	50	97	73	83	80	83	81
Protein catabolism	17	60	47	53	97	73	83	85	86	85
Gene expression	200	60	41	49	88	58	70	75	81	78
Localization	39	39	47	43	61	69	65	67	79	72
Transcription	60	24	52	33	49	80	61	57	78	66
Binding	153	56	63	59	65	62	63	65	81	72
Regulation	90	47	69	55	52	67	58	62	73	67
Positive regulation	220	70	27	39	75	25	38	55	74	63
Negative regulation	125	42	46	44	54	38	45	68	82	74
Average		51	48	47	**71**	61	63	68	**80**	**73**

Table 4: CRF sequence labeling results.

Type of evaluation	Coverage %
Exact boundary matching	79%
Soft boundary matching	82%

performance was evaluated on two fronts in terms of this overlap.

(i) *Exact boundary matching*: this involves exact label matching and exact trigger phrase match.

(ii) *Soft boundary matching*: this involves exact label matching and partial trigger phrase match, allowing 1-word window on either side of the actual trigger phrase.

Checking of the above constraints was a combination of template matching and manually filtering of abstracts. Table 4 gives an estimate of the coverage. Soft boundary matching increases the coverage by around 3%. Table 3 gives the overall evaluation of CRF with respect to GENIA corpus. With regards to the CRF results, accuracy for *positive regulation* is comparatively low. Also, the test instances for *positive regulation* were more than any other event type. So this reduced average precision to some extent.

A detailed analysis of the results showed that around 3% tags were labeled incorrectly in terms of the event type. There were some cases where it was not certain whether an event should be marked as *regulation* or *positive regulation*. Some examples include "the expression of LAL-mRNA," where "LAL-mRNA" refers to a gene. As per examples seen in the training data, the template of the form "expression of <gene name>" generally indicates presence of a *Gene expression* event. Hence, more analysis may be need to exactly filter out such annotations as true negatives or deliberately induced false positives.

7. Discussion and Conclusions

PubMed is one of the most well known and used citation indexes for the Life Sciences. It provides basic keyword searches and benefits largely from a hierarchically organized set of indexing terms, MeSH, that are semi-automatically assigned to each article. PubMed also enables quick searches for related publications given one or more articles deemed relevant by the user. Some research tools provide additional cross-referencing of entities to databases such as UniProt or to the GeneOntology. They also try to identify relations between entities of the same or different types, such as protein-protein interactions, functional protein annotations, or gene-disease associations. GoPubMed [43] guides users in their everyday searches by mapping articles to concept hierarchies, such as the Gene Ontology and MeSH. For each concept found in abstracts returned by the initial user query, GoPubMed computes a rank based on occurrences of that concept. Thus, users can quickly grasp which terms occur frequently, providing clues for relevant topics and relations, and refine subsequent queries by focusing on particular concepts, discarding others.

In this paper, we presented BioEve Search framework, which can help identify important relationships between entities such as drugs, diseases, and genes by highlights them during the search process. Thereby, allowing the researcher not only to navigate the literature, but also to see entities and the relations they are involved in immediately, without having to fully read the article. Nonetheless, we envision future extensions to provide a more complete and mainstream service and here are few of these next steps.

Keeping the search index up-to-date and complete: we are adding a synchronization module that will frequently check with Medline for supplement articles as they are published; these will typically be in the range of 2500–4500 new articles per day. Frequent synchronization is necessary to keep BioEve abreast with Medline collection and give users the access to the most recent articles.

Normalizing and grounding of entity names: as the same gene/protein can be referred by various names and symbols (e.g., the TRK-fused gene is also known as TF6; TRKT3; FLJ36137; TFG), a user searching for any of these names should find results mentioning any of the others. Removal of duplicates and cleanup of nonbiomedical vocabulary that occurs in the entity tag clouds will further improve navigation and search results.

Cross-referencing with biomedical databases: we want to cross-reference terms indexed with biological databases. For example, each occurrence of a gene could be linked to EntrezGene and OMIM; cell lines can be linked and enriched with ATCC.org's cell line database; we want to cross-reference disease names with UMLS and MeSH to provide access to ontological information. To perform this task of entity normalization, we have previously developed Gnat [6], which handles gene names. Further entity classes that exhibit relatively high term ambiguity with other classes or within themselves are diseases, drugs, species, and GeneOntology terms ("Neurofibromatosis 2" can refer to the disease or gene).

Conflict of Interests

To the authors knowledge, there is no conflict of interest with name "BioEve" or with any trademarks.

Acknowledgments

The authors like to thank Jeorg Hakenberg, Chintan Patel, and Sheela P. Kanwar for valuable discussions, ideas, and help with writing this paper. They also wish to thank the researchers who provided an initial user review and gave them valuable feedback.

References

[1] S. Pyysalo, *A dependency parsing approach to biomedical text mining*, Ph.D. thesis, 2008.

[2] D. Rebholz-Schuhmann, H. Kirsch, M. Arregui, S. Gaudan, M. Riethoven, and P. Stoehr, "EBIMed—text crunching to gather facts for proteins from Medline," *Bioinformatics*, vol. 23, no. 2, pp. e237–e244, 2007.

[3] C. Plake, T. Schiemann, M. Pankalla, J. Hakenberg, and U. Leser, "ALIBABA: pubMed as a graph," *Bioinformatics*, vol. 22, no. 19, pp. 2444–2445, 2006.

[4] U. Leser and J. Hakenberg, "What makes a gene name? Named entity recognition in the biomedical literature," *Briefings in Bioinformatics*, vol. 6, no. 4, pp. 357–369, 2005.

[5] H. Xu, J. W. Fan, G. Hripcsak, E. A. Mendonça, M. Markatou, and C. Friedman, "Gene symbol disambiguation using knowledge-based profiles," *Bioinformatics*, vol. 23, no. 8, pp. 1015–1022, 2007.

[6] J. Hakenberg, C. Plake, R. Leaman, M. Schroeder, and G. Gonzalez, "Inter-species normalization of gene mentions with GNAT," *Bioinformatics*, vol. 24, no. 16, pp. i126–i132, 2008.

[7] K. Oda, J. D. Kim, T. Ohta et al., "New challenges for text mining: mapping between text and manually curated pathways," *BMC Bioinformatics*, vol. 9, supplement 3, article S5, 2008.

[8] M. E. Califf and R. J. Mooney, "Relational learning of pattern-match rules for information extraction," in *Working Notes of AAAI Spring Symposium on Applying Machine Learning to Discourse Processing*, pp. 6–11, AAAI Press, Menlo Park, Calif, USA, 1998.

[9] N. Kushmerick, D. S. Weld, and R. B. Doorenbos, "Wrapper induction for information extraction," in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI '97)*, pp. 729–737, 1997.

[10] L. Schubert, "Can we derive general world knowledge from texts?" in *Proceedings of the 2nd International Conference on Human Language Technology Research*, pp. 94–97, Morgan Kaufmann, San Francisco, Calif, USA, 2002.

[11] M. Friedman and D. S. Weld, "Efficiently executing information-gathering plans," in *Proceedings of the 15th International Joint Conference on Artificial Intelligence (IJCAI '97)*, pp. 785–791, Nagoya, Japan, 1997.

[12] R. Bunescu, R. Ge, R. J. Kate et al., "Comparative experiments on learning information extractors for proteins and their interactions," *Artificial Intelligence in Medicine*, vol. 33, no. 2, pp. 139–155, 2005.

[13] W. Daelemans, S. Buchholz, and J. Veenstra, "Memory-based shallow parsing," in *Proceedings of the Conference on Natural Language Learning (CoNLL '99)*, vol. 99, pp. 53–60, 1999.

[14] E. Brill, "A simple rule-based part-of-speech tagger. In Proceedings of ANLP-92," in *Proceedings of the 3rd Conference on Applied Natural Language Processing*, pp. 152–155, Trento, Italy, 1992.

[15] A. Mikheev and S. Finch, "A workbench for finding structure in texts," in *Proceedings of the Applied Natural Language Processing (ANLP '97)*, Washington, DC, USA, 1997.

[16] M. Craven and J. Kumlien, "Constructing biological knowledge bases by extracting information from text sources," in *Proceedings of the 7th International Conference on Intelligent Systems for Molecular Biology*, pp. 77–86, AAAI Press, 1999.

[17] K. Seymore, A. McCallum, and R. Rosenfeld, "Learning hidden markov model structure for information extraction," in *Proceedings of the AAAI Workshop on Machine Learning for Information Extraction*, 1999.

[18] L. Hunter, Z. Lu, J. Firby et al., "OpenDMAP: an open source, ontology-driven concept analysis engine, with applications to capturing knowledge regarding protein transport, protein interactions and cell-type-specific gene expression," *BMC Bioinformatics*, vol. 9, article 78, 2008.

[19] T. Ono, H. Hishigaki, A. Tanigami, and T. Takagi, "Automated extraction of information on protein-protein interactions from the biological literature," *Bioinformatics*, vol. 17, no. 2, pp. 155–161, 2001.

[20] C. Blaschke, M. A. Andrade, C. Ouzounis, and A. Valencia, "Automatic extraction of biological information from scientific text: protein-protein interactions," AAAI, pp. 60–67.

[21] C. Friedman, P. Kra, H. Yu, M. Krauthammer, and A. Rzhetsky, "GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles," *Bioinformatics*, vol. 17, no. 1, pp. S74–S82, 2001.

[22] A. Rzhetsky, I. Iossifov, T. Koike et al., "GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data," *Journal of Biomedical Informatics*, vol. 37, no. 1, pp. 43–53, 2004.

[23] D. P. A. Corney, B. F. Buxton, W. B. Langdon, and D. T. Jones, "BioRAT: extracting biological information from full-length papers," *Bioinformatics*, vol. 20, no. 17, pp. 3206–3213, 2004.

[24] G. Leroy, H. Chen, and J. D. Martinez, "A shallow parser based on closed-class words to capture relations in biomedical text," *Journal of Biomedical Informatics*, vol. 36, no. 3, pp. 145–158, 2003.

[25] M. Krallinger, F. Leitner, C. Rodriguez-Penagos, and A. Valencia, "Overview of the protein-protein interaction annotation extraction task of BioCreative II," *Genome Biology*, vol. 9, no. 2, article S4, 2008.

[26] J. D. Kim, T. Ohta, S. Pyysalo, Y. Kano, and J. Tsujii, "Overview of BioNLP'09 shared task on event extraction," in *Proceedings of the Workshop Companion Volume for Shared Task (BioNLP '09)*, pp. 1–9, Association for Computational Linguistics, Boulder, Colo, USA, 2009.

[27] A. Yakushiji, Y. Tateisi, Y. Miyao, and J. Tsujii, "Event extraction from biomedical papers using a full parser," *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pp. 408–419, 2001.

[28] J. Ding, D. Berleant, J. Xu, and A. W. Fulmer, "Extracting Biochemical Interactions from MEDLINE Using a Link Grammar Parser," in *Proceedings of the 15th IEEE International Conference on Tools with Artificial Intelligence*, pp. 467–471, November 2003.

[29] H. Cunningham, *Information Extraction, Automatic*, Encyclopedia of Language and Linguistics, 2nd edition, 2005.

[30] O. Etzioni, M. Cafarella, D. Downey et al., "Methods for domain-independent information extraction from the web: an experimental comparison," in *Proceedings of the 19th National Conference on Artificial Intelligence (AAAI '04)*, pp. 391–398, AAAI Press, Menlo Park, Calif, USA, July 2004.

[31] M. Cafarella, D. Downey, S. Soderland, and O. Etzioni, "KnowItNow: fast, scalable information extraction from the web," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 563–570, Association for Computational Linguistics, Morristown, NJ, USA, 2005.

[32] O. Etzioni, M. Banko, S. Soderland, and D. S. Weld, "Open information extraction from the web," *Communications of the ACM*, vol. 51, no. 12, pp. 68–74, 2008.

[33] M. Cafarella and O. Etzioni, "A search engine for natural language applications," in *Proceedings of the International Conference on World Wide Web (WWW '05)*, pp. 442–452, ACM, New York, NY, USA, 2005.

[34] R. White, B. Kules, and S. Drucker, "Supporting exploratory search, introduction, special issue, communications of the ACM," *Communications of the ACM*, vol. 49, no. 4, pp. 36–39, 2006.

[35] W. T. Fu, T. G. Kannampallil, and R. Kang, "Facilitating exploratory search by model-based navigational cues," in *Proceedings of the 14th ACM International Conference on Intelligent User Interfaces (IUI '10)*, pp. 199–208, ACM, New York, NY, USA, February 2010.

[36] J. Koren, Y. Zhang, and X. Liu, "Personalized interactive faceted search," in *Proceedings of the 17th International Conference on World Wide Web (WWW '08)*, pp. 477–485, ACM, April 2008.

[37] V. Sinha and D. R. Karger, "Magnet: supporting navigation in semistructured data environments," in *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD '05)*, pp. 97–106, ACM, June 2005.

[38] M. Hearst, "Design recommendations for hierarchical faceted search interfaces," in *Proceedings of the ACM Workshop on Faceted Search (SIGIR '06)*, 2006.

[39] S. Stamou and L. Kozanidis, "Towards faceted search for named entity queries," *Advances in Web and Network Technologies, and Information Management*, vol. 5731, pp. 100–112, 2009.

[40] D. Tunkelang, *Faceted Search*, Morgan & Claypool, 2009.

[41] B. Settles, "ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text," *Bioinformatics*, vol. 21, no. 14, pp. 3191–3192, 2005.

[42] J. Lafferty and F. Pereira, "Conditional random fields: probabilistic models for segmenting and labeling sequence data," in *Proceedings of the 18th International Conference on Machine Learning (ICML '01)*, 2001.

[43] A. Doms and M. Schroeder, "GoPubMed: exploring PubMed with the gene ontology," *Nucleic Acids Research*, vol. 33, no. 2, pp. W783–W786, 2005.

Hindawi Publishing Corporation
Advances in Bioinformatics
Volume 2012, Article ID 391574, 17 pages
doi:10.1155/2012/391574

Review Article

Applications of Natural Language Processing in Biodiversity Science

Anne E. Thessen,[1] Hong Cui,[2] and Dmitry Mozzherin[1]

[1] *Center for Library and Informatics, Marine Biological Laboratory, 7 MBL Street, Woods Hole, MA 02543, USA*
[2] *School of Information Resources and Library Science, University of Arizona, Tucson, AZ 85719, USA*

Correspondence should be addressed to Anne E. Thessen, athessen@mbl.edu

Received 4 November 2011; Accepted 15 February 2012

Academic Editor: Jörg Hakenberg

Centuries of biological knowledge are contained in the massive body of scientific literature, written for human-readability but too big for any one person to consume. Large-scale mining of information from the literature is necessary if biology is to transform into a data-driven science. A computer can handle the volume but cannot make sense of the language. This paper reviews and discusses the use of natural language processing (NLP) and machine-learning algorithms to extract information from systematic literature. NLP algorithms have been used for decades, but require special development for application in the biological realm due to the special nature of the language. Many tools exist for biological information extraction (cellular processes, taxonomic names, and morphological characters), but none have been applied life wide and most still require testing and development. Progress has been made in developing algorithms for automated annotation of taxonomic text, identification of taxonomic names in text, and extraction of morphological character information from taxonomic descriptions. This manuscript will briefly discuss the key steps in applying information extraction tools to enhance biodiversity science.

1. Introduction

Biologists are expected to answer large-scale questions that address processes occurring across broad spatial and temporal scales, such as the effects of climate change on species [1, 2]. This motivates the development of a new type of data-driven discovery focusing on scientific insights and hypothesis generation through the novel management and analysis of preexisting data [3, 4]. Data-driven discovery presumes that a large, virtual pool of data will emerge across a wide spectrum of the life sciences, matching that already in place for the molecular sciences. It is argued that the availability of such a pool will allow biodiversity science to join the other "Big" (i.e., data-centric) sciences such as astronomy and high-energy particle physics [5]. Managing large amounts of heterogeneous data for this Big New Biology will require a cyberinfrastructure that organizes an open pool of biological data [6].

To assess the resources needed to establish the cyberinfrastructure for biology, it is necessary to understand the nature of biological data [4]. To become a part of the cyberinfrastructure, data must be ready to enter a digital data pool. This means data must be digital, normalized, and standardized [4]. Biological data sets are heterogeneous in format, size, degree of digitization, and openness [4, 7, 8]. The distribution of data packages in biology can be represented as a hollow curve [7] (Figure 1). To the left of the curve are the few providers producing large amounts of data, often derived from instruments and born digital such as in molecular biology. To the right of the curve are the many providers producing small amounts of data. It is estimated that 80% of scientific output comes from these small providers [7]. Generally called "small science," these data are rarely preserved [9, 10]. Scientific publication, a narrative explanation derived from primary data, is often the only lasting record of this work.

The complete body of research literature is a major container for much of our knowledge about the natural world and represents centuries of investment. The value of this information is high as it reflects observations that are difficult to replace if they are replaceable at all [7]. Much of the information has high relevance today, such as records on

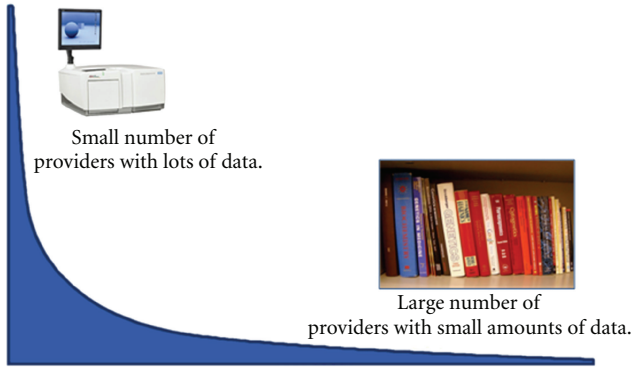


Figure 1: The long tail of biology. Data quantity, digitization, and openness can be described using a hyperbolic (hollow) curve with a small number of providers providing large quantities of data, and a large number of individuals providing small quantities of data.

the historical occurrence of species that will help us better understand shifting abundances and distributions. Similarly, taxonomy, with its need to respect all nomenclatural acts back to the 1750s, needs to have access to information contained exclusively within this body of literature. Unfortunately, this knowledge has been presented in the narrative prose such that careful reading and annotation are required to make use of any information [11] and only a subset has been migrated into digital form.

The number of pages of the historical biodiversity literature is estimated to be approximately hundreds of millions [12]. Currently, over 33 million pages of legacy biology text are scanned and made available online through the Biodiversity Heritage Library (http://www.biodiversitylibrary.org/) and thousands of new digital pages are published every month in open-access biology journals (estimated based on 216 journals publishing approx 10 articles per month of less than 10 pages; http://www.doaj.org/doaj?cpid=67&func=subject). Additional biologically focused digital literature repositories can be found here (http://www.library.illinois.edu/nhx/resources/digitalresourcecatalogs.html).

The information is in human-readable form but is too much for a person to transform into a digital data pool. Machines can better handle the volume, but cannot determine which elements of the text have value. In order to mobilize the valuable content in the literature, we need innovative algorithms to translate the entirety of the biological literature into a machine-readable form, extract the information with value, and feed it in a standards-compliant form into an open data pool. This paper discusses the application of natural language processing algorithms to biodiversity science to enable data-driven discovery.

2. Overview

2.1. Information Extraction. Research addressing the transformation of natural language text into a digital data pool is generally labeled as "information extraction" (IE). An IE task typically involves a corpus of source text documents to be acted upon by the IE algorithm and an extraction template that describes what will be extracted. For a plant character IE task, (e.g., [13]), a template may consist of taxon name, leaf shape, leaf size, leaf arrangement, and so forth (Table 1). The characteristics of the source documents and the complexity of the template determine the difficulty level of an IE task. More complex IE tasks are often broken down to a series (stacks) of sub tasks, with a later subtask often relying on the success of an earlier one. Table 2 illustrates typical subtasks involved in an IE task. Note, not all IE tasks involve all of these subtasks. Examples of information extraction tools for biology (not including biodiversity science) can be found in Table 3.

The IE field has made rapid progress since the 1980s with the Message Understanding Conferences (MUCs) and has become very active since the 1990s due largely to the development of the World Wide Web. This has made available huge amounts of textual documents and human-prepared datasets (e.g., categorized web pages, databases) in an electronic format. Both can readily be used to evaluate the performance of an IE system. The massive production of digital information demands more efficient, computer-aided approaches to process, organize, and access the information. The urgent need to extract interesting information from large amounts of text to support knowledge discovery was recognized as an application for IE tools (e.g., identifying possible terrorists or terrorism attacks by extracting information from a large amount of email messages). For this reason,

Table 1: From Tang and Heidorn [13]. An example template for morphological character extraction.

Template slots	Extracted information
Genus	Pellaea
Species	mucronata
Distribution	Nev. Calif.
Leaf shape	ovate-deltate
Leaf margin	dentate
Leaf apex	mucronate
Leaf base	
Leaf arrangement	clustered
Blade dimension	
Leaf color	
Fruit/nut shape	

Table 2: Information extraction tasks outlined by the MUCs and their descriptions.

Task	Description
Named entity	Extracts names of entities
Coreference	Links references to the same entity
Template element	Extracts descriptors of entities
Template rotation	Extracts relationships between entities
Scenario template	Extracts events

Table 3: Existing IE systems for biology [17–26].

System	Approach	Structure of Text	Knowledge in	Application domain	Reference
AkanePPI	shallow parsing	sentence-split, tokenized, and annotated		protein interactions	[17]
EMPathIE	pattern matching	text	EMP database	enzymes	[18]
PASTA	pattern matching	text	biological lexicons	protein structure	[19]
BioIE	pattern matching	xml	dictionary of terms	biomedicine	[20]
BioRAT	pattern matching, sub-language driven	could be xml, html, text or asn.1, can do full-length pdf papers (converts to text)	dictionary for protein and gene names, dictionary for interactions, and synonyms; text pattern template	biomedicine	[21]
Chilibot	shallow parsing	not sure what was used in paper, but could be xml, html, text or asn.1	nomenclature dictionary	biomedicine	[22]
Dragon Toolkit	mixed syntactic semantic	text	domain ontologies	genomics	[23]
EBIMed	pattern matching	xml	dictionary of terms	biomedicine	[24]
iProLINK	shallow parsing	text	protein name dictionary, ontology, and annotated corpora	proteins	[25]
LitMiner	mixed syntactic semantic	web documents		Drosophila research	[26]

IE and other related research have acquired another, more general label "text data mining" (or simply "text mining").

Information extraction algorithms are regularly evaluated based on three metrics: recall, precision, and the F score. Consider an algorithm trained to extract names of species from documents being run against a document containing the words: cat, dog, chicken, horse, goat, and cow. The recall would be the ratio of the number of "species words" extracted to the number in the document (6). So, an algorithm that only recognized cat and dog would have low recall (33%). Precision is the percentage of what the algorithm extracts that is correct. Since both cat and dog are species words, the precision of our algorithm would be 100% despite having a low recall. If the algorithm extracted all of the species words from the document, it would have both high precision and recall, but if it also extracts other words that are not species, then it would have low precision and high recall. The F score is an overall metric calculated from precision and recall when precision and recall are considered equally important:

$$F \text{ score} = 2\left(\frac{(\text{precision} * \text{recall})}{(\text{precision} + \text{recall})}\right). \quad (1)$$

Before we review current IE systems for biodiversity science, we will first present a reference system architecture for a model IE system that covers the entire process of an IE application (Figure 2). In reviewing variant systems, we will refer to this reference architecture.

The blue-shaded areas in Figure 2 illustrate an IE system. The inputs to the IE system include source documents in a digital format (element number 1 in Figure 2), an IE template which describes the IE task (2) and knowledge entities to perform the task (3). If documents are not in a digital format, OCR technologies can be used to make the transition (4; see below section on digitization), but then it is necessary to correct OCR errors before use (5). In this model system, we use "IE template" to refer not only to those that are well defined such as the leaf character template example in Table 1, but also those more loosely defined. For example, we also consider lists of names and characters to be IE templates so the reference system can cover Named Entity Recognition systems (see below for examples) and character annotation systems (see below for examples). Knowledge entities include, for example, dictionaries, glossaries, gazetteers, or ontologies (3). The output of an IE system is often data in a structured format, illustrated as a database in the diagram (6). Ideally the structured format conforms to one of many data standards (7), which can range from relational database schemas to RDF. The arrow from Knowledge Entities to Extracted Data illustrates that, in some cases, the extracted data can be better interpreted with the support of knowledge entities (like annotation projects such as phenoscape, http://phenoscape.org/wiki/Main_Page). The arrow from Data Standards to Extracted Data suggests the same.

NLP techniques are often used in combination with extraction methods (including hand-crafted rules and/or

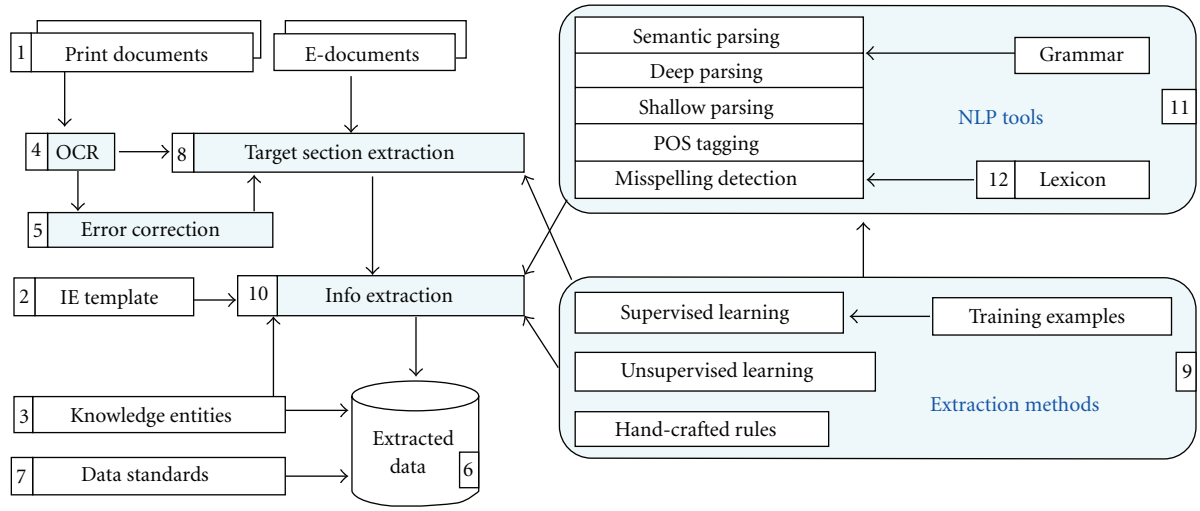


Figure 2: A reference system architecture for an example IE system. Numbers correspond to the text.

machine learning methods). Often the input documents contain text that are not relevant to an IE task [14]. In these cases, the blocks of text that contain extraction targets need to be identified and extracted first to avoid the waste of computational resources (8). An IE method is often used for this purpose (9). From the selected text, a series of tasks may be performed to extract target information (10) and produce final output (6; see also IE subtasks in Table 2). This is often accomplished by first applying NLP techniques (11) and then using one or a combination of extraction methods (9). The arrow from extraction methods to NLP tools in Figure 2 indicates that machine learning and hand-crafted rules can be used to adapt/improve NLP tools for an IE task by, for example, extracting domain terms to extend the lexicon (12) used by a syntactic parser or even create a special purpose parser [15]. One important element that is not included in the model (Figure 2) is the human curation component. This is important for expert confirmation that extraction results are correct.

2.2. Natural Language Processing. IE is an area of application of natural language processing (NLP). NLP enables a computer to read (and possibly "understand") information from natural language texts such as publications. NLP consists of a stack of techniques of increasing sophistication to progressively interpret language, starting with words, progressing to sentence structure (syntax or syntactic parsing), and ending at sentence meaning (semantics or semantic parsing) and meaning within sequences of sentences (discourse analysis). Typically an NLP technique higher in the stack (discourse analysis) utilizes the techniques below it (syntactic parsing). A variety of NLP techniques have been used in IE applications, but most only progress to syntactic parsing (some special IE applications specifically mentioned in this paper may not use any of the techniques). More sophisticated techniques higher in the stack (semantic parsing and discourse analysis) are rarely used in IE applications because they are highly specialized that is, cannot be reliably applied in general applications and are more computationally expensive.

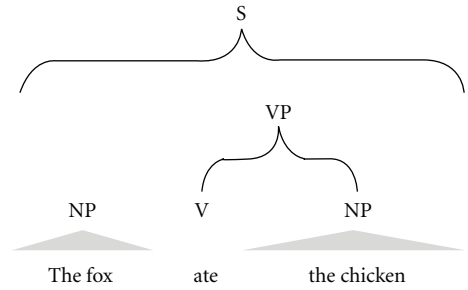


Figure 3: An example of shallow parsing. Words and a sentence (S) are recognized. Then, the sentence is parsed into noun phrases (NP), verbs (V), and verb phrases (VP).

Syntactic parsing can be shallow or deep. Shallow syntactic parsing (also called "chunking") typically identifies noun, verb, preposition phrases, and so forth in a sentence (Figure 3), while deep syntactic parsing produces full parse trees, in which the syntactic function (e.g., Part of Speech, or POS) of each word or phrase is tagged with a short label (Figure 4). The most commonly used set of POS tags used is the Penn Treebank Tag Set (http://bulba.sdsu.edu/jeanette/thesis/PennTags.html), which has labels for different parts of speech such as adjective phrases (ADJP), plural nouns (NNP), and so forth. Not all shallow parsers identify the same set of phrases. GENIA Tagger, for example, identifies adjective phrases (ADJP), adverb phrases (ADVP), conjunctive phrases (CONJP), interjections (INTJ), list markers (LST), noun phrases (NP), prepositional phrases (PP), participles (PRT), subordinate clauses (SBAR), and verb phrases (VP). Some shallow parsing tools are the Illinois Shallow Parser (http://cogcomp.cs.illinois.edu/page/software_view/13) the Apache OpenNLP (http://incubator.apache.org/opennlp/index.html), and GENIA Tagger

Original text:

leaf blade obovate to nearly orbiculate, 3–9 × 3–8 cm, leathery, base obtuse.

Shallow parsing result :

[NP leaf blade] [VP obovate] [VP to nearly orbiculate], [NP 3–9 × 3–8 cm], [NP leathery], [NP base obtuse].

Deep parsing result:

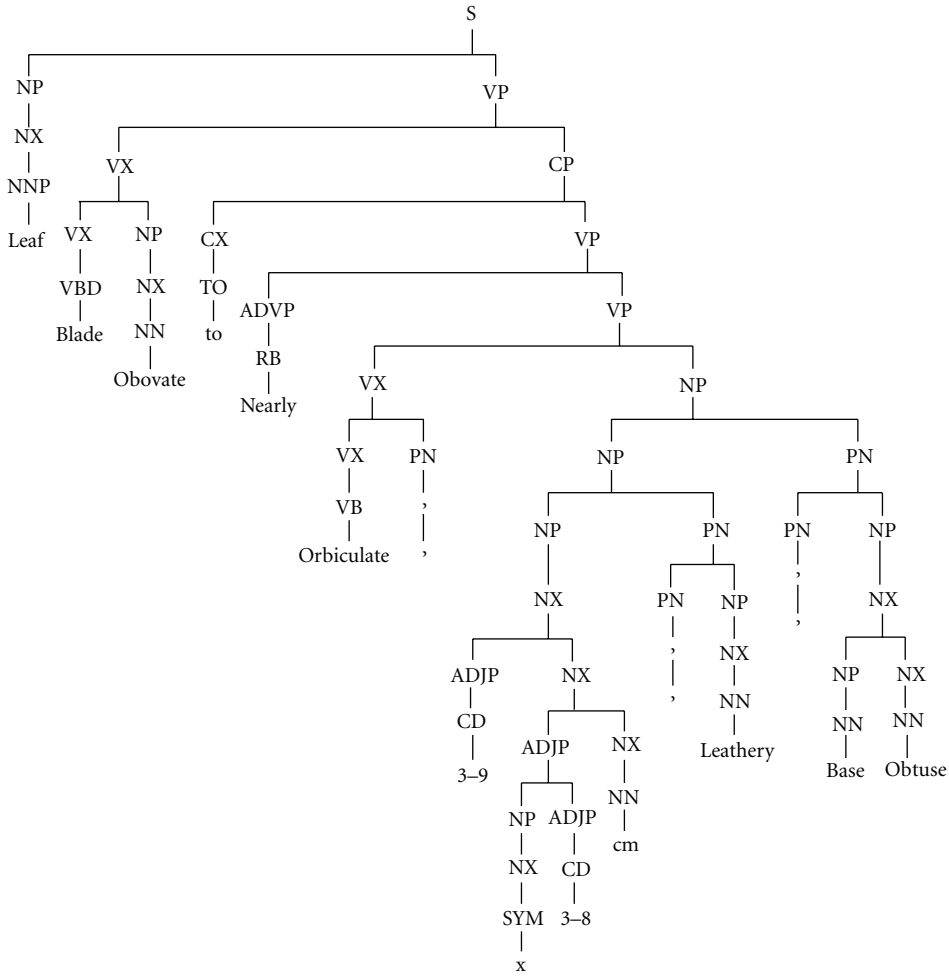


Figure 4: Shallow-vs-Deep-Parsing. The shallow parsing result produced by GENIA Tagger (http://text0.mib.man.ac.uk/software/geniatagger/). The deep parsing result produced by Enju Parser for Biomedical Domain (http://www-tsujii.is.s.u-tokyo.ac.jp/enju/demo.html). GENIA Tagger and Enju Parser are products of the Tsujii Laboratory of the University of Tokyo and optimized for biomedical domain. Both Parsing results contain errors, for example "obovate" should be an ADJP (adjective phrase), but GENIA Tagger chunked it as a VP (verb phrase). "blade" is a noun, but Enju parser parsed it as a verb (VBD). This is not to criticize the tools, but to point out language differences in different domains could have a significant impact on the performance of NLP tools. Parsers trained for a general domain produce erroneous results on morphological descriptions [16].

(http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger/). Deep parsing tools include Stanford Parser (http://nlp.stanford.edu/software/lex-parser.shtml), Link Parser (http://www.link.cs.cmu.edu/link/) and Enju Parser (http://www-tsujii.is.s.u-tokyo.ac.jp/enju/). A majority of IE applications use the shallow parsing technique, but the use of deep parsing techniques is on the rise in biology applications. This is driven in part because shallow parsing is not adequate to extract information from biology text [27–29].

Several NLP approaches are available for IE applications in biology that go beyond shallow parsing and are not mutually exclusive.

(1) *Pattern matching* approaches exploit basic patterns in text to extract information. An example pattern is "enzyme activates protein" or X activates Y. The computer would look for the specific text pattern and assume that all X are enzymes or all Y are

proteins. Dictionary-based IE is a variant of pattern matching that focuses on finding words in text that are contained in a dictionary previously given to the computer. For example, the computer might be given a list of enzyme names (such as the UM-BBD list of enzyme names, http://umbbd.msi.umn.edu/servlets/pageservlet?ptype; X in previous example). Once the enzyme name is located, the computer can infer the pattern that it "activates Y." Another variant of pattern matching is the preposition-based parsing which focuses on finding prepositions like "by" and "of" and filling a basic template with information surrounding that preposition. An example of this would be "Y is activated by X." Pattern matching suffers from the difficulty in accounting for the wide array of linguistic patterns used in text (X activates Y, Y is activated by X, Y was activated by X, X activated Y, Y is activated via X, X, which activates Y, etc.). Many of these systems extract phrases or sentences instead of structured facts, which limits their usefulness for further informatics. An example system that uses pattern matching is given in Krauthammer et al. [37].

(2) *Full parsing* approaches expand on shallow parsing to include an analysis of sentence structure (i.e., syntax, see Figure 3). The biggest challenge with this approach is the special language of biology-specific texts. Most existing full-parsing systems are designed to handle general language texts, like news articles. The approach is also limited by grammar mistakes in the literature, which are often due to nonnative English speakers. Full parsing often runs into ambiguity due to the many ways a sentence (even moderately complex) can be interpreted by a machine. Sentence fragments, such as titles or captions, can also cause problems. UniGene Tabulator is an example of a full parser for biology [38].

(3) *Probability-based* approaches offer a solution to the linguistic variability that confounds full parsing. These approaches use weighted grammar rules to decrease sensitivity to variation. The weights are assigned through processing of a large body of manually annotated text. Probabilistic grammars are used to estimate the probability that a particular parse tree will be correct or the probability that a sentence or sentence fragment has been recognized correctly. Results can be ranked according to the probabilities. Nasr and Rambow give an example of a probability-based parser [39].

(4) *Mixed syntactic-semantic* approaches take advantage of syntactic and semantic knowledge together. This essentially combines part-of-speech taggers with named-entity recognition, such as in the BANNER system [40]. This removes reliance on lexicons and templates. This approach will be discussed further below.

(5) *Sub language-driven* approaches use the specialized language of a specific community. A specialized sub language typically has a set of constraints that determine vocabulary, composition, and syntax that can be translated into a set of rules for an algorithm. Algorithms for use in processing biology text must cope with specialized language and the telegraphic sentence structure found in many taxonomic works. Being unaware of a sub language will often lead to incorrect assumptions about use of the language. Metexa is an example of a tool that uses a specialized sub language in the radiology domain [41].

NLP techniques are often used as a (standard) initial text processing procedure in an IE application. Once a computer has an understanding of the syntactic and/or semantic meaning of the text, other methods, such as manually derived rules or machine learning based methods, are then often used for further information extraction.

2.3. Machine Learning. Machine learning has been used in IE applications since 1990s. It is a process by which a machine (i.e., computer algorithm) improves its performance automatically with experience [42]. Creating extraction rules automatically by machine learning are favored over creating them manually because the hand-crafted rules take longer to create and this time accumulates for each new document collection [43]. As a generic method, machine-learning applications may be found in all aspects of an IE system, ranging from learning lexicons for a syntactic parser, classifying and relating potential extraction targets, to fitting extracted entities into an extraction template.

Learning can take various forms including rule sets, decision trees, clustering algorithms, linear models, Bayesian networks, artificial neural networks, and genetic algorithms (which are capable of mimicking chromosome mutations). Some machine-learning algorithms (e.g., most classification algorithms such as decision trees, naïve Baysian, Support Vector Machines) rely on substantial "training" before they can perform a task independently. These algorithms fall in the category of "supervised machine learning." Some other algorithms (e.g., most clustering algorithms) require little or no training at all, so they belong to the "unsupervised machine learning" category. Due to the considerable cost associated with preparing training examples, one research theme in machine learning is to investigate innovative ways to reduce the amount of training examples required by supervised learning algorithms to achieve the desired level of performance. This gave rise to a third category of machine learning algorithms, "semisupervised." Co-training is one of the learning approaches that falls into this category. Co-training refers to two algorithms that are applied to the same task, but learn about that task in two different ways. For example, an algorithm can learn about the contents of a web site by (1) reading the text of the web site or (2) reading the text of the links to the web site. The two bodies of text are different, but refer to the same thing (i.e., the web site). Two different algorithms can be used to learn about the web site, feed each other machine-made training examples

ORIGINAL TEXT
Leaf blade orbiculate, 3–9 × 3–8 cm, leathery, base obtuse
EXPECTED EXTRACTION RESULT
leaf {**leafShape: orbiculate**}
leaf {**bladeDimension: 3–9 × 3–8 cm**}

Box 1

(which reduces the requirements of human-made training examples), and often make each other better. However, co-training requires two independent views of the same learning task and two independent learners. Not all learning tasks fulfill these requirements. One line of research in NLP that uses co-training is word sense disambiguation [44]. We are not aware of the use of this learning approach in biodiversity information extraction. The best learning algorithm for a certain task is determined by the nature of the task and characteristics of source data/document collection, so it is not always possible to design an unsupervised or semisupervised algorithm for a learning task (i.e., an unsupervised algorithm to recognize human handwriting may not be possible).

The form of training examples required by a supervised algorithm is determined largely by the learning task and the algorithm used. For example, in Tang and Heidorn [13], the algorithm was to learn (automatically generate) rules to extract leaf properties from plant descriptions. A training example used in their research as well as the manually derived, correct extraction is in Box 1 (examples slightly modified for human readability).

By comparing original text (italics) and the text in bold, the algorithm can derive a set of candidate extraction rules based on context. The algorithm would also decide the order that the extraction rules may be applied according to the rules' reliability as measured with training examples. The more reliable rules would be utilized first. Two extraction rules generated by the Tang and Heidorn [13] algorithm are shown in Box 2. Rule 1 extracts from the original text any leaf shape term (represented by <leafShape>) following a term describing leaf blade (represented by <PartBlade>) and followed by a comma (,) as the leafShape (represented by the placeholder $1). Rule 2 extracts any expression consisting of a range and length unit (represented by <Range><LengUnit>) that follows a comma (,) and is followed by another comma (,) and a leaf base term (represented by <PartBase>) as the bladeDimension.

These rules can then be used to extract information from new sentences not included in the original training example. Box 3 shows how the rules match a new statement, and are applied to extract new leafShape and bladeDimension values.

This example illustrates a case where words are the basic unit of processing and the task is to classify words by using the context where they appear (*obovate* is identified as a leaf shape because it follows the phrase "leaf blade").

In some applications, for example, named entity recognition (e.g., recognizing a word/phrase as a taxon name), an extraction target may appear in any context (e.g., a taxon name may be mentioned anywhere in a document). In these applications, the contextual information is less helpful in classifying a word/phrase than the letter combinations within the names. In NetiNeti, for example, a Naïve Baysian algorithm (a supervised learning algorithm based on Bayes conditional probability theorem) uses letter combinations to identify candidate taxon names [32]. When several training examples indicate names like *Turdus migratorius* are taxon names, NetiNeti may learn that a two-word phrase with the first letter capitalized and the last word ending with "*us*" (e.g., *Felis catus*) is probably a taxon name, even though *Felis catus* has never appeared in training examples.

Supervised learning algorithms can be more difficult to use in biology largely because compiling large training datasets can be labor intensive, which decreases the adaptability and scalability of an algorithm to new document collections. Hundreds of controlled vocabularies exist for biological sciences, which can provide some training information to an algorithm but are often not comprehensive [16].

Unsupervised learning algorithms do not use training examples. These algorithms try to find hidden structure in unlabeled data using characteristics of the text itself. Well-known unsupervised learning algorithms include clustering algorithms, dimensionality reduction, and self-organization maps, to name a few. Cui et al. Boufford [14] designed an unsupervised algorithm to identify organ names and organ properties from morphological description sentences. The algorithm took advantage of a recurring pattern in which plural nouns that start a sentence are organs and a descriptive sentence starts with an organ name followed by a series of property descriptors. These characteristics of descriptive sentences allow an unsupervised algorithm to discover organ names and properties.

The procedure may be illustrated by using a set of five descriptive statements taken from Flora of North America (Box 4).

Because *roots* is a plural noun and starts statement 1 (in addition, the words *rooting* or *rooted* are not seen in the entire document collection, so *roots* is unlikely a verb) the algorithm infers *roots* is an organ name. Then, what follows it (i.e., *yellow*) must be a property. The algorithm remembers *yellow* is a property when it encounters statement 2 and it then infers that *petals* is an organ. Similarly, when it reads statement 3, because *petals* is an organ, the algorithm infers *absent* is a property, which enables the algorithm to further infer *subtending bracts* and *abaxial hastular* in statements 4 and 5 are organs. This example shows that by utilizing the description characteristics, the algorithm is able to learn that *roots*, *petals*, *subtending bracts*, and *abaxial hastular* are organ names and *yellow* and *absent* are properties, without using any training examples, dictionaries, or ontologies.

Because not all text possesses the characteristics required by the algorithm developed by Cui et al. [14], it cannot be directly applied to all taxon descriptions. However, because descriptions with those characteristics do exist in large numbers and because of the low overhead (in terms of preparing training examples) of the unsupervised learning algorithm, it is argued that unsupervised learning should be

Rule 1: Pattern:: ∗ <PartBlade> (<leafShape>), ∗
Output:: leaf{leafShape: $1}
Rule 2: Pattern::∗ <PartBlade> ∗, (<Range><LengUnit>), <PartBase>
Output:: leaf{bladeDimension: $1}

Box 2

NEW TEXT
Leaf blade obovate, 1–3 × 1-2 cm, base rounded
Rule 1: Output:: leaf{leafShape: obovate}
Rule 2: Output:: leaf{bladeDimension: 1–3 × 1-2 cm}

Box 3

1. **roots** *yellow* to medium brown or black, thin.
2. **petals** *yellow* or white
3. **petals** *absent*
4. **subtendingbracts** *absent*
5. **abaxialhastular** *absent*

Box 4

exploited when possible, such as when preparing text for a supervised learning task [16].

3. Review of Biodiversity Information Extraction Systems

Our review describes the features of each system existing at the time of this writing. Many of the systems are being constantly developed with new features and enhanced capabilities. We encourage the readers to keep track of the development of these systems.

3.1. Digitization. The first step to making older biological literature machine readable is digitization (number 4 in Figure 2). Book pages can be scanned as images of text and made into pdf files, but cannot be submitted to NLP processing in this form. To make the text accessible, it must be OCRed (Optical Character Recognition) to translate the image of text (such as .pdf) into actual text (such as .txt). The Biodiversity Heritage Library is in the process of digitizing 600,000 pages of legacy text a month, making them available as pdf image files and OCR text files [45]. Most modern publications are available as pdf and html files from the publisher (and thus do not need to be scanned or OCRed). Images of text can be run through software designed to OCR files on desktop computers or as a web service (i.e., http://www.onlineocr.net/). OCR of handwriting is very different from that of text and can be quite difficult as there are as many handwriting styles as there are people. However, this type of OCR can be very important because significant portions of biodiversity data are only available as handwriting, such as museum specimen labels and laboratory notebooks. Algorithms do exist and are used for OCR of handwritten cities, states, and zip codes on envelopes and handwritten checks [46, 47].

OCR is not a perfect technology. It is estimated that >35% of taxon names in BHL OCR files contain an error [45, 48, 49]. This is skewed, however, as older documents that use nonstandard fonts carry the majority of the errors [49]. Biodiversity literature can be especially difficult to OCR as they often have multiple languages on the same page (such as Latin descriptions), an expansive historical record going back to the 15th Century (print quality and consistency issues), and an irregular typeface or typesetting [48]. OCR is poor at distinguishing indentation patterns, bold, and italicized text, which can be important in biodiversity literature [50, 51]. The current rate of digitization prohibits manual correction of these errors. Proposed solutions include components of crowd-sourcing manual corrections and machine-learning for automated corrections [48].

OCR errors may be overcome by using "fuzzy" matching algorithms that can recognize the correct term from the misspelled version. TAXAMATCH is a fuzzy matching algorithm for use in taxonomy [52]. The need for a "fuzzy matching" algorithm for detection of similar names is apparent for functions such as search, federation of content, and correction of misspellings or OCR errors. TAXAMATCH is a tool that uses phonetic- and nonphonetic-based near-match algorithms that calculate the distance of the given letter combination to a target name included in a reference database [52]. A letter combination with a close proximity to a target name is proposed as a fuzzy match. This system is being successfully used to increase hits in species databases [52] and is optimized for human typos rather than OCR errors. The php version of this code is available through Google code (http://code.google.com/p/taxamatch-webservice/) and a Ruby version is available through git hub (https://github.com/GlobalNamesArchitecture/taxamatch_rb).

3.2. Annotation. Once text has been digitized, it can be annotated in preparation for an IE task or for use as training data for algorithms (Figure 2 number 8). Both aims require different levels of annotation granularity, which can be accomplished manually or automatically using annotation software. A low level of granularity (coarse) is helpful for identifying blocks of text useful for IE. As mentioned before, not all text is useful for every IE task. In the practice of systematics, taxonomists need text containing nomenclatural acts which may be discovered and annotated automatically through terms such as "sp. nov." and "nov. comb." Annotation of these text blocks is helpful for algorithms designed to extract information about species. A finer granularity

is needed for training data annotation. Words or phrases within a taxonomic work may be annotated as a name, description, location, and so forth. High granularity is more helpful for training a machine-learning algorithm but imposes a larger cost in time needed to do the manual annotation. There must be a balance between level of granularity and amount of manual investment which is determined by the specific goals at hand.

Manual annotation is very time consuming but can be assisted with annotation software. Several software packages aid with this.

taXMLit. TaXMLit is an interface to allow annotation of taxonomic literature [51]. It was developed using botanical and zoological text, but also works well on paleontological text. This system is designed for annotation of text elements such as "description" and "locality." This system requires a fairly large amount of human intervention and is not widely accepted.

GoldenGATE. GoldenGATE is an annotation tool for marking up taxonomic text in XML according to taxonX schema (http://plazi.org/files/GoldenGATE_V2_end_user_manual.pdf, [53]. Most of the annotation is done semi-automatically, with users checking the correctness of the annotations in the GoldenGATE editor that facilitates manual XML mark up. There are several plugins available for GoldenGATE, including modules for annotation specific to ZooTaxa and new plugins can be relatively easily added. The system is implemented in JAVA. This system performs best with text marked up with basic html tags (such as paragraph and header) and high-quality OCR.

ABNER. ABNER is an entity recognition algorithm designed specifically for the biomedical literature [54]. It uses a conditional random fields (CRF) model. This is a type of Bayesian statistics, wherein the computer uses characteristics of the text to determine the probability that a given term should be annotated as a given class. In this case, the available classes are: protein, DNA, RNA, Cell line, and Cell type. A human uses a point-and-click interface to confirm the algorithm results and add the annotation.

OnTheFly. OnTheFly is a text annotator that automatically finds and labels names of proteins, genes, and other small molecules in Microsoft Office, pdf, and text documents [55]. A user submits a file through the interface and it converts the file of interest into html and sends it to the Reflect tool. This tool looks for names and synonyms of proteins and small molecules to annotate as such [56]. It uses a list of 5.8 million molecule names from 373 organisms and returns matching terms. Clicking on an annotated term returns a pop-up window with additional information. In addition, this tool can create a graphical representation of the relationships between these entities using the STITCH database [57].

Phenex. Phenex was designed for use in the phenotypic literature [58]. It is a user interface that aids in manual annotation of biological text using terms from existing ontologies. Phenex allows users to annotate free text or NEXUS files. A core function of this software is to allow users to construct EQ (Entity:Quality) statements representing phenotypes. An EQ statement consists of two parts, a character (entity) and state (quality). The character is described using a term from an anatomy ontology and the state of that character is described using a term from a quality ontology (see, e.g., [59]). An example would be supraorbital bone:sigmoid. The fact that sigmoid is a shape is inferred from the PATO ontology and thus does not have to be specifically mentioned in the EQ statement (within [59] see Figure 1). Users can load the ontology containing the terms they want to use for annotation into Phenex which has an auto-complete function to facilitate work. The Phenex GUI provides components for editing, searching, and graphical displays of terms. This software is open source, released under the MIT license (http://phenoscape.org/wiki/Phenex).

3.3. Names Recognition and Discovery. A taxonomic name is connected to almost every piece of information about an organism, making names near universal metadata in biology (see Rod Page's iphylo blog entry http://iphylo.blogspot.com/2011/04/dark-taxa-genbank-in-post-taxonomic.html for an exception). This can be exploited to find and manage nearly all biological data. No life-wide, comprehensive list of taxonomic names exists, but the Global Names Index (GNI) holds 20 million names and NameBank (http://www.ubio.org/index.php?pagename=namebank) holds 10 million names. There are also exclusive lists of taxonomically creditable names such as the Catalogue of Life (CoLP) and the Interim Register of Marine and Non-marine Genera (IRMNG). These lists hold 1.3 million and 1.6 million names, respectively.

Taxonomic names discovery (or Named Entity Recognition in computer science parlance) can be achieved through several approaches. *Dictionary-based approaches* rely on an existing list of names. These systems try to find names on the list directly in the text. The major drawback of this approach in biology is that there is no comprehensive list of names and terms including all misspellings, variants, and abbreviations. Dictionary-based approaches can also miss synonyms and ambiguous names. Some algorithms have been developed to aid dictionary-based approaches with recognizing variants of names in the list (e.g., see algorithms described below). *Rule-based approaches* work by applying a fixed set of rules to a text. This approach is capable of dealing with variations in word order and sentence structure in addition to word morphology. The major drawback is that the rule sets are handmade (and, therefore, labor intensive) and are rarely applicable to multiple domains. *Machine-learning approaches* use rule sets generated by the machine using statistical procedures (such as Hidden Markov Models). In this approach, algorithms are trained on an annotated body of text in which names are tagged by hand. The algorithms can be applied to text in any discipline as long as appropriate training data are available. All of these approaches have strengths and weaknesses, so they are often combined in final products.

Table 4: Performance metrics for the names recognition and morphological character extraction algorithms reviewed. Recall and precision values may not be directly comparable between the different algorithms. NA: not available [30].

Tool	Recall	Precision	Test Corpora	Reference
TaxonGrab	>94%	>96%	Vol. 1 Birds of the Belgian Congo by Chapin	[31]
FAT	40.2%	84.0%	American Seashells by Abbott	[32]
Taxon Finder	54.3%	97.5%	American Seashells by Abbott	[32]
Neti Neti	70.5%	98.9%	American Seashells by Abbott	[32]
LINNAEUS	94.3%	97.1%	LINNAEUS gold standard data set	[33]
Organism Tagger	94.0%	95.0%	LINNAEUS gold standard data set	[34]
X-tract	NA	NA	Flora of North America	[35]
Worldwide Botanical Knowledge Base	NA	NA	Flora of China	http://wwbota.free.fr/
Terminator	NA	NA	16 nematode descriptions	http://www.math.ucdavis.edu/~milton/genisys/terminator.html
MultiFlora	mid 60%	mid 70%	Descriptions of Ranunculus spp. from six Floras	http://intranet.cs.man.ac.uk/ai/public/MultiFlora/MF1.html
MARTT	98.0%	58.0%	Flora of North America and Flora of China	[30]
WHISK	33.33% to 79.65%	72.52% to 100%	Flora of North America	[13]
CharaParser	90.0%	91.0%	Flora of North America	[36]

Several algorithms have been developed that are capable of identifying and discovering known and unknown (to the algorithm) taxon names in free text. These are discussed below and their performance metrics are given in Table 4.

TaxonGrab. TaxonGrab identifies names by using a combination of nomenclatural rules and a list (dictionary) of non-taxonomic English terms [31]. As most taxonomic names do not match words in common parlance, the dictionary can be used as a "black list" to exclude terms. This is not always the case because some Latin names match vernacular names, such as bison and *Bison bison*. The algorithm scans text for terms that are not found in the black list. It treats these as candidate names. These terms are then compared to the capitalization rules of Linnaean nomenclature. Algorithms of this type have low precision because misspelled, non-English words, medical, or legal terms would be flagged as a candidate name. However, these terms can be iteratively added to the black list, improving future precision. This method does have the advantage of not requiring a complete list of species names, but can only be used on English texts. Later, several additional rules were added to create a new product, FAT [60]. FAT employs "fuzzy" matching and structural rules sequentially so that each rule can use the results of the last. The TaxonGrab code is available at SourceForge, but the FAT code is not. FAT is a part of the plazi.org toolset for markup of taxonomic text.

TaxonFinder. TaxonFinder identifies scientific names in free text by comparing the name to several lists embedded into the source code ([61], Leary personal comments). These lists are derived from a manually curated version of NameBank (http://www.ubio.org/index.php?pagename=namebank). A list of ambiguous names was compiled from words that are names, but are more often used in common parlance, like pluto or tumor. TaxonFinder breaks documents into words and compares them to the lists individually. When it encounters a capitalized word, it checks the "genus" and "above-genus" name lists. If the word is in the above-genus list, but not in the ambiguous name list, it is returned as a name. If it is in the genus list, the next word is checked to see if it is in lower case or all caps and to see if it is in the "species-or-below" name list. If it is, then the process is repeated with the next word until a complete polynomial is returned. If the next word is not in the list, then the previous name is returned as a genus. TaxonFinder is limited to dictionaries and thus will not find new names or misspellings but can discover new combinations of known

names. This system can have both high precision and recall with a higher score in precision (more false negatives than false positives). A previous version of TaxonFinder, FindIT (http://www.ubio.org/tools/recognize.php), had the ability to identify authorship by recognizing the reference (usually a taxonomist's name), which TaxonFinder does not do (http://code.google.com/p/taxon-name-processing/wiki/nameRecognition). A new, Apache Lucene-based name indexer is now available from GBIF which is based on TaxonFinder (http://tools.gbif.org/namefinder/). The source code for TaxonFinder is available at Google code (http://code.google.com/p/taxon-finder/).

NetiNeti. NetiNeti takes a more unsupervised approach to names extraction [32]. The system uses natural language processing techniques involving probabilistic classifiers (Naive Bayes classifier by default) to recognize scientific names in an arbitrary document. The classifier is trained to recognize characteristics of scientific names as well as the context. The algorithm uses "white list" and "black list" detection techniques in a secondary role. As a result, scientific names not mentioned in a white list or names with OCR errors or misspellings are found with great accuracy. Some of the limitations of NetiNeti include an inability to identify genus names less than four letters long, the assumption of one letter abbreviations of genera, and limitation of contextual information available to one word on either side of a candidate name. The code of this tool is written in Python and is going to be released under GPL2 license at https://github.com/mbl-cli/NetiNeti.

Linnaeus. This is a list-based system designed specifically for identifying taxonomic names in biomedical literature and linking those names to database identifiers [33]. The system recognizes names contained in a white list (based on the NCBI classification and a custom set of synonyms) and resolves them to an unambiguous NCBI taxonomy identifier within the NCBI taxonomy database (http://www.ncbi.nlm.nih.gov/books/NBK21100/). In this way, multiple names for one species are normalized to a single identifier. This system is capable of recognizing and normalizing ambiguous mentions, such as abbreviations (*C. elegans*, which refers to 41 species) and acronyms (CMV, which refers to 2 species). Acronyms that are not listed within the NCBI classification are discovered using the Acromine service [62] and a novel acronym detector built into LINNAEUS that can detect acronym definitions within text (in the form of "species (acronym)"). Ambiguous mentions that are not resolvable are assigned a probability of how likely the mention refers to a species based on the relative frequency of nonambiguous mentions across all of MEDLINE. Applying a black list of species names that occur commonly in the English language when not referring to species (such as the common name spot) greatly reduces false positives. LINNAEUS can process files in XML and txt formats and give output in tab-separated files, XML, HTML and MySQL database tables. This code is available at SourceForge (http://sourceforge.net/projects/linnaeus/).

OrganismTagger. This system uses the NCBI taxonomy database (http://www.ncbi.nlm.nih.gov/books/NBK21100/) to generate semantically enabled lists and ontology components for organism name extraction from free text [34]. These components are connected to form a work flow pipeline using GATE (the General Architecture for Text Engineering; [63, 64]). These components are a combination of rule-based and machine-learning approaches to discover and extract names from text, including strain designations. To identify strains not in the NCBI taxonomy database, OrganismTagger uses a "strain classifier," a machine-learning (SVM model) approach trained on manually annotated documents. After the strain classifier is applied, organism names are first detected, then normalized to a single canonical name and grounded to a specific NCBI database ID. The semantic nature of this tool allows it to output data in many different formats (XML, OWL, etc.). This code along with supporting materials is available under an open source license at http://www.semanticsoftware.info/organism-tagger.

3.4. Morphological Character Extraction. Morphological characters of organisms are of interest to systematists, evolutionary biologists, ecologists, and the general public. The examples used in Figure 4 are typical of morphological descriptions. The kinds of language used in biodiversity science has the following characteristics that make it difficult for general-purpose parsers to process [15, 65, 66].

(1) Specialized Language. Most scientific terms are not in the lexicons of existing parsers. Even if they were, biological terms are more ambiguous than general English [67]. General English has 0.57% ambiguous terms while gene names have 14.2% ambiguity. Taxonomic homonyms are 15% at the genus level (http://www.obis.org.au/irmng/irmng_faq/). Life Science literature also relies heavily on abbreviations [68]. There were over 64,000 new abbreviations introduced in 2004 in the biomedical literature alone and an average of one new abbreviation every 5–10 abstracts [69]. Dictionaries, such as the Dictionary of Medical Acronyms and Abbreviations can help, but most dictionaries contain 4,000 to 32,000 terms, which is only a fraction of the estimated 800,000 believed to exist [69, 70]. This means that dictionary-based approaches will not scale to work in biology.

(2) Diversity. Descriptions are very diverse across taxon groups. Even in one group, for example, plants, variations are large. Lydon et al. [71] compared and contrasted the descriptions of five common species in six different English language Floras and found the same information in all sources only 9% of the time. They also noted differences in terminology usage across Floras.

(3) Syntax differences. Many species descriptions are in telegraphic sublanguage (that lacks of verbs) but there are also many descriptions conforming to more standard English syntax. Parsers expecting standard English syntax often mistake other groups of words

for verbs when parsing telegraphic sublanguage because they expect to see verbs in a sentence. There is not typically standardized syntax across different taxon groups or even within the same group.

Taylor [15, 72] manually constructed a grammar and a lexicon of 2000 characters, and character states (1500 from Radford [73] and 500 from descriptive text) to parse the Flora of New South Wales (4 volumes) and volume 19 of the Flora of Australia. The goal of parsing these Floras was to create sets of organism part, character, and character state from each description. These statements can be extracted from morphological taxon descriptions using the hand-crafted parser to get a machine-readable set of facts about organism characteristics. While the sublanguage nature of the plant descriptions used by Taylor [15, 72] made it easier to construct the grammar and lexicon manually, the author acknowledged the limited coverage they could be expected to achieve (60–80% recall was estimated based on manual examination of output). Algorithms for machine-aided expansion of the lexicon were suggested; however, at the time automated creation of rules was believed to be too difficult.

Since Taylor [15, 72], a variety of methods have been used to extract morphological traits from morphological descriptions. Their performance metrics are given in Table 4.

X-Tract. X-tract [35] was an interactive tool to extract morphological information from Flora of North America (FNA) descriptions available as a print and HTML version. X-tract used HTML tags embedded in the FNA pages to identify the morphological description sections. It used a glossary to classify each word in a description as structure (i.e., organs or part of organs) or character states. If a word was a character state, its corresponding characters were looked up in the glossary. Then, X-tract created a form to display the structures, substructures, characters, and character states extracted from a document for a user to review, modify, and save to a database. Evaluation of the extraction accuracy or the extent of user intervention was not provided.

Worldwide Botanical Knowledge Base. Jean-Marc Vanel initiated a project called Worldwide Botanical Knowledge Base, which also takes the approach of parsing plus glossary/lexicon. It marks up morphological descriptions at sentence level (e.g., leaf blade obovate is marked as "leaf blade") without extracting detailed character information. It stores extracted information in XML files instead of a relational database as Taylor [15, 72] and Abascal and Sánchez [35]. The project aims to support queries on species descriptions in botanical databases. The database search seems to have stopped working (http://jmvanel.free.fr/protea.html). The parser was reported to work on Flora of China and it can be downloaded from the website (http://wwbota.free.fr/). However, as of the time of this publication, the authors were unable to use the parser.

Terminator. Diederich, Fortuner and Milton [74] developed a system called Terminator, which used a hand-crafted plant glossary that amounts to an ontology including structure names, characters and character states to support character extraction. The extraction process was a combination of fuzzy keyword match and heuristic extraction rules. Because Terminator was an interactive system (i.e., a human operator selects correct extractions), the evaluation was done on 16 descriptions to report the time taken to process them. Extraction performance was evaluated only on 1 random sample: for non-numerical characters, 55% of the time a perfect structure/character/value combination was among the first 5 candidates suggested by the system.

MultiFlora. Similar to previous works, Wood, Lydon, and colleagues' MultiFlora project (http://intranet.cs.man.ac.uk/ai/public/MultiFlora/MF1.html) started with manual analysis of description documents. They created an ontology manually, which included classes of organs (i.e., petal) and features (i.e., yellow) linked by properties (i.e., hasColor). They also manually created a gazetteer, which included terms referring to the organs and features that served as a lookup list. The prototype MultiFlora system used a combination of keyword matching, internal and contextual pattern matching, and shallow parsing techniques provided by GATE to extract organ and feature information from a small collection of morphological descriptions (18 species descriptions, recall, and precision were in the range of mid 60% to mid 70%; [66, 75]). While the work of Wood, Lydon, and colleagues shows that using descriptions from different sources can be used to improve recall, the authors acknowledged that organs not included in the manually-created gazetteer/ontology have to be marked as "unknown." The extraction results were output in RDF triples and used to build a knowledgebase about plants, which is not related to Worldwide Botanical Knowledge Base reviewed earlier. RDF is a type of programming language that allows a user to make machine readable assertions in the form of an RDF triple. The EQ format mentioned earlier is a similar format used in biology. The advantage to using ontology-supported RDF/EQ is that multiple data providers can use the same ontological identifier for the same term. In this way, statements become machine-readable and can be linked regardless of the source. With ontological support, machine-based logic reasoning has become possible. An immediate application of this type of reasoning and a pool of RDF triples describing species morphology is a specimen identification key. RDF is supported by more recent biodiversity IE systems as an output format.

MARTT. MARTT [76] is an automated description markup system employing a supervised machine-learning algorithm. The system marks up a description sentence-by-sentence with tags that indicate the subject, for example, "stem" is tagged in the text statement "stem solitary." MARTT along with a test collection is downloadable from http://sites.google.com/site/biosemanticsproject/project-progress-wiki. Wei [77] conducted an exploratory study of the application of information fusion techniques to taxonomic descriptions. It confirmed Wood et al. [75] finding that combining multiple descriptions of the same

species from different sources and different taxonomic ranks can provide the researchers more complete information than any single description. Wei used MARTT [76] and a set of heuristic rules to extract character information from descriptions of taxa published in both FNA and Flora of China (FoC) and categorized the extracted information between the two sources as either identical, equivalent, subsumption, complementary, overlap, or conflict. Non-conflict information from both sources was then merged together. The evaluation was conducted involving 13 human curators verifying results generated from 153 leaf descriptions. The results show that the precisions for genus level fusion, species level fusion, FNA genus-species fusion, and FoC genus-species fusion were 77%, 63%, 66%, and 71%, respectively. The research also identified the key factors that contribute to the performance of the system: the quality of the dictionary (or the domain knowledge), the variance of the vocabulary, and the quality of prior IE steps.

WHISK. Tang and Heidorn [13] adapted WHISK [78] to extract morphological character and other information from several volumes of FNA to show that IE helps the information retrieval system SEARFA (e.g., retrieval of relevant documents). The "pattern matching" learning method used by WHISK is described in Section 2. The pattern matching algorithm was assisted by a knowledge base created by manually collecting structure and character terms from training examples. The IE system was evaluated on a relatively small subset of FNA documents and it was evaluated on different template slots (see Table 1 for examples of template slots) separately. Different numbers of training and/or test examples were used for different slots (training examples ranged from 7 to 206, test examples ranged from 6 to 192) and the performance scores were obtained from one run (as opposed to using the typical protocol for supervised learning algorithms). The system performed perfectly on nonmorphological character slots (Genus, Species, and Distribution). The recall on morphological character slots (Leaf shape, Leaf margin, Leaf apex, Leaf base, Leaf arrangement, Blade dimension, Leaf color, and Fruit/nut shape) ranged from 33.33% to 79.65%. The precision ranged from 75.52% to 100%. Investigation of human user performance on plant identification using internet-based information retrieval systems showed that even with imperfect extraction performance, users were able to make significantly more identifications using the information retrieval system supported by the extracted character information than using a keyword-based full-text search system.

CharaParser. All IE systems reviewed above relied on manually created vocabulary resources, whether they are called lexicons, gazetteers, or knowledge bases. Vocabularies are a fundamental resource on which more advanced syntactic and semantic analyses are built. While manually collecting terms for a proof-of-concept system is feasible, the manual approach cannot be scaled to the problem of extracting morphological traits of all taxa. Cui, Seldon & Boufford [14] proposed an unsupervised bootstrapping based algorithm (described in Section 2) that can extract 93% of anatomical terms and over 50% character terms from text descriptions without any training examples. This efficient tool may be used to build vocabulary resources that are required to use various IE systems on new document collections.

This unsupervised algorithm has been used in two IE systems [36, 79]. One of the systems used intuitive heuristic rules to associate extracted character information with appropriate anatomical structures. The other system (called CharaParser) adapted a general-purpose syntactic parser (Stanford Parser) to guide the extraction. In addition to structures and character extraction, both systems extract constraints, modifiers, and relations among anatomical structures (e.g., head *subtended by* distal leaves; pappi *consist of* bristles) as stated in a description. Both systems were tested on two sets of descriptions from volume 19 of FNA and Part H of Treatise on Invertebrate Paleontology (TIP); each set consisted of over 400 descriptions. The heuristic rule-based system achieved precision/recall of 63%/60% on the FNA evaluation set and 52%/43% on the TIP evaluation set on character extraction. CharaParser performed significantly better and achieved precision/recall of 91%/90% on the FNA set and 80%/87% on the TIP set. Similar to Wood et al. [66], Cui and team found the information structure of morphological descriptions was too complicated to be represented in a typical IE template (such as Table 1). Wood et al. [66] designed an ontology to hold the extracted information, while Cui and team used XML to store extracted information (Figure 5). CharaParser is expected to be released as an open-source software in Fall 2012. Interested readers may contact the team to obtain a trial version before its release.

3.5. Integrated IE Systems. Tang and Heidorn [13] supervised learning IE system, MutiFlora, and the CharaParser system, all reviewed before, can be described using the reference model depicted in Figure 2. Here, we describe another system that integrates formal ontologies. This is the text mining system that is currently under development by the Phenoscape project (http://www.phenoscape.org/). The goal of Phenoscape is to turn text phenotype descriptions to EQ expressions [80] to support machine reasoning of scientific knowledge as a transforming way of conducting biological research. In this application, EQ expressions may be considered both the IE template and a data standard. The input to the Phenoscape text mining system is digital or OCRed phylogenetic publications. The character descriptions are targeted (1 character description = 1 character statement + multiple character state statements) and used to form the taxon-character matrix. The target sections are extracted by student assistants using Phenex and put into NeXML (http://www.nexml.org/) format. NeXML is an exchange standard for representing phyloinformatic data. It is inspired by the commonly used NEXUS format, but more robust and easier to process. There is one NeXML file for a source text. NeXML files are the input to CharaParser, which performs bootstrapping-based learning (i.e., unsupervised learning) and deep parsing to extract information and output candidate EQ expressions. CharaParser learns lexicons of anatomy terms and character terms from description

```
(a) Original sentence:
   principal cauline well distributed, gradually reduced distally, bases of proximal cauline winged-petiolate or
   sessile, based of distal cauline expanded and clasping margins sometimes spinier than those of proximal;
(b) Extraction Result in XML:
   <?xml version="1.0" encoding="utf-8"?>
   <statmentn id="83.txt-6">
   <structure id="o1" name="leaf" constraint="principal cauline">
     <character name="arrengement" value="distributed" modifier="well" />
     <character name="size" value="reduced" modifier="gradually;distally" />
   </structure>
   <structure id="o2" name="base">
     <character name="archilecture" value="winged-potiolate" />
     <character name="archilecture" value="sessile" />
   </structure>
   <structure id="o3" name="leaf" constraint="proximal cauline">
   <relation id="r1" name="part_of" from="o2" to="o3" negation="false" />
   <structure id="o4" name="base">
     <character name="size" value="expanded" />
     <character name="archilecture" value="clapsing" />
   </structure>
   <structure id="o5" name="leaf" constraint="distal cauline" />
   <relation id="r2" name="part_of" from="o4" to="o5" negation="false" />
   <structure id="o6" name="margin">
    <character name="archilecture" value="spinier" modifier="sometimes" constraint="thanmargins"
    constraintid="o7" />
   </structure>
   <structure id="o7" name="margin" />
   <structure id="o8" name="leaf" constraint="proximal" />
   <relation id="r3" name="part_of" from="o7" to="o8" negation="false" />
   </statment>
```

Figure 5: Extraction result from a descriptive sentence.

collections. Learned terms are reviewed by biologist curators (many OCR errors are detected during this step). Terms that are not in existing anatomy ontologies are proposed to the ontologies for addition. The lexicons and ontologies are the knowledge entities that the text mining system iteratively uses and enhances. With new terms added to the ontologies, the system replaces the terms in candidate EQ statements with term IDs from the ontologies. For example, [E]tooth [Q]large is turned into [E]TAO: 0001625 [Q]PATO: 0001202. The candidate EQ expressions are reviewed and accepted by biologist curators using Phenex. Final EQ expressions are loaded into the Phenoscape Knowledge base at http://kb.phenoscape.org/. This EQ populated knowledge base supports formal logical reasoning. At the time of writing, the developing work is ongoing to integrate CharaParser with Phenex to produce an integrated text-mining system for Phenoscape. It is important to notice that the applicability of Phenex and CharaParser is not taxon specific.

4. Conclusion

NLP approaches are capable of extracting large amounts of information from free text. However, biology text presents a unique challenge (compared to news articles) to machine-learning algorithms due to its ambiguity, diversity, and specialized language. Successful IE strategies in biodiversity science take advantage of the Linnaean binomial structure of names and the structured nature of taxon descriptions. Multiple tools currently exist for fuzzy matching of terms, automated annotation, named-entity recognition, and morphological character extraction that use a variety of approaches. None have yet been used on a large scale to extract information about all life, but several, such as CharaParser, show potential to be used in this way. Further improvement of biodiversity IE tools could be achieved through increased participation in the annual BioCreative competitions (http://www.biocreative.org/) and assessing tool performance on publicly available document sets so that comparison between systems (and thus identification of methods that have real potential to address biodiversity IE problems) becomes easier.

A long-term vision for the purpose of making biodiversity data machine readable is the compilation of semantic species descriptions that can be linked into a semantic web for biology. An example of semantic species information can be found at TaxonConcept.org. This concept raises many questions concerning semantics which are outside the scope of this paper, such as what makes a "good" semantic description of a species. Many of these issues are technical and are being addressed within the computer science community. There are two data pathways that need to be developed to achieve the semantic web for biology. One is a path going forward, in which new data are made machine-readable from the beginning of a research project. The model of mobilizing data many years after collection with little to no data management planning during collection is not sustainable or desirable going into the future. Research is being applied to this area and publishers, such as Pensoft, are working to capture machine-readable data about species at the point of publication. The other is a path for mobilizing data that have already been collected. NLP holds much promise in helping with the second path.

Mobilizing the entirety of biodiversity knowledge collected over the past 250 years is an ambitious goal that requires meeting several challenges from both the taxonomic and technological fronts. Considering the constantly changing nature of biodiversity science and the constraints of NLP algorithms, best results may be achieved by drawing information from high quality modern reviews of taxonomic groups rather than repositories of original descriptions. However, such works can be rare or nonexistent for some taxa. Thus, issues such as proper aggregation of information extracted from multiple sources on a single subject (as mentioned above) still need to be addressed. In addition, demanding that a modern review be available somewhat defeats the purpose of applying NLP to biodiversity science.

While using a modern review may be ideal when available, it should not be required for information extraction.

Biodiversity science, as a discipline, is being asked to address numerous challenges related to climate change, biodiversity loss, and invasive species. Solutions to these problems require discovery and aggregation of data from the entire pool of biological knowledge including what is contained exclusively in print holdings. Digitization and IE on this scale is unprecedented. Unsupervised algorithms hold the greatest promise for achieving the scalability required because they do not require manually generated training data. However, most successful IE algorithms use combinations of supervised and unsupervised strategies and multiple NLP approaches because not all problems can be solved with an unsupervised algorithm. If the challenge is not met, irreplaceable data from centuries of research funded by billions of dollars may be lost. The annotation and extraction algorithms mentioned in this manuscript are key steps toward liberating existing biological data and even serve as preliminary evidence that this goal can be achieved.

Acknowledgments

The authors would like to thank Dr. David J. Patterson, Dr. Holly Bowers, and Mr. Nathan Wilson for thoughtful comments on an early version of this manuscript and productive discussion. This work was funded in part by the MacArthur Foundation Grant to the Encyclopedia of Life, the National Science Foundation Data Net Program Grant no. 0830976, and the National Science Foundation Emerging Front Grant no. 0849982.

References

[1] B. Wuethrich, "How climate change alters rhythms of the wild," *Science*, vol. 287, no. 5454, pp. 793–795, 2000.

[2] W. E. Bradshaw and C. M. Holzapfel, "Genetic shift in photoperiodic response correlated with global warming," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 25, pp. 14509–14511, 2001.

[3] National Academy of Sciences, "New biology for the 21st Century," *Frontiers in Ecology and the Environment*, vol. 7, no. 9, article 455, 2009.

[4] A. E. Thessen and D. J. Patterson, "Data issues in life science," *ZooKeys*, vol. 150, pp. 15–51, 2011.

[5] A. Hey, *The Fourth Paradigm: Data-Intensive Scientific Discovery*, 2009, http://iw.fh-potsdam.de/fileadmin/FB5/Dokumente/forschung/tagungen/i-science/TonyHey_-__eScience_Potsdam__Mar2010____complete_.pdf.

[6] L. D. Stein, "Towards a cyberinfrastructure for the biological sciences: progress, visions and challenges," *Nature Reviews Genetics*, vol. 9, pp. 678–688, 2008.

[7] P. B. Heidorn, "Shedding light on the dark data in the long tail of science," *Library Trends*, vol. 57, no. 2, pp. 280–299, 2008.

[8] Key Perspectives Ltd, "Data dimensions: disciplinary differences in research data sharing, reuse and long term viability," Digital Curation Centre, 2010, http://scholar.google.com/scholar?hl=en&q=Data+Dimensions:+disciplinary+differences+in+research+data-sharing,+reuse+and+long+term+viability.++&btnG=Search&as_sdt=0,22&as_ylo=&as_vis=0#0.

[9] A. Vollmar, J. A. Macklin, and L. Ford, "Natural history specimen digitization: challenges and concerns," *Biodiversity Informatics*, vol. 7, no. 2, 2010.

[10] P. N. Schofield, J. Eppig, E. Huala et al., "Sustaining the data and bioresource commons," *Research Funding*, vol. 330, no. 6004, pp. 592–593, 2010.

[11] P. Groth, A. Gibson, and J. Velterop, "Anatomy of a Nanopublication," *Information Services & Use*, vol. 30, no. 1-2, pp. 51–56, 2010.

[12] M. Kalfatovic, "Building a global library of taxonomic literature," in *28th Congresso Brasileiro de Zoologia Biodiversidade e Sustentabilidade*, 2010, http://www.slideshare.net/Kalfatovic/building-a-global-library-of-taxonomic-literature.

[13] X. Tang and P. Heidorn, "Using automatically extracted information in species page retrieval," 2007, http://scholar.google.com/scholar?hl=en&q=Tang+Heidorn+2007+using+automatically+extracted&btnG=Search&as_sdt=0,22&as_ylo=&as_vis=0#0.

[14] H. Cui, P. Selden, and D. Boufford, "Semantic annotation of biosystematics literature without training examples," *Journal of the American Society for Information Science and Technology*, vol. 61, pp. 522–542, 2010.

[15] A. Taylor, "Extracting knowledge from biological descriptions," in *Proceedings of 2nd International Conference on Building and Sharing Very Large-Scale Knowledge Bases*, pp. 114–119, 1995.

[16] H. Cui, "Competency evaluation of plant character ontologies against domain literature," *Journal of the American Society for Information Science and Technology*, vol. 61, no. 6, pp. 1144–1165, 2010.

[17] Y. Miyao, K. Sagae, R. Sætre, T. Matsuzaki, and J. Tsujii, "Evaluating contributions of natural language parsers to protein-protein interaction extraction," *Bioinformatics*, vol. 25, no. 3, pp. 394–400, 2009.

[18] K. Humphreys, G. Demetriou, and R. Gaizauskas, "Two applications of information extraction to biological science journal articles: enzyme interactions and protein structures," in *Proceedings of the Pacific Symposium on Biocomputing (PSB '00)*, vol. 513, pp. 505–513, 2000.

[19] R. Gaizauskas, G. Demetriou, P. J. Artymiuk, and P. Willett, "Protien structures and information extraction from biological texts: the pasta system," *Bioinformatics*, vol. 19, no. 1, pp. 135–143, 2003.

[20] A. Divoli and T. K. Attwood, "BioIE: extracting informative sentences from the biomedical literature," *Bioinformatics*, vol. 21, no. 9, pp. 2138–2139, 2005.

[21] D. P. A. Corney, B. F. Buxton, W. B. Langdon, and D. T. Jones, "BioRAT: extracting biological information from full-length papers," *Bioinformatics*, vol. 20, no. 17, pp. 3206–3213, 2004.

[22] H. Chen and B. M. Sharp, "Content-rich biological network constructed by mining PubMed abstracts," *Bmc Bioinformatics*, vol. 5, article 147, 2004.

[23] X. Zhou, X. Zhang, and X. Hu, "Dragon toolkit: incorporating auto-learned semantic knowledge into large-scale text retrieval and mining," in *Proceedings of the19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI '07)*, pp. 197–201, October 2007.

[24] D. Rebholz-Schuhmann, H. Kirsch, M. Arregui, S. Gaudan, M. Riethoven, and P. Stoehr, "EBIMed—text crunching to gather facts for proteins from Medline," *Bioinformatics*, vol. 23, no. 2, pp. e237–e244, 2007.

[25] Z. Z. Hu, I. Mani, V. Hermoso, H. Liu, and C. H. Wu, "iProLINK: an integrated protein resource for literature

mining," *Computational Biology and Chemistry*, vol. 28, no. 5-6, pp. 409–416, 2004.

[26] J. Demaine, J. Martin, L. Wei, and B. De Bruijn, "LitMiner: integration of library services within a bio-informatics application," *Biomedical Digital Libraries*, vol. 3, article 11, 2006.

[27] M. Lease and E. Charniak, "Parsing biomedical literature," in *Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCNLP '05)*, Jeju Island, Korea, 2005.

[28] S. Pyysalo and T. Salakoski, "Lexical adaptation of link grammar to the biomedical sublanguage: a comparative evaluation of three approaches," *BMC Bioinformatics*, vol. 7, supplement 3, article S2, 2006.

[29] L. Rimell and S. Clark, "Porting a lexicalized-grammar parser to the biomedical domain," *Journal of Biomedical Informatics*, vol. 42, no. 5, pp. 852–8865, 2009.

[30] H. Cui, "Converting taxonomic descriptions to new digital formats," *Biodiversity Informatics*, vol. 5, pp. 20–40, 2008.

[31] D. Koning, I. N. Sarkar, and T. Moritz, "TaxonGrab: extracting taxonomic names from text," *Biodiversity Informatics*, vol. 2, pp. 79–82, 2005.

[32] L. M. Akella, C. N. Norton, and H. Miller, "NetiNeti: discovery of scientific names from text using machine learning methods," 2011.

[33] M. Gerner, G. Nenadic, and C. M. Bergman, "LINNAEUS: a species name identification system for biomedical literature," *BMC Bioinformatics*, vol. 11, article 85, 2010.

[34] N. Naderi and T. Kappler, "OrganismTagger: detection, normalization and grounding of organism entities in biomedical documents," *Bioinformatics*, vol. 27, no. 19, pp. 2721–2729, 2011.

[35] R. Abascal and J. A. Sánchez, "X-tract: structure extraction from botanical textual descriptions," in *Proceeding of the String Processing & Information Retrieval Symposium & International Workshop on Groupware*, pp. 2–7, IEEE Computer Society, Cancun , Mexico, September 1999.

[36] H. Cui, "CharaParser for fine-grained semantic annotation of organism morphological descriptions," *Journal of the American Society for Information Science and Technology*, vol. 63, no. 4, pp. 738–754, 2012.

[37] M. Krauthammer, A. Rzhetsky, P. Morozov, and C. Friedman, "Using BLAST for identifying gene and protein names in journal articles," *Gene*, vol. 259, no. 1-2, pp. 245–252, 2000.

[38] L. Lenzi, F. Frabetti, F. Facchin et al., "UniGene tabulator: a full parser for the UniGene format," *Bioinformatics*, vol. 22, no. 20, pp. 2570–2571, 2006.

[39] A. Nasr and O. Rambow, "Supertagging and full parsing," in *Proceedings of the 7th International Workshop on Tree Adjoining Grammar and Related Formalisms (TAG '04)*, 2004.

[40] R. Leaman and G. Gonzalez, "BANNER: an executable survey of advances in biomedical named entity recognition," in *Proceedings of the Pacific Symposium on Biocomputing (PSB '08)*, pp. 652–663, Kona, Hawaii, USA, January 2008.

[41] M. Schröder, "Knowledge-based processing of medical language: a language engineering approach," in *Proceedings of the16th German Conference on Artificial Intelligence (GWAI '92)*, vol. 671, pp. 221–234, Bonn, Germany, August-September 1992.

[42] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann Series in Data Management Systems, Morgan Kaufmann, 2nd edition, 2005.

[43] C. Blaschke, L. Hirschman, and A. Valencia, "Information extraction in molecular biology," *Briefings in Bioinformatics*, vol. 3, no. 2, pp. 154–165, 2002.

[44] A. Jimeno-Yepes and A. R. Aronson, "Self-training and co-training in biomedical word sense disambiguation," pp. 182–183.

[45] C. Freeland, "An evaluation of taxonomic name finding & next steps in Biodiversity Heritage Library (BHL) developments," *Nature Precedings*, 2009, http://precedings.nature.com/documents/3372/version/1.

[46] A. Kornai, "Experimental hmm-based postal ocr system," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '97)*, vol. 4, pp. 3177–3180, April 1997.

[47] A. Kornai, K. Mohiuddin, and S. D. Connell, "Recognition of cursive writing on personal checks," in *Proceedings of the 5th International Workshop on Frontiers in Handwriting Recognition*, pp. 373–378, Citeseer, Essex, UK, 1996.

[48] C. Freeland, "Digitization and enhancement of biodiversity literature through OCR, scientific names mapping and crowdsourcing.," in *BioSystematics Berlin*, 2011, http://www.slideshare.net/chrisfreeland/digitization-and-enhancement-of-biodiversity-literature-through-ocr-scientific-names-mapping-and-crowdsourcing.

[49] A. Willis, D. King, D. Morse, A. Dil, C. Lyal, and D. Roberts, "From XML to XML: the why and how of making the biodiversity literature accessible to researchers," in *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC '10)*, pp. 1237–1244, European Language Resources Association (ELRA), Valletta, Malta, May 2010.

[50] F. Bapst and R. Ingold, "Using typography in document image analysis," in *Proceedings of Raster Imaging and Digital Typography (RIDT '98)*, pp. 240–251, Saint-Malo, France, March-April 1998.

[51] A. L. Weitzman and C. H. C. Lyal, *An XML Schema for Taxonomic Literature—TaXMLit*, 2004, http://www.sil.si.edu/digitalcollections/bca/documentation/taXMLitv1-3Intro.pdf.

[52] T. Rees, "TAXAMATCH, a "fuzzy" matching algorithm for taxon names, and potential applications in taxonomic databases," in *Proceedings of TDWG*, 2008, pp. 35, http://www.tdwg.org/fileadmin/2008conference/documents/Proceedings2008.pdf#page=35.

[53] G. Sautter, K. Böhm, and D. Agosti, "Semi-automated xml markup of biosystematic legacy literature with the goldengate editor," in *Proceedings of the Pacific Symposium on Biocomputing (PSB '07)*, pp. 391–402, World Scientific, 2007.

[54] B. Settles, "ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text," *Bioinformatics*, vol. 21, no. 14, pp. 3191–3192, 2005.

[55] G. A. Pavlopoulos, E. Pafilis, M. Kuhn, S. D. Hooper, and R. Schneider, "OnTheFly: a tool for automated document-based text annotation, data linking and network generation," *Bioinformatics*, vol. 25, no. 7, pp. 977–978, 2009.

[56] E. Pafilis, S. I. O'Donoghue, L. J. Jensen et al., "Reflect: augmented browsing for the life scientist," *Nature Biotechnology*, vol. 27, no. 6, pp. 508–510, 2009.

[57] M. Kuhn, C. von Mering, M. Campillos, L. J. Jensen, and P. Bork, "STITCH: interaction networks of chemicals and proteins," *Nucleic Acids Research*, vol. 36, no. 1, pp. D684–D688, 2008.

[58] J. P. Balhoff, W. M. Dahdul, C. R. Kothari et al., "Phenex: ontological annotation of phenotypic diversity," *Plos ONE*, vol. 5, no. 5, article e10500, 2010.

[59] W. M. Dahdul, J. P. Balhoff, J. Engeman et al., "Evolutionary characters, phenotypes and ontologies: curating data from the systematic biology literature," *Plos ONE*, vol. 5, no. 5, Article ID e10708, 2010.

[60] G. Sautter, K. Bohm, and D. Agosti, "A combining approach to find all taxon names (FAT) in legacy biosystematics literature," *Biodiversity Informatics*, vol. 3, pp. 46–58, 2007.

[61] P. R. Leary, D. P. Remsen, C. N. Norton, D. J. Patterson, and I. N. Sarkar, "UbioRSS: tracking taxonomic literature using RSS," *Bioinformatics*, vol. 23, no. 11, pp. 1434–1436, 2007.

[62] N. Okazaki and S. Ananiadou, "Building an abbreviation dictionary using a term recognition approach," *Bioinformatics*, vol. 22, no. 24, pp. 3089–3095, 2006.

[63] K. Bontcheva, V. Tablan, D. Maynard, and H. Cunningham, "Evolving gate to meet new challenges in language engineering," *Natural Language Engineering*, vol. 10, no. 3-4, pp. 349–373, 2004.

[64] H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan, C. Ursu et al., *Developing Language Processing Components with GATE (A User Guide)*, University of Sheffield, 2006.

[65] E. Fitzpatrick, J. Bachenko, and D. Hindle, "The status of telegraphic sublanguages," in *Analyzing Language in Restricted Domains: Sublanguage Description and Processing*, pp. 39–51, 1986.

[66] M. Wood, S. Lydon, V. Tablan, D. Maynard, and H. Cunningham, "Populating a database from parallel texts using ontology-based information extraction," in *Natural Language Processing and Information Systems*, vol. 3136, pp. 357–365, 2004.

[67] L. Chen, H. Liu, and C. Friedman, "Gene name ambiguity of eukaryotic nomenclatures," *Bioinformatics*, vol. 21, no. 2, pp. 248–256, 2005.

[68] H. Yu, W. Kim, V. Hatzivassiloglou, and W. J. Wilbur, "Using MEDLINE as a knowledge source for disambiguating abbreviations and acronyms in full-text biomedical journal articles," *Journal of Biomedical Informatics*, vol. 40, no. 2, pp. 150–159, 2007.

[69] J. T. Chang and H. Schutze, "Abbreviations in biomedical text," in *Text Mining for Biology and Biomedicine*, pp. 99–119, 2006.

[70] J. D. Wren and H. R. Garner, "Heuristics for identification of acronym-definition patterns within text: towards an automated construction of comprehensive acronym-definition dictionaries," *Methods of Information in Medicine*, vol. 41, no. 5, pp. 426–434, 2002.

[71] S. Lydon and M. Wood, "Data patterns in multiple botanical descriptions: implications for automatic processing of legacy data," *Systematics and Biodiversity*, vol. 1, no. 2, pp. 151–157, 2003.

[72] A. Taylor, "Using prolog for biological descriptions," in *Proceedings of The 3rd international Conference on the Practical Application of Prolog*, pp. 587–597, 1995.

[73] A. E. Radford, *Fundamentals of Plant Systematics*, Harper & Row, New York, NY, USA, 1986.

[74] J. Diederich, R. Fortuner, and J. Milton, "Computer-assisted data extraction from the taxonomical literature," 1999, http://math.ucdavis.edu/~milton/genisys.html.

[75] M. Wood, S. Lydon, V. Tablan, D. Maynard, and H. Cunningham, "Using parallel texts to improve recall in IE," in *Proceedings of Recent Advances in Natural Language Processing (RANLP '03)*, pp. 505–512, Borovetz, Bulgaria, 2003.

[76] H. Cui and P. B. Heidorn, "The reusability of induced knowledge for the automatic semantic markup of taxonomic descriptions," *Journal of the American Society for Information Science and Technology*, vol. 58, no. 1, pp. 133–149, 2007.

[77] Q. Wei, *Information fusion in taxonomic descriptions*, Ph.D. thesis, University of Illinois at Urbana-Champaign, Champaign, Ill, USA, 2011.

[78] S. Soderland, "Learning information extraction rules for semi-structured and free text," *Machine Learning*, vol. 34, no. 1, pp. 233–272, 1999.

[79] H. Cui, S. Singaram, and A. Janning, "Combine unsupervised learning and heuristic rules to annotate morphological characters," *Proceedings of the American Society for Information Science and Technology*, vol. 48, no. 1, pp. 1–9, 2011.

[80] P. M. Mabee, M. Ashburner, Q. Cronk et al., "Phenotype ontologies: the bridge between genomics and evolution," *Trends in Ecology and Evolution*, vol. 22, no. 7, pp. 345–350, 2007.