OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

Education

Biomedical Text Mining and Its Applications

Raul Rodriguez-Esteban*

Pfizer Research Technology Center, Cambridge, Massachusetts, United States of America

Introduction

This tutorial is intended for biologists and computational biologists interested in adding text mining tools to their bioinformatics toolbox. As an illustrative example, the tutorial examines the relationship between progressive multifocal leukoencephalopathy (PML) and antibodies. Recent cases of PML have been associated to the administration of some monoclonal antibodies such as efalizumab [1]. Those interested in a further introduction to text mining may also want to read other reviews [2–4].

Understanding large amounts of text with the aid of a computer is harder than simply equipping a computer with a grammar and a dictionary. A computer, like a human, needs certain specialized knowledge in order to understand text. The scientific field that is dedicated to train computers with the right knowledge for this task (among other tasks) is called natural language processing (NLP). Biomedical text mining (henceforth, text mining) is the subfield that deals with text that comes from biology, medicine, and chemistry (henceforth, biomedical text). Another popular name is BioNLP, which some practitioners use as synonymous with text mining.

Biomedical text is not a homogeneous realm [5]. Medical records are written differently from scientific articles, sequence annotations, or public health guidelines. Moreover, local dialects are not uncommon [6]. For example, medical centers develop their own jargons and laboratories create their idiosyncratic protein nomenclatures. This variability means, in practice, that text mining applications are tailored to specific types of text. In particular, for reasons of availability and cost, many are designed for scientific abstracts in English from Medline.

Main Concepts

Terms

A *term* is a name used in a specific domain, and a *terminology* is a collection of terms. Terms abound in biomedical text, where they constitute important building blocks. Some examples of terms are the names of cell types, proteins, medical devices, diseases, gene mutations, chemical names, and protein domains [7]. Due to their importance, text miners have worked to design algorithms that recognize terms (see examples in Figure 1). The task of recognizing terms is also called *named entity recognition* in the text mining literature, although this NLP task is broader and goes beyond recognition of terms. Although the concept of term is intuitive (or, perhaps, *because* it is intuitive), terms are hard to define precisely [8]. For example, the text "early progressive multifocal leukoencephalopathy" could possibly refer to any, or all, of these disease terms: "early progressive multifocal leukoencephalopathy," "progressive multifocal leukoencephalopathy," "multifocal leukoencephalopathy," and "leukoencephalopathy." To overcome such dilemmas, text miners ask experts to identify terms within collections of text such as sets of selected Medline abstracts. These annotations are then used to train a computer by example, so that the computer can emulate the knowledge experts deploy when they read biomedical text. This pedagogical method, "teaching by example," is a common approach used in many text mining tasks and it is more generally called supervised training. (Alternatively, text miners create rules using expert knowledge.) Thus, text miners rely heavily on collections of text (corpora) that have been annotated by experts (see compilations of corpora: http://www2.informatik.hu-berlin.de/~hakenber/links/benchmarks.html;http://compbio.uchsc.edu/ccp/corpora/obtaining.shtml). Before beginning a text mining task, it is advisable to limit the scope of the task to a corpus made of a set of documents around the topic of interest. In our case, a PML corpus could comprise all the Medline abstracts that mention the term "progressive multifocal leukoencephalopathy," because this is an unambiguous term. Another relevant corpus to consider could be the ImmunoTome [9], which is focused on immunology.

Text miners are interested in terminologies that have been built manually. These controlled terminologies have notable roles in biomedicine, for example, the HUGO gene nomenclature, the ICD disease classification, or the Gene Ontology. Many of these terminologies are more than just a flat list of terms. Some include term synonyms (thesauri) or relations between terms (taxonomies, ontologies). For text miners, their usefulness comes from their ability to link to information. Once a text is mapped to one of these terminologies, a bridge is opened between the text and other resources. This usefulness justifies efforts such as the National Library of Medicine's manual mapping of Medline abstracts to the Medical Subject Headings (MeSH) terminology. In our example, MeSH can be used to make the PML corpus more focused by restricting it only to abstracts with the MeSH term "leukoencephalopathy, progressive multifocal." Controlled terminologies can be used to annotate results from experiments and databases [10]. Text miners attempt to make such mappings automatically. For example, a task called *gene normalization* consists in recognizing names of genes in text and mapping them to their corre-

Citation: Rodriguez-Esteban R (2009) Biomedical Text Mining and Its Applications. PLoS Comput Biol 5(12): e1000597. doi:10.1371/journal.pcbi.1000597

Editor: Fran Lewitter, Whitehead Institute, United States of America

Published December 24, 2009

Copyright: © 2009 Raul Rodriguez-Esteban. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: The author received no specific funding for this work.

Competing Interests: The author has declared that no competing interests exist.

* E-mail: raul.rodriguez-esteban@pfizer.com

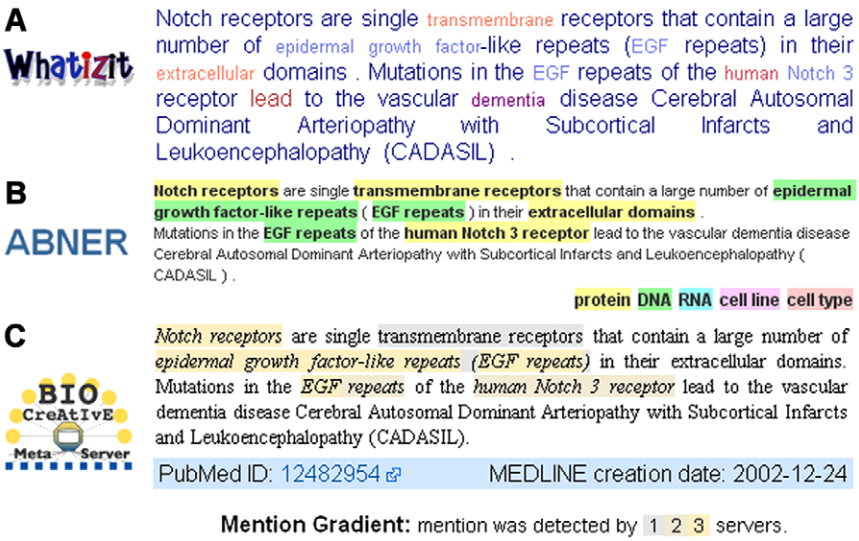

Figure 1. Examples of term recognition. (A) Text marked with protein (blue), disease (crimson), Gene Ontology (bright red), chemical (dark red), and species (red) terms by Whatizit [15] with the *whatizitEBIMedDiseaseChemicals* pipeline. (B) Text marked with protein and cell line terms by ABNER [16]. (C) Protein terms identified by the prototype BIOCreAtIvE metaserver [68]. In the example shown, the metaserver combines the output of systems hosted in three servers.
doi:10.1371/journal.pcbi.1000597.g001

sponding gene identifiers (e.g., Entrez Gene ID). Thus, using gene normalization it is possible to identify all the abstracts in Medline that mention a given gene from Entrez Gene [11].

Because there are many controlled terminologies, some terminologies have been created to map between them. For example, the BioThesaurus [12] is a compilation of protein synonyms from several terminologies. The Unified Medical Language System (UMLS) [13,14] is a grand compilation of more than 120 terminologies and close to 4 million terms. Despite UMLS's size, all controlled terminologies are incomplete, because new terms are created too quickly to keep them up to date. Furthermore, all have gaps and areas of emphasis that conflict with the needs of users.

Tools for Terms

Whatizit [15] is a tool that recognizes several types of terms. It can be accessed through a Web interface, Web services, or a streamed servlet. Abner [16] is a standalone application that recognizes five types of terms: protein, DNA, RNA, cell line, and cell type. More specialized term recognition has been used, for example, for databases such as LSAT [17] for alternative transcripts and PepBank [18] for peptides. Text miners have also used terminologies to enrich PubMed's search capabilities. Some recent search engines are semedico [19], novo|seek [20], and GoPubMed/GoGene [21,22].

Relationships

After recognizing terms, the natural next step is to look for relationships between terms. The simplest method to identify relationships is using the *co-occurrence* assumption: terms that appear in the same texts tend to be related. For example, if a protein is mentioned often in the same abstracts as a disease, it is reasonable to hypothesize that the protein is involved in some aspect of the disease. The degree of co-occurrence can be quantified statistically to rank and eliminate statistically weak co-occurrences (see Box 1). An example using GoGene [22] can illustrate the use of simple co-occurrence, MeSH terms, and gene normalization. The query *"leukoencephalopathy, progressive multifocal"[mh]* in GoGene returns all the genes mentioned in Medline abstracts annotated with the MeSH term for PML. The genes that appear most often are likely to be related to PML. Those that appear disproportionately more often for PML than for other diseases are likely to be more specific to PML.

Better evidence than co-occurrence comes from relationships that are described explicitly [23]. For example, the sentence "We describe a *PML* in a 67-year-old woman with a destructive polyarthritis associated with *anti-JO1 antibodies* treated with corticosteroids" [24] describes an explicit link between PML and anti-JO1 antibodies. We can simplify this relationship into a triplet of two terms

Box 1. The strength of a relationship. The confidence in a fact that comes from text can be qualified by the level of certainty of the assertion where the fact was found or by the strength of the evidence pointed [71]. Since facts do not stand alone, this confidence depends also on the fact's consistency with related facts [72]. In the case of co-occurrence of two terms t_1 and t_2, the simplest confidence metric is the count c of texts that include both terms, $c(t_1 \wedge t_2)$ (for a PPI example, see [73]). This measure can be normalized by the possibility of random co-occurrences due to the sheer popularity of one or both terms. For example,

$$\frac{c(t_1 \wedge t_2)}{c(t_1)c(t_2)}.$$

Pointwise mutual information (PMI) is similarly derived as

$$PMI(t_1,t_2) = \log_2\left(\frac{p(t_1 \wedge t_2)}{p(t_1)p(t_2)}\right),$$

where p, in this case, is c divided by the total number of texts. More generally, different measures can be drawn from the 2×2 contingency table that encompasses the counts of texts that include the two terms, $c(t_1 \wedge t_2)$, only one term ($c(t_1 \wedge \neg t_2)$ and $c(\neg t_1 \wedge t_2)$), and none, $c(\neg t_1 \wedge \neg t_2)$. Using this contingency table, Medgene [32] compared the merit of different statistical measures for gene-disease associations such as chi-square analysis, Fisher's exact probabilities, relative risk of gene, and relative risk of disease. More heuristic methods have been devised that use manually adjusted weights for different types of co-occurrence [36].

and a verb: PML *is associated with* anti-JO1 antibodies. To create the triplet, the verb can be identified with the aid of a part-of-speech (POS) tagger. An example of a POS tagger for biomedical text is MedPost [25]. This triplet representation is powerful due to its simplicity, but it omits crucial details from the original article, such as the fact that the evidence comes from a clinical case study.

A heavily studied area in text mining concerns the relationships known as protein-protein interactions (PPI). Using the triplet representation, PPI can be depicted as network graphs with the proteins as nodes and the verbs as edges (see Figure 2). When analyzing text-mined interaction networks, it is important to understand the information that underpins them. For example, interactions can be direct (physical) or indirect, depending on the verb (examples of direct verbs are *to bind*, *to stabilize*, *to phosphorylate*; examples of indirect verbs are *to induce*, *to trigger*, *to block*) [26]. The different nature of the protein interactions described in the literature reflects in part the experimental methodology employed and the nature of the interaction itself. A common way to capture the textual variations is by exhaustively identifying all the patterns that appear and writing a set of rules that capture them [27,28]. For example, a simple pattern to capture phosphorylations might involve, sequentially, a kinase name, a form of the verb *to phosphorylate*, and a substrate name [29,30].

Tools for Relationships

To see co-occurrence in action, try FACTA [31]. MedGene and BioGene [32,33] use co-occurrence for gene prioritization. Gene prioritization tools such as Endeavour [34] and G2D [35] use text as well as other data sources. PolySearch [36] uses heuristic weighting of different co-occurrence measures and includes a detailed guide to implementation and vocabularies. Anni [37] uses textual profiles instead of co-occurrence to measure relationship between terms. For PPI, iHOP [38] is the most popular tool. RLIMS-P [30] uses linguistic patterns to detect the kinase, substrate, and phosphosite in a phosphorylation. E3Miner [39] detects ubiquitinations, including contextual information.

Discovery

Besides finding relationships, text miners are also interested in *discovering* relationships. Due to the size of the literature, scientists miss links between their work and other, related work. Swanson called these links "undiscovered public knowledge." In a classic example he found by careful reading 11 links between magnesium and migraine that had been neglected [40]. One method to discover relationships is based on transitive inference [41]. Simply stated, if A is linked to B, and B is linked to C, then there is a chance that A is linked to C. PPI networks are, at the core, an example of transitive inference. Arrowsmith [42] is a basic discovery tool that compares two literature sets to find links between them. Applying Arrowsmith to the literature for PML and antibodies yields the immunomodulator tacrolimus, a calcineurin inhibitor, among the top hits. Tacrolimus affects the production of several proteins depicted in Figure 2, such as IL-2.

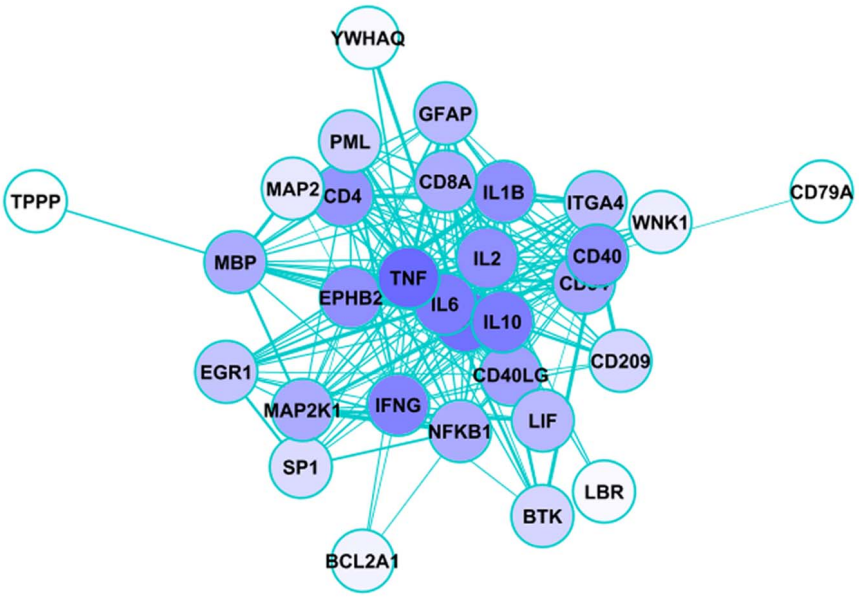

Figure 2. Example of text-mined PPI network. The nodes are proteins identified using the query: *"leukoencephalopathy, progressive multifocal"[mh] antibody[pubmed]* in GoGene [22]. The query retrieves gene symbols mapped to PubMed abstracts that include the keyword *antibody* and the MeSH term *leukoencephalopathy, progressive multifocal* (PML). The gene list was exported to SIF format and the gene symbols extracted and used to query PPI using iHOP Web services [69]. Only those iHOP interactions with at least two co-occurrences and confidence above zero were considered. The network was plotted using Cytoscape [70]. The node color is based on the number of interactions (node degree).
doi:10.1371/journal.pcbi.1000597.g002

Quality

The most common measure of output quality in text mining is the F-measure, which is the harmonic mean of two other measures, precision and recall. These three measures can be described with the analogy of searching for needles in a haystack. After a manual search of a haystack, our hands end up full with valuable needles but also with some useless straws. Recall is based on the number of needles found. High recall means that we have found most of the needles for which we were looking. Precision, however, is based on the number of both needles and straws. High precision means that we have retrieved far more needles than straws. Both high precision and high recall are desirable, and a high F-measure reflects both because it is the harmonic mean. Optimizing the F-measure of a text mining application is often different from optimizing the accuracy, because there are usually few needles and large amounts of hay in the haystack. An application that identifies the whole haystack as being only hay is quite accurate but misses all the needles.

It is important to ponder over the way an application has been evaluated before assessing its F-measure [43], and especially to consider how realistic the evaluation was. The F-measure is not an absolute value. The larger a haystack is, the more difficult it is to find needles. In other words, a low F-measure might reflect a harder task, not a worse application. Moreover, text mined applications may perform differently in different types of text and this may be reflected in lower F-measures than advertised. When the F-measure attainable is not high enough, one solution is to use text mining as a filter. A filter needs high recall, but only moderate precision, to reduce the amount of hay without affecting the needles.

Filtering with text mining is used as a preliminary step in databases such as MINT [44], DIP [45], and BIND [46]. Filtering is followed by human *curation*, which involves the review and assessment of results to reduce hay and, hopefully, provide feedback to improve the filtering. The feedback loop between text mining and curation can have an incremental positive impact in output results [47].

Comprehensiveness

Doing comprehensive text mining means considering all sources of information—Medline and beyond. The abstract conveys an article's main findings, but many other pieces of information are elsewhere in the full text, figures, tables, supplementary information, references, databases, Web sites, and multimedia files. In particular, the full text is critical for information that rarely appears in abstracts, such as experimental measurements. A more comprehensive PML corpus would include full text articles, however despite the surge in open access articles (see the Directory of Open Access Journals, www.doaj.org; [48]), the majority of published articles have access and processing restrictions. PubMed Central [49] is the main source of open access articles, and the specialized search engines BioText [50], Yale Image Finder [51], and Figurome [52] search PubMed Central figures and tables. A search for "progressive multifocal leukoencephalopathy" in the Yale Image Finder yields only one figure, while a search for "PML" yields a large number of hits, most of them not relevant because PML is an ambiguous acronym.

Text and DNA

Considering text as a sequence of symbols as informative as a protein's DNA sequence is the underlying premise of many text mining tools for bioinformatics. For example, the linguistic similarity between protein corpora (sets of texts built around proteins) correlates with the BLAST score between those same proteins [53]. Text that is used in articles or database annotations to describe a protein can be used for protein clustering and to predict structure [54], subcellular localization, and function [55]. For example, a protein corpus of a protein located in the nucleus uses a vocabulary that is somewhat different from a corpus built around a secreted protein. These vocabulary differences can be used to predict the subcellular localization of a protein of unknown location. One way to measure vocabulary differences is to represent the texts as vectors of word counts. The word counts can be normalized by the size of the text they come from and the vectors compared using, for example, Euclidean distance (for more, see [56]). To reduce vector dimensionality, some words can be grouped using a method called stemming. A simple example of stemming is converting plural nouns into singular form and verbs into infinitive form (a widely used stemming algorithm is the Porter stemmer [57]). Additional simplification can be achieved via tokenization, because some words can be separated into constitutive elements called tokens. In English, however, most words are a single token. An example of a word of two tokens is *don't*.

Text mining applications for bioinformatics [58] include subcellular localization prediction such as Sherloc and Epiloc [59,60] and protein clustering such as TXTGate [61]. Thus, text mining tools can be used for annotating biological databases in the same fashion other bioinformatics tools are used.

More Tools

An extensive list of text mining applications is maintained in http://zope.bioinfo.cnio.es/bionlp_tools/ [62]. A growing number of tools are being developed under a standard framework called UIMA, which comprises NLP as well as BioNLP tools [63].

Conclusion

Text mining tools are increasingly more accessible to biologists and computational biologists and these can often be applied to answer scientific questions in combination with other bioinformatics tools. Getting acquainted with them is a first step towards grasping the possibilities of text mining and towards venturing into the algorithms described in the literature. One way to get started on this path is by looking at examples such as [64–67].

Acknowledgments

I would like to thank Rohitha P. SriRamaratnam for comments on the manuscript.

References

1. Sobell JM, Weinberg JM (2009) Patient fatalities potentially associated with efalizumab use. J Drugs Dermatol 8: 215.
2. Cohen KB, Hunter L (2008) Getting started in text mining. PLoS Comput Biol 4: e20. doi:10.1371/journal.pcbi.0040020.
3. Rzhetsky A, Seringhaus M, Gerstein MB (2009) Getting started in text mining: part two. PLoS Comput Biol 5: e1000411. doi:10.1371/journal.pcbi.1000411.
4. Rzhetsky A, Seringhaus M, Gerstein M (2008) Seeking a new biology through text mining. Cell 134: 9–13.
5. Friedman C, Kra P, Rzhetsky A (2002) Two biomedical sublanguages: a description based on the theories of Zellig Harris. J Biomed Inform 35: 222–235.
6. Netzel R, Perez-Iratxeta C, Bork P, Andrade MA (2003) The way we write. EMBO Rep 4: 446–451.
7. Krauthammer M, Nenadic G (2004) Term identification in the biomedical literature. J Biomed Inform 37: 512–526.
8. Tanabe L, Xie N, Thom LH, Matten W, Wilbur WJ (2005) GENETAG: a tagged corpus for gene/protein named entity recognition. BMC Bioinformatics 6 Suppl 1: S3.
9. Kabiljo R, Shepherd AJ (2008) Protein name tagging in the immunological domain. Proceedings of the Third International Symposium on Semantic Mining in Biomedicine (SMBM 2008). pp 141–144.
10. Lu X, Zhai C, Gopalakrishnan V, Buchanan BG (2004) Automatic annotation of protein motif function with Gene Ontology terms. BMC Bioinformatics 5: 122.
11. Morgan AA, Lu Z, Wang X, Cohen AM, Fluck J, et al. (2008) Overview of BioCreative II gene normalization. Genome Biol 9 Suppl 2: S3.
12. Liu H, Hu ZZ, Zhang J, Wu C (2006) BioThesaurus: a web-based thesaurus of protein and gene names. Bioinformatics 22: 103–105. Available: http://pir.georgetown.edu/pirwww/iprolink/biothesaurus.shtml.
13. Bangalore A, Thorn KE, Tilley C, Peters L (2003) The UMLS knowledge source server: an object model for delivering UMLS data. AMIA Annu Symp Proc. pp 51–55. Available: http://www.nlm.nih.gov/research/umls/.
14. Aronson AR (2001) Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp. pp 17–21. Available: http://mmtx.nlm.nih.gov/.
15. Rebholz-Schuhmann D, Arregui M, Gaudan S, Kirsch H, Jimeno A (2008) Text processing through web services: calling Whatizit. Bioinformatics 24: 296–298. Available: http://www.ebi.ac.uk/webservices/whatizit/info.jsf.
16. Settles B (2005) ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. Bioinformatics 21: 3191–3192. Available: http://pages.cs.wisc.edu/~bsettles/abner/.
17. Shah PK, Bork P (2006) LSAT: learning about alternative transcripts in MEDLINE. Bioinformatics 22: 857–865. Available: http://www.bork.embl.de/LSAT.
18. Shtatland T, Guettler D, Kossodo M, Pivovarov M, Weissleder R (2007) PepBank–a database of peptides based on sequence text mining and public peptide data sources. BMC Bioinformatics 8: 280. Available: http://pepbank.mgh.harvard.edu/.
19. Wermter J, Tomanek K, Hahn U (2009) High-performance gene name normalization with GeNo. Bioinformatics 25: 815–821. Available: http://www.semedico.org/.
20. Alonso-Allende R (2009) Accelerating searches of research grants and scientific literature with novo|seek. Nat Methods 6. Advertising feature. Available: http://www.novoseek.com/.
21. Doms A, Schroeder M (2005) GoPubMed: exploring PubMed with the Gene Ontology. Nucleic Acids Res 33: W783–W786. Available: http://www.gopubmed.com.
22. Plake C, Royer L, Winnenburg R, Hakenberg J, Schroeder M (2009) GoGene: gene annotation in

the fast lane. Nucleic Acids Res 37(Web Server issue). pp W300–W304. Available: http://www.gopubmed.org/gogene/.
23. Shatkay H, Pan F, Rzhetsky A, Wilbur WJ (2008) Multi-dimensional classification of biomedical text: toward automated, practical provision of high-utility text to diverse users. Bioinformatics 24: 2086–2093.
24. Viallard JF, Lazaro E, Ellie E, Eimer S, Camou F, et al. (2007) Improvement of progressive multifocal leukoencephalopathy after cidofovir therapy in a patient with a destructive polyarthritis. Infection 35: 33–36.
25. Smith L, Rindflesch T, Wilbur WJ (2004) MedPost: a part-of-speech tagger for bioMedical text. Bioinformatics 20: 2320–2321. Available: http://www.ncbi.nlm.nih.gov/staff/lsmith/MedPost.html.
26. Santos C, Eggle D, States DJ (2005) Wnt pathway curation using automated natural language processing: combining statistical methods with partial and full parse for knowledge extraction. Bioinformatics 21: 1653–1658.
27. Friedman C, Kra P, Yu H, Krauthammer M, Rzhetsky A (2001) GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. Bioinformatics 17 Suppl 1: S74–S82.
28. Blaschke C, Valencia A (2001) The potential use of SUISEKI as a protein interaction discovery tool. Genome Inform 12: 123–134.
29. Hu ZZ, Narayanaswamy M, Ravikumar KE, Vijay-Shanker K, Wu CH (2005) Literature mining and database annotation of protein phosphorylation using a rule-based system. Bioinformatics 21: 2759–2765.
30. Yuan X, Hu ZZ, Wu HT, Torii M, Narayanaswamy M, et al. (2006) An online literature mining tool for protein phosphorylation. Bioinformatics 22: 1668–1669. Available: http://pir.georgetown.edu/pirwww/iprolink/rlimsp.shtml.
31. Tsuruoka Y, Tsujii J, Ananiadou S (2008) FACTA: a text search engine for finding associated biomedical concepts. Bioinformatics 24: 2559–2560. Available: http://text0.mib.man.ac.uk/software/facta/.
32. Hu Y, Hines LM, Weng H, Zuo D, Rivera M, et al. (2003) Analysis of genomic and proteomic data using advanced literature mining. J Proteome Res 2: 405–412. Available: http://medgene.med.harvard.edu/MEDGENE/.
33. Rolfs A, Hu Y, Ebert L, Hoffmann D, Zuo D, et al. (2008) A biomedically enriched collection of 7000 human ORF clones. PLoS ONE 3: e1528. Available: http://biogene.med.harvard.edu/BIOGENE/.
34. Aerts S, Lambrechts D, Maity S, Van Loo P, Coessens B, et al. (2006) Gene prioritization through genomic data fusion. Nat Biotechnol 24: 537–544. Available: http://homes.esat.kuleuven.be/~bioiuser/endeavour/endeavour.php.
35. Perez-Iratxeta C, Wjst M, Bork P, Andrade MA (2005) G2D: a tool for mining genes associated with disease. BMC Genet 6: 45.
36. Cheng D, Knox C, Young N, Stothard P, Damaraju S, et al. (2008) PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites. Nucleic Acids Res 36: W399–W405. Available: http://wishart.biology.ualberta.ca/polysearch/index.htm.
37. Jelier R, Schuemie MJ, Veldhoven A, Dorssers LC, Jenster G, et al. (2008) Anni 2.0: a multipurpose text-mining tool for the life sciences. Genome Biol 9: R96. Available: http://www.biosemantics.org/index.php?page = anni-2-0.
38. Hoffmann R, Valencia A (2004) A gene network for navigating the literature. Nat Genet 36: 664. Available: http://www.ihop-net.org/.
39. Lee H, Yi GS, Park JC (2008) E3Miner: a text mining tool for ubiquitin-protein ligases. Nucleic Acids Res 36: W416–W422. Available: http://e3miner.biopathway.org.
40. Swanson DR (1988) Migraine and magnesium: eleven neglected connections. Perspect Biol Med 31: 526–557.
41. Weeber M, Kors JA, Mons B (2005) Online tools to support literature-based discovery in the life sciences. Brief Bioinform 6: 277–286.
42. Smalheiser NR, Torvik VI, Zhou W (2009) Arrowsmith two-node search interface: a tutorial on finding meaningful links between two disparate sets of articles in MEDLINE. Comput Meth Program Biomed 94: 190–197. Available: http://arrowsmith.psych.uic.edu /cgi-bin /arrowsmith_uic/start.cgi.
43. Caporaso JG, Deshpande N, Fink JL, Bourne PE, Cohen KB, et al. (2008) Intrinsic evaluation of text mining tools may not predict performance on realistic tasks. Pac Symp Biocomput. pp 640–651.
44. Zanzoni A, Montecchi-Palazzi L, Quondam M, Ausiello G, Helmer-Citterich M, et al. (2002) MINT: a Molecular INTeraction database. FEBS Lett 513: 135–140.
45. Marcotte EM, Xenarios I, Eisenberg D (2001) Mining literature for protein-protein interactions. Bioinformatics 17: 359–363.
46. Donaldson I, Martin J, de Bruijn B, Wolting C, Lay V, et al. (2003) PreBIND and Textomy–mining the biomedical literature for protein-protein interactions using a support vector machine. BMC Bioinformatics 4: 11.
47. Rodriguez-Esteban R, Iossifov I, Rzhetsky A (2006) Imitating manual curation of text-mined facts in biomedicine. PLoS Comput Biol 2: e118. doi:10.1371/journal.pcbi.0020118.
48. Wadman M (2009) Open-access policy flourishes at NIH. Nature 458: 690–691.
49. Vastag B (2000) NIH launches PubMed Central. J Natl Cancer Inst 92: 374. Available: http://www.ncbi.nlm.nih.gov/pmc/.
50. Hearst MA, Divoli A, Guturu H, Ksikes A, Nakov P, et al. (2007) BioText Search Engine: beyond abstract search. Bioinformatics 23: 2196–2197. Available: http://biosearch.berkeley.edu/.
51. Xu S, McCusker J, Krauthammer M (2008) Yale Image Finder (YIF): a new search engine for retrieving biomedical images. Bioinformatics 24: 1968–1970. Available: http://krauthammerlab.med.yale.edu/imagefinder/.
52. Rodriguez-Esteban R, Iossifov I (2009) Figure mining for biomedical research. Bioinformatics 25: 2082–2084.
53. Yandell MD, Majoros WH (2002) Genomics and natural language processing. Nat Rev Genet 3: 601–610.
54. Koussounadis A, Redfern OC, Jones DT (2009) Improving classification in protein structure databases using text mining. BMC Bioinformatics 10: 129.
55. Pandev G, Kumar V, Steinbach M (2006) Computational approaches for protein function prediction: a survey. Technical Report 06-028, Department of Computer Science and Engineering, University of Minnesota, Twin Cities.
56. Manning CD, Schutze H (1999) Foundations of Statistical Natural Language Processing MIT Press.
57. Van Rijsbergen CJ, Robertson SE, Porter MF (1980) New models in probabilistic information retrieval. Tech. Rep. 5587. British Library. Available: http://tartarus.org/~martin/PorterStemmer/.
58. Krallinger M, Valencia A (2005) Text-mining and information-retrieval services for molecular biology. Genome Biol 6: 224.
59. Shatkay H, Höglund A, Brady S, Blum T, Dönnes P, et al. (2007) SherLoc: high-accuracy prediction of protein subcellular localization by integrating text and protein sequence data. Bioinformatics 23: 1410–1417. Available: http://www-bs.informatik.uni-tuebingen.de/Services/SherLoc2/.
60. Brady S, Shatkay H (2008) EpiLoc: a (working) text-based system for predicting protein subcellular location. Pac Symp Biocomput. pp 604–615. Available: http://epiloc.cs.queensu.ca/.
61. Glenisson P, Coessens B, Van Vooren S, Mathys J, Moreau Y, et al. (2004) TXTGate: profiling gene groups with text-based information. Genome Biol 5: R43. Available: http://tomcat.esat.kuleuven.be/txtgate/.
62. Krallinger M, Hirschman L, Valencia A (2008) Linking genes to literature: text mining, information extraction, and retrieval applications for biology. Genome Biol 9: S8. Available: http://zope.bioinfo.cnio.es/bionlp_tools/.
63. Kano Y, Baumgartner WA Jr, McCrohon L, Ananiadou S, Cohen KB, et al. (2009) U-Compare: share and compare text mining tools with UIMA. Bioinformatics 25: 1997–1998. Available: http://u-compare.org/.
64. Ramialison M, Bajoghli B, Aghaallaei N, Ettwiller L, Gaudan S, et al. (2008) Rapid identification of PAX2/5/8 direct downstream targets in the otic vesicle by combinatorial use of bioinformatics tools. Genome Biol 9: R145.
65. Natarajan J, Berrar D, Dubitzky W, Hack C, Zhang Y, et al. (2006) Text mining of full-text journal articles combined with gene expression analysis reveals a relationship between sphingosine-1-phosphate and invasiveness of a glioblastoma cell line. BMC Bioinformatics 7: 373.
66. Leach SM, Tipney H, Feng W, Baumgartner WA, Kasliwal P, et al. (2009) Biomedical discovery acceleration, with applications to craniofacial development. PLoS Comput Biol 5: e1000215. doi:10.1371/journal.pcbi.1000215.
67. Campillos M, Kuhn M, Gavin AC, Jensen LJ, Bork P (2008) Drug target identification using side-effect similarity. Science 321: 263–266.
68. Leitner F, Krallinger M, Rodriguez-Penagos C, Hakenberg J, Plake C, et al. (2008) Introducing meta-services for biomedical information extraction. Genome Biol 9 Suppl 2: S6. Available: http://bcms.bioinfo.cnio.es/.
69. Fernández JM, Hoffmann R, Valencia A (2007) iHOP web services. Nucleic Acids Res 35(Web Server issue). pp W21–W26.
70. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, et al. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Research 13: 2498–2504. Available: http://www.cytoscape.org/.
71. Wilbur WJ, Rzhetsky A, Shatkay H (2006) New directions in biomedical text annotation: definitions, guidelines and corpus construction. BMC Bioinformatics 7: 356.
72. Rzhetsky A, Zheng T, Weinreb C (2006) Self-correcting maps of molecular pathways. PLoS One 1: e61. doi:10.1371/journal.pone.0000061.
73. Jenssen TK, Laegreid A, Komorowski J, Hovig E (2001) A literature network of human genes for high-throughput analysis of gene expression. Nat Genet 28: 21–28.