Text Mining for Metabolic Pathways, Signaling Cascades, and Protein Networks

Robert Hoffmann,[1] Martin Krallinger,[1] Eduardo Andres,[1] Javier Tamames,[2] Christian Blaschke,[2] Alfonso Valencia[1]*

(Published 10 May 2005)

Introduction

Databases and repositories containing information on molecular interactions, metabolic pathways, and signaling cascades face a number of challenges, including intuitive exploration and visualization. In this context, cross-linking of database information to the original sources (that is, direct references to the literature) is a key issue. In the past decade, text-mining and information-extraction methods have been developed in response to some of these challenges.

Database annotators and domain experts use text-mining tools to retrieve relevant content from text repositories and to filter for facts of potential biological relevance. Most important, the use of text-mining methods can standardize the curation process for databases. In the near future, we can imagine that novel approaches will be able to summarize complex information and to handle and update facts annotated in biological databases by finding their relationships to information stored in heterogeneous text sources. Apart from database curators, biologists and biomedical researchers in general will also increasingly benefit from the use of text-mining systems to access and extract information, reproduce the reasoning behind database information, and ultimately assist researchers in generating novel hypotheses and models.

Here, we review the present state of the art in text mining and describe the main technical and scientific bottlenecks to the extraction of information from biomedical texts and the developments that are currently available (Table 1).

	Basic features	URL
Repositories		
PubMed/Entrez	Biomedical citation retrieval system	www.ncbi.nlm.nih.gov/entrez
GENIA corpus	Annotated corpus related to human blood cell transcription factors	www-tsujii.is.s.u-tokyo.ac.jp/GENIA/
BioCreative corpus	Corpus of protein annotation relevant text passages	www.pdg.cnb.uam.es/BioLINK/BioCreative.eval.html
Assessments		
BioCreative challenge	Text mining of protein names and annotations	www.pdg.cnb.uam.es/BioLINK/BioCreative.eval.html
KDD challenge	Information extraction of Drosophila gene expression information	www.biostat.wisc.edu/~craven/kddcup/tasks.html
TREC Genomics track	IR, document classification and question answering in biology domain	ir.ohsu.edu/genomics/
NLPBA challenge	Protein and gene name identification	www.genisis.ch/~natlang/JNLPBA04
Information retrieval		
PubMed/Entrez	Biomedical literature retrieval tool	http://www.ncbi.nlm.nih.gov/entrez
XplorMed	Iterative retrieval and extraction of abstracts	www.bork.embl-heidelberg.de/xplormed/
Google Scholar	Scholar literature search engine	scholar.google.com
CrossRef search	Full content search engine	www.crossref.org/crossrefsearch.html
Name recognition		
AbGene	Protein/gene name tagger	ftp://ftp.ncbi.nlm.nih.gov/pub/tanabe
GAPSCORE	Protein/gene name tagger	bionlp.stanford.edu/gapscore
NLProt	Protein/gene name tagger	cubic.bioc.columbia.edu/services/nlprot
Protein (set) function		
PubGene	Text mining tool for microarrays	www.pubgene.org
MedMiner	Extract gene relevant sentences	discover.nci.nih.gov/textmining/main.jsp
iProLINK	Protein annotation and tagging	pir.georgetown.edu/iprolink
Textpresso	*C. elegans* literature IR/IE tool	medblast.sibsnet.org
KAT	Annotate proteins from scientific references	www.bork.embl-heidelberg.de/kat
Protein interactions		
Chilibot	Relationship extraction tool	www.chilibot.net
GeneScene	IE of regulatory pathways	econport.arizona.edu:8080/NetVis/index.html
PreBIND	Classifier of protein interaction documents	bind.ca
Protein network exploration		
iHOP	Literature-based gene and protein network	www.pdg.cnb.uam.es/UniPub/iHOP/
Knowledge discovery		
ARROWSMITH	Extended MEDLINE search tool	kiwi.uchicago.edu
BITOLA	Literature-based biomedical discovery system	www.mf.uni-lj.si/bitola

Table 1. The main text-mining repositories and systems that are currently available.

Complex Nature of Gene and Protein Names

Text-mining techniques depend on the correct identification of entities such as protein and gene names, chemical compounds, and diseases. This basic step, however, has turned out to be extremely difficult, because the biomedical literature is flooded

[1]Protein Design Group, National Center for Biotechnology, CNB-CSIC, Darwin 3, Cantoblanco, 28049 Madrid, Spain. [2]BioAlmai, Tres Cantos, Madrid, Spain.

*Corresponding author. E-mail: valencia@cnb.uam.es

with short names, acronyms, gene and protein synonyms, and names with multiple meanings (homonyms). Distinguishing between specific protein and general protein family names is another serious difficulty that complicates the mapping between names and their corresponding biological database entries (such as protein or nucleic acid sequences).

According to the recent BioCreative assessment (*1*), the best systems available can only recover 80% (recall) of the protein or gene names in biomedical text with an accuracy of about 80% (precision). Indeed, the problem of entity identification in biology has been found to be harder than the identification of names in areas such as economics or news wire services.

Despite all efforts to assemble dictionaries and establish nomenclature standards, official gene names still do not provide a solution to the problem of name detection. In 1994, only 36% of the human genes were mentioned by their official names according to the Human Genome Organization (HUGO) nomenclature, and by 2004 this percentage had increased only to about 43% (*2*). It seems that the dynamics of synonym creation and usage are as vigorous as the evolution of genes and proteins (*3*), so static nomenclatures and dictionaries will always lag behind. Thus, community efforts to establish a standard vocabulary will probably not succeed unless publishers decide to enforce it, as they have done with the standard deposition of sequences, structures, and expression profiles.

Interactions Between Proteins and Genes

The extraction of associations between proteins is the first logical step toward the reconstruction of biological pathways. The underlying assumption is that protein names tend to appear together within a given text segment if they display a biological relationship (*4*). This task has attracted considerable attention during the past decade, but it has turned out to be more difficult than anticipated (*5*). Current techniques use sentences as the basic context for co-occurrence analysis (*6*), although some analyze whole abstracts (*7*) or passages extracted from full-text articles (*8*).

The next problem is the characterization of the biological significance of the interactions and the classification of the interactions into biologically meaningful groups or types. Blaschke *et al*. (*9*, *10*) proposed the use of a controlled set of expressions or frames to classify the various biological relations between proteins and genes. These frames were expressions of the type: "complex of protein x and protein y," "phosphorylation of protein x by protein y," or "protein x binds protein y." Other systems, such as GENIES (*11*), use basic natural language processing techniques for the classification of the type of interactions between proteins. The differences between using a controlled set of expressions and natural language processing and learning techniques result in greater precision in the first case, because only a small number of predefined frames is used, and in greater recall in the second case, because it allows the automatic discovery of new association expressions or verbs (*12*, *13*). The problem of organizing, summarizing, and presenting this information in a biologically meaningful manner still remains a major difficulty for all of these systems.

Some of the strategies for the detection of co-occurrences have been applied to the detection of indirect relations. This process of inference is equivalent to the one used for the classical discovery of magnesium for the treatment of migraine, based on sentences such as "magnesium loss can have an effect on stress," "stress is associated with migraines," "magnesium is a natural calcium channel blocker," and "channel blockers prevent some migraines" (*14*), or for the relation of Raynaud's disease to dietary fish oil (*15*).

The first systematic exploration of this strategy showed that neighboring genes in the literature network of a given year will have a higher chance of being mentioned together explicitly in the year after the first publication years [0.06% of all genes with a network distance of two steps, but only 0.01% of all genes with a distance of four steps, were subsequently mentioned together (*16*)]. This means that the interaction network can be used to predict new biologically meaningful relations in the absence of a direct textual connection. This approach yielded the proposed relationship of Ntc20 and Ntc30 as part of the spliceosome (*17*).

The Special Case of Metabolic Pathways

The discovery of relations between enzymes in metabolic pathways faces specific problems, because enzymes acting in successive steps of a reaction are rarely mentioned together within the same text passage. Exploration of the main metabolic databases [EcoCyc (*18*) and KEGG (*19*)] showed that only 26% of the successive steps in pathways correspond to proteins that co-occur in PubMed abstracts, whereas in 44% of cases, the information about successive steps can be deduced through intermediate chemical compounds that are the product of one reaction and a substrate in the following one (*20*).

Two observations are remarkable: (i) Half of the information in these metabolic databases cannot be automatically traced to its origin in papers, and (ii) there is additional information not currently contained in the metabolic databases that can be extracted by automatic screening of PubMed abstracts, such as propionate metabolism and threonine anaerobic degradation. These observations again point to the need to incorporate text-mining tools in the process of database annotation to guarantee the connection between database annotations and the original sources of information [see (*21*) for a description of the origin of these problems].

From Chains of Interactions to Networks

The description of proteins connected in ordered pathways can be extended to a general model of connectivity in a protein network, using the type of representation made familiar by the high-throughput proteomics experiments.

Our group has developed a freely accessible system, called iHOP (Information Hyperlinked over Proteins), which provides a network of genes and proteins that co-occur in the PubMed biomedical literature (*22*). Navigation across interrelated sentences within this network is closer to human intuition than conventional keyword searches and allows for stepwise and controlled acquisition of information. Additionally, iHOP is beginning to provide direct links to the IntAct (*23*) protein interaction database and allow for the superimposition of external experimental information onto the textual network. In this way, it becomes possible to explore novel and existing knowledge simultaneously.

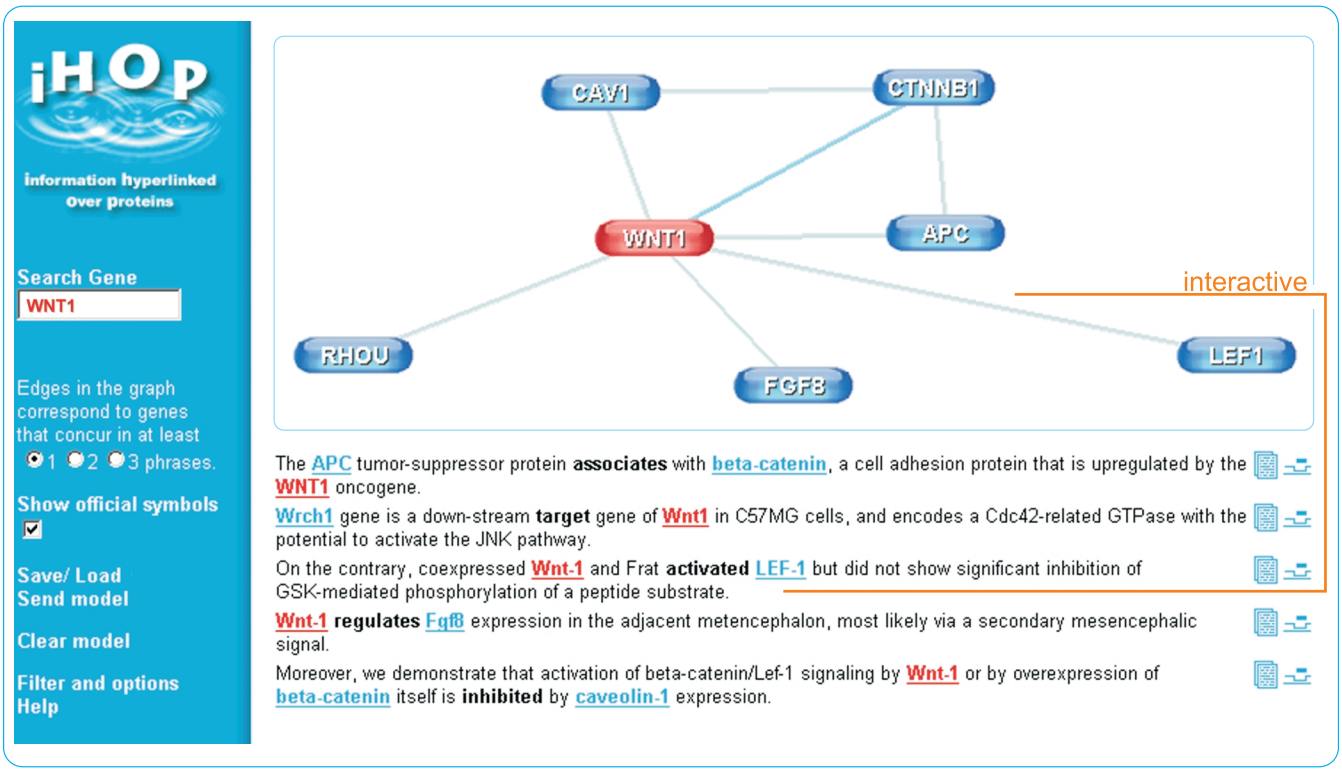

Fig. 1. In the course of navigation through the iHOP system, interesting sentences can be collected into a logbook or gene model and are dynamically represented as a graph. This graph represents the condensed result of a literature search but also remains hyperlinked to the corresponding sentences. In this way, users can familiarize themselves with the newly acquired information in an interactive manner and further extend the model. The iHOP server is publicly accessible at www.pdg.cnb.uam.es/UniPub/iHOP/.

Next Challenge: Discovery and Generalization

Pathways and networks are more than a set of connected entities, because they carry out a common biological function for which they have been selected in evolution. The methods for detecting general functions are still missing from the current systems for detecting protein interactions and exploring protein networks described above.

The problem of detecting the function common to proteins participating in pathways and cascades is to some extent equivalent to the problem of characterizing the function common to a group of genes with similar expression patterns (*24*). Two factors complicate this issue: (i) A common function could be described in many different ways by different authors [for example, the Gene Ontology (GO) (*25*) concept "cytokinesis" can be mentioned as "cell division" in free text (BioCreative citation 1)], or (ii) a common function has yet to be described for the proteins or genes of interest. Initial systems were based on functional keywords (*26*), although the development of ontologies, such as GO, provided a more useful source of biological concepts for the annotation of groups of genes (*27*). The most obvious approach to this problem is to link the heterogeneous textual information collected for each protein to the concepts that build up those ontologies. The results of the protein annotation extraction task of BioCreative (*28*) show the limitations of current approaches to mapping functional descriptions detected in text to the corresponding classes in the GO ontology. Only about half of the text passages containing information to annotate could be retrieved by the participating teams. The main difficulty encountered by those systems was the large number of ways in which to formulate functional terms in texts, the lack of a high-quality training set, and the lack of natural language–like synonyms for those terms in GO. Therefore, the best solution that current text-mining methods can offer is to integrate information extraction techniques with the expertise of the users in the process of exploration of the literature sources of pathways and networks (Fig. 1).

References

1. A. Yeh, A. Morgan, M. Colosimo, L. Hirschman, *BioCreAtIvE Task 1A: Gene Mention Finding Evaluation* (MITRE Corporation, Bedford, MA, 2005).
2. J. Tamames *et al.*, unpublished data.
3. R. Hoffmann, A. Valencia, Life cycles of successful genes. *Trends Genet.* **19**, 79–81 (2003).
4. B. J. Stapley, G. Benoit, Biobibliometrics: Information retrieval and visualization from co-occurrences of gene names in Medline abstracts. *Pac. Symp. Biocomput.* **2000**, 529–540 (2000).

5. C. Blaschke, L. Hirschman, A. Valencia, Information extraction in molecular biology. *Brief. Bioinf.* **3**, 154–165 (2002).
6. T. Ono, H. Hishigaki, A. Tanigami, T. Takagi, Automated extraction of information on protein-protein interactions from the biological literature. *Bioinformatics* **17**, 155–161 (2001).
7. T. K. Jenssen, A. Laegreid, J. Komorowski, E. Hovig, A literature network of human genes for high-throughput analysis of gene expression. *Nat. Genet.* **28**, 21–28 (2001).
8. M. Krallinger, M. Padron, A. Valencia, A sentence sliding window approach to extract protein annotations from biomedical articles. *BMC Bioinf.* in press.
9. C. Blaschke, M. A. Andrade, C. Ouzounis, A. Valencia, Automatic extraction of biological information from scientific text: Protein-protein interactions. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 60–67 (1999).
10. C. Blaschke, A. Valencia, The frame-based module of the Suiseki information extraction system. *IEEE Intell. Syst.* **17**, 14–20 (2002).
11. C. Friedman, P. Kra, H. Yu, M. Krauthammer, A. Rzhetsky, GENIES: A natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics* **17**, S74–S82 (2001).
12. V. Hatzivassiloglou, W. Weng, Learning anchor verbs for biological interaction patterns from published text articles. *Int. J. Med. Inf.* **67**, 19–32 (2002).
13. T. Sekimizu, H. S. Park, J. Tsujii, Identifying the interaction between genes and gene products based on frequently seen verbs in Medline abstracts. *Genome Inf. Ser. Workshop Genome Inf.* **9** 62–71 (1998).
14. D. R. Swanson, Migraine and magnesium: Eleven neglected connections. *Perspect. Biol. Med.* **31**, 526–557 (1988).
15. N. R. Smalheiser, D. R. Swanson, Using ARROWSMITH: A computer-assisted approach to formulating and assessing scientific hypotheses. *Comput. Methods Programs Biomed.* **57**, 149–153 (1998).
16. C. Blaschke, A. Valencia, unpublished data.
17. C. Blaschke, A. Valencia, The potential use of SUISEKI as a protein interaction discovery tool. *Genome Inf. Ser. Workshop Genome Inf.* **12**, 123–134 (2001).
18. M. Kanehisa S. Goto, S. Kawashima, Y. Okuno, M. Hattori, The KEGG resource for deciphering the genome. *Nucleic Acids Res.* **32**, D277–D280 (2004).
19. I. M. Keseler, J. Collado-Vides, S. Gama-Castro, J. Ingraham, S. Paley, I. T. Paulsen, M. Peralta-Gil, P. D. Karp, EcoCyc: A comprehensive database resource for *Escherichia coli*. *Nucleic Acids Res.* **33**, D334–D337 (2005).
20. E. Andres *et al.*, in preparation.
21. A. Valencia, Search and retrieve. Large-scale data generation is becoming increasingly important in biological research. But how good are the tools to make sense of the data? *EMBO Rep.* **3**, 396–400 (2002).
22. R. Hoffmann, A. Valencia, A gene network for navigating the literature. *Nat. Genet.* **36**, 664–664 (2004).
23. H. Hermjakob, L. Montecchi-Palazzi, C. Lewington, S. Mudali, S. Kerrien, S. Orchard, M. Vingron, B. Roechert, P. Roepstorff, A. Valencia, H. Margalit, J. Armstrong, A. Bairoch, G. Cesareni, D. Sherman, R. Apweiler, IntAct: An open source molecular interaction database. *Nucleic Acids Res.* **32**, D452–D455 (2004).
24. C. Blaschke, J. C. Oliveros, A. Valencia, Mining functional information associated with expression arrays. *Funct. Integr. Genom.* **1**, 256–268 (2001).
25. Gene Ontology Consortium, The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* **32**, D258–D261 (2004).
26. M. A. Andrade, A. Valencia, Automatic annotation for biological sequences by extraction of keywords from MEDLINE abstracts. Development of a prototype system. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **5**, 25–32 (1997).
27. F. Al-Shahrour, R. Diaz-Uriarte, J. Dopazo, FatiGO: A web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics* **20**, 578–580 (2004).
28. C. Blaschke, L. E. Andres, M. Krallinger, A. Valencia, Evaluation of Bio Creative assessment of task 2. *BMC Bioinf.*, in press.
29. Many thanks to L. Hirschman and A. S. Yeh (MITRE Corporation) for their efforts in organizing BioCreative (task 1). The work of our group described here was in part supported by grants from the European Commission (ORIEL IST-2001-32688, TEMBLOR QLRT-2001-00015, and Biosapiens LSHC-CT-2003-505265), by a European Molecular Biology Organization grant for the organization of BioCreative, and by a research contract between Consejo Superior Investigaciones and BioAlma.

Citation: R. Hoffmann, M. Krallinger, E. Andres, J. Tamames, C. Blaschke, A. Valencia, Text mining for metabolic pathways, signaling cascades, and protein networks. *Sci. STKE* **2005**, pe21 (2005).