Briefings in Bioinformatics Advance Access published October 10, 2006

doi:10.1093/bib/bbl037

# Mining literature for systems biology

*Phoebe M Roberts*

Submitted: 2nd June 2006; Received (in revised form): 6th September 2006

## Abstract

Currently, literature is integrated in systems biology studies in three ways. Hand-curated pathways have been sufficient for assembling models in numerous studies. Second, literature is frequently accessed in a derived form, such as the concepts represented by the Medical Subject Headings (MeSH) and Gene Ontologies (GO), or functional relationships captured in protein–protein interaction (PPI) databases; both of these are convenient, consistent reductions of more complex concepts expressed as free text in the literature. Moreover, their contents are easily integrated into computational processes required for dealing with large data sets. Last, mining text directly for specific types of information is on the rise as text analytics methods become more accurate and accessible. These uses of literature, specifically manual curation, derived concepts captured in ontologies and databases, and indirect and direct application of text mining, will be discussed as they pertain to systems biology.

**Keywords:** *Medical Subject Headings; gene ontology; interactome; pathways; networks*

## INTRODUCTION

Two unprecedented phenomena are occurring in the biomedical space: the increasing rate at which papers are published [1,2], and the genome-wide experimental coverage made possible by the advent of high-throughput technologies, including yeast two-hybrid, microarray and mass spectrometry [3]. To stay abreast of one's field of expertise, or to analyze the results from high-throughput data sets, it is essential to tap into the ever-growing repository of existing knowledge that is housed in the scientific literature. To do so in an efficient way requires methods that can reduce complexity without compromising the integrity of published data. A central aim of text analytics is to transform scientific literature from free-form densely written articles into high quality structured data from which proteins, diseases and species and their inter-relationships are easily accessed (for recent reviews, see [4,5]). To conduct such a transformation requires identifying relevant literature, which is known as information retrieval or text classification. Classes of related terms, or entities, such as proteins, diseases or tissues, are extracted from the text. Relationships among entities are extracted for mining or quantification. Performing these tasks manually does not scale well for analysis of results generated by high-throughput approaches like microarrays or two-hybrid analysis. High-volume literature analysis requires the technologies collectively known as text analytics.

Just as text analytics aims to manage complexity of the literature, systems biology attempts to manage biological complexity by considering integrated pathways previously studied only in isolation, thereby better representing what happens *in vivo* [6]. The 'four M's', measurement, mining, modeling and manipulation, describe the series of events that comprise a systematic study. A system is *manipulated* by perturbation, the effects are *measured* using high-throughput methods, the data are *mined* and *modeled* to create a new hypothetical system, which is then subjected to further manipulation to test emergent hypotheses [7]. As the source of meaningful concepts familiar to the biologist, the scientific literature is an essential element of mining and modeling. Literature-derived data imparts a descriptive aspect not inherent to sequence identification numbers or obscure official gene names.

High-throughput methods have been in place long enough to raise concern whether one, or even two, large-scale experiments directed at the same problem sufficiently distinguish rare events from false positives. There is mounting evidence that integrating several large-scale data sets can provide answers

Corresponding author. Biogen Idec, Inc., 14 Cambridge Center, Cambridge MA 02142, USA. Tel: 617-914-7033; Fax: 617-679-2306; E-mail: Phoebe.Roberts@biogenidec.com

**Dr Roberts** is a scientist with the Biogen Idec Library, where she provides text analytics support for research projects. She received her PhD in molecular biology from the University of Chicago.

that are drowned out by noise in the absence of multiple experimental approaches. For instance, candidate genes for several mitochondrial disorders were identified by combining no less than eight genome-wide data sets, when traditional methods would have required sequencing megabases of DNA to pinpoint candidate loci [8]. In another example, hypotheses generated by integrating high-throughput data sets and information from literature-derived databases yielded new insights into yeast galactose utilization not evident after decades of scrutiny [9].

Text analytics can be thought of as another high-throughput weapon in the system biologist's arsenal, with all the power and complications that such approaches afford. The challenge of interpreting any high-throughput data set is distinguishing signals from noise [10,11]. This is as true of text mining data sets as it is of transcript profiling or yeast two-hybrid analysis. In a microarray experiment, there are genes whose expression undergoes dramatic changes in response to perturbation, and many more genes that are expressed at or around the level of background. Text mining analyses also have strong and weak signals, and they can depend on, for example, the number of times a concept is stated [12] or how simply it is phrased in the literature [13,14]. Nonetheless, transforming the unstructured information in scientific publications into structured data, that can be mapped to gene identifiers and other well-defined entities, is essential for reaping the rewards of the millions of person-years that went into generating the published literature available today.

This review covers how free text has been structured for use in high-throughput studies, and how literature is used in systems biology. Text analytics are currently used to generate resources that support systems biology studies, but they are rarely applied directly. Based on current methods of literature use, text mining technologies and resources underutilized in systems biology are recommended, as are improvements to text analytics to encourage more widespread use in systems-wide studies.

## STRUCTURING THE LITERATURE: ONTOLOGIES, THESAURI AND DATABASES

For systems biologists to mine the literature effectively, they must be able to retrieve relevant articles using commonly used keywords that unify related entities. An obvious way to homogenize related concepts is to flag them with the same standardized term. If the standardized terms are organized in a hierarchical structure that reflects known relationships, even greater understanding is conveyed by standardization. An ontology is a set of terms organized to define their relationships, creating a formal representation of knowledge [15]. As such, it is the logical solution to transform biology into a machine-readable depiction of life as we know it. Term collections less structured than ontologies include taxonomies, controlled vocabularies, thesauri and dictionaries. Differences among these are subtle and will collectively be referred to as terminologies [16]. A terminology's utility in information retrieval and entity extraction is in part dependent on the synonyms attached to principle terms. MeSH (Medical Subject Headings) is a 24 000 term terminology developed by the National Library of Medicine (NLM) to structure the seventeen million records in MEDLINE, and it is frequently used in text analytics for information retrieval and mining relationships [17,18]. Curators at the NLM review each article entering MEDLINE, then select terms from the MeSH hierarchy that best capture the aims of the article, thereby summarizing it using a set of controlled vocabulary terms. Curating citations with MeSH terms facilitates searching by unifying similar concepts and by disclosing additional information not mentioned specifically in the title or abstract. Numerous studies illustrate the power of MeSH in information retrieval, citing improvements in efficiency and recall [19,20]. Accurate selection of annotation terms from large terminologies requires a well-defined ontology, explicit instructions for curation and domain expertise [21,22].

The MeSH terminology has some coverage of how proteins function in cellular systems, but it is not exhaustive. To more accurately capture what happens at the cellular and molecular level, the Gene Ontology (GO) was developed [23]. Originally founded by three groups who supported species-specific databases, the GO Consortium has expanded to represent fourteen organisms with 14 000 terms that describe biological process (BP), molecular function (MF) and cellular component (CC) [22]. Just as MeSH terms are assigned to individual scientific articles to describe their content, GO terms are assigned to proteins to illustrate what they do (MF), where they do it (CC) and to what end (BP).

**Table 1:** A non-exhaustive list of ontologies, terminologies and their uses

| Ontology/Terminology | Used to Annotate | URL | Reference |
|---|---|---|---|
| MeSH[a] | scientific publications | http://www.nlm.nih.gov/mesh/meshhome.html | [24] |
| GO[a] | proteins | http://www.geneontology.org/ | [22] |
| IUBMB EC numbers | enzymes | http://www.chem.qmul.ac.uk/iubmb/enzyme/ | |
| COG | orthologous protein groups | http://www.ncbi.nlm.nih.gov/COG/new/ | [49] |
| SNOMED[a] | clinical phenomena | http://www.nlm.nih.gov/research/umls/sources.by.categories.html | [24] |
| NCI thesaurus[a] | oncological phenomena | http://www.nlm.nih.gov/research/umls/sources.by.categories.html | [24] |
| OMIM[a] | genetic disorders | http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM | [50] |
| NCBI Taxonomy[a] | species | http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Taxonomy | [33] |
| UWDA[a] | anatomy | http://www.nlm.nih.gov/research/umls/sources.by.categories.html | [24] |
| Sequence Ontology | sequences | http://song.sourceforge.net/ | [22] |
| MIPS FunCat | proteins | http://mips.gsf.de/projects/funcat | [51] |

[a] Part of UMLS (http://www.nlm.nih.gov/research/umls/umlsmain.html) (see text for details).

Table 1 lists ontologies and terminologies commonly used in biomedical research and, to a lesser extent, in systems biology. MeSH and GO have been merged with other terminologies to form the Unified Medical Language System (UMLS) [24]. To identify formal ontological relationships among terms in the UMLS, the Semantic Network was created, thereby adding additional computational accessibility to the free text annotated with UMLS-affiliated terminologies [16].

## TEXT MINING FOR BUILDING SYSTEMS BIOLOGY RESOURCES

Modeling biological systems usually includes methodology that is gene- or protein-centric, such as transcript profiling or proteomic analysis. Reliably identifying protein names in the literature is an active area of text mining research with a wide range of applications, including microarray analysis, biomarker discovery and database curation. Identifying genes and proteins in the literature is prohibitively difficult because of multiple names for a single gene, shared synonyms among genes and names that are common or biomedical terms [25,26]. There are efforts afoot to standardize gene nomenclature, including the HUGO Gene Nomenclature Committee for human gene names, but the scientific community has not adopted standardized terms; estimates of official human gene name use range from 18–44% of all mentions of human genes in a subset of PubMed abstracts [26,27]. Accurately identifying proteins presents a formidable barrier between information in the literature and automated interpretation of high-throughput data. Although text analytics has not solved the problem of protein identification, there is movement in the right direction. Strategies include assembling dictionaries of gene names and synonyms, expanding acronyms to differentiate among shared acronyms, and including terms like ‘gene’ or ‘protein’ as part of the search [5,25,28]. A comparison of five strategies to improve protein name recognition in MEDLINE shows that all five strategies are required for optimal resolution of relevant literature attributed to a unique genetic locus and its products [25].

Once entities like genes or pathways are recognized, automatically extracting how they are related is another active field in text analytics. Co-occurrence of two entities in a document is often used to identify relationships, but it has some limitations. Wren *et al.* [29] calculated that co-occurrence in an abstract reflects a real relationship 58% of the time, whereas if two terms are in the same sentence, they are related 83% of the time. The increase in accuracy comes at a cost to sensitivity, as 43% of true relationships would be missed by requiring terms to co-occur in a sentence [29]. Nonetheless, co-occurrence has been described as the least labor-intensive method of automatically linking concepts [5].

In the aforementioned co-occurrence example, sentence bounds were used to improve the odds that an automated relationship prediction would be correct. Text mining using natural language processing (NLP) not only uses sentence structure, but also employs parts of speech and phrase recognition to identify certain relationships among entities in a sentence (reviewed in [30]). The utility of overlaying part-of-speech tagging has been quantified [12,14], and it varies with the objective of the text mining exercise. In order to classify literature relevant to stem cell research, nouns found in relevant documents are more effective than verbs or adjectives at segregating stem cell articles from the rest of MEDLINE [12]. Reliably identifying gene-disease relationships is more effective for those related by verbs than by prepositions [14]. Co-occurrence was also an effective method of automatically detecting gene-disease relationships, but only if gene and disease terms were within three words of one another [14]. Together, these results show that using multiple methods with high precision and low recall can collectively improve the number of articles retrieved without sacrificing relevance to the topic in question.

Considering the challenges inherent to correctly identifying gene names and consistently applying GO terms to proteins, combining the two to automatically assign GO terms to proteins based on literature excerpts is a formidable task. The BioCreAtIvE challenge was established to provide a community forum for comparing strategies to automate this task [31]. Although correct identification of proteins in the literature was reasonably successful, accurate annotation of those proteins with a relevant GO term proved difficult. CC, MF and BP terms were assigned correctly 11, 9 and 7% of the time, respectively [31]. These results underscore the difficulty of fully automating extraction of information frequently used by systems biologists, and underscore the need to simplify manual intervention by a domain expert for rapid review of text mining data sets.

## USING ONTOLOGIES AND DATABASES IN SYSTEMS BIOLOGY

Genome sequencing has created nearly complete gene catalogs for several species, and databases like UniProt and Entrez house protein compendia as they relate to genomic sequence, diseases, orthologs, variants and other kinds of annotation [32,33]. These central clearinghouses of proteins are logical homes for functional annotation derived from the literature, such as GO terms. Because systems biology concerns the interplay among constituent parts, a list of parts, however well annotated, is often insufficient. This observation is supported by a search of the systems biology literature for resources from which

**Table 2:** Sources of pathway information in a non-exhaustive list of integrative systems biology studies

| Reference | Species | Databases[a] | Pattern (build or Validate) |
|---|---|---|---|
| [52] | Yeast | DIP | Validate |
| [8] | Human, mouse | SwissProt(Pfam), SGD(GO), OMIM | Validate |
| [34] | Yeast, worm, fly, human | DIP, MINT, BIND, MIPS, HPRD, OMIM, Flybase, Wormbase, MGD, Inparanoid, Custom | Build, validate |
| [9] | Yeast | DIP, BIND, SGD(GO) | Validate |
| [35] | Human | Custom, MGD(GO), DIP, MINT, BIND, MIPS, HPRD | Validate |
| [46] | Yeast | MIPS, SGD, Custom | Validate |
| [37] | Yeast, *e. coli*, *h. influenza*, *h. pylori* | KEGG(EC), MIPS(CYGD), SGD, Enzyme nomenclature, YPD, ERGO, Swissprot, Custom | Build |
| [40] | Worm | Wormbase(GO) | Validate |
| [53] | Many | N/A, Custom | Build |
| [38] | Worm | Custom | Build, validate |
| [54] | Yeast | Custom, Ensemble, SMART domains | Build |
| [55] | Human, mouse, rat, dog | TRANSFAC | Validate |
| [56] | Mouse | Custom | Validate |
| [39] | Human | Custom, db from [57] | Build |
| [45] | Mouse | Custom curated db | Validate |
| [58] | Human | Custom | Build |
| [36] | Yeast | GO, DIP, BIND | Validate |
| [59] | Fly | GO | Build |
| [60] | Human | Custom | Build |
| [61] | Yeast | YPD(GO), SGD(GO), Curagen interactome | Validate |

[a] See corresponding reference for database details. 'Custom' refers to unspecified method of curating information from the literature.

annotation is taken, some results of which are listed in Table 2. To compile this list, references reporting integration of at least two high-throughput approaches (not including literature-based approaches) yielding testable hypotheses were curated for their use of literature to mine or model the system in question. Combining literature mining (directly or with the use of ontologies) with a single high-throughput method is becoming commonplace and is reviewed elsewhere [1,5,34]. Table 2 shows frequent use of pathway and Protein–Protein Interaction (PPI) databases such as the Kyoto Encyclopedia of Genes and Genomes (KEGG), Biomolecular Interaction Network Database (BIND), the Molecular Interactions database (MINT), Database of Interaction Proteins (DIP), and the Munich Information Center for Protein Sequences (MIPS) Protein-Protein Interaction Database, which provide the next level of programmatically accessible information by formalizing relationships among uniquely identifiable constituents. The GO is also frequently mentioned. Computationally accessible annotation like GO allows one to ask, for instance, if physical interaction significantly correlates with MF, subcellular localization, or BP, thereby laying the groundwork for better algorithms to add predicted annotation to uncharacterized proteins [9,35,36].

Using the 21 references from Table 2, two patterns of literature usage were apparent. The ‘build’ pattern describes the use of literature curation to reconstruct well-studied pathways made up of known protein interactions and their effects. Curation methods and criteria for including or excluding proteins in the pathway are rarely specified, and authors frequently supplement existing databases with additional information from the literature [34,37]. High-throughput experiments are conducted, and new information in the form of relationships with uncharacterized proteins or uncharacterized relationships among known proteins is added and tested [38,39]. The ‘validate’ usage pattern involves building networks from integrated data sets, and consulting the literature (usually captured indirectly in PPI databases) to test if the model accurately reproduces well-characterized pathways [40]. In sum, existing resources frequently tapped in systems biology require additional information from the literature to address specific needs of the experimental design. In the following section, examples from Table 2 will be used to illustrate how text analytics can be used to efficiently and comprehensively mine relationships from text.

The Bimolecular Interaction Network Database (BIND) is a key resource in systems biology studies (Table 2). BIND has provided an explicit description of literature mining methods used to generate the database, including co-occurrence, sentence structure, text classification and gene dictionaries [41]. Abstracts that contain two and only two protein names in the same sentence are retrieved for curation. This stringent requirement accounts for some of the differences between annotation in BIND *vs* other interaction databases (for degree of overlap, see [34]), and highlights the challenge of capturing in an automated fashion the myriad ways the English language allows expression of similar concepts. To further increase relevance, presence of terms associated with PPI are also used to rank articles.

Additional efficiency in database creation is provided at the curation interface, including highlighting of key terms (e.g. protein names and interaction terms), as well as isolation of the relevant sentence, which may eliminate the need to read an entire abstract. BIND staff quantified a 70% reduction in curation time with a concomitant decrease in the amount of time spent curating each article [41]. Getting rid of the noise is not only economical and efficient, but also improves the quality of evaluation. A study of visual searches for rare items showed that there were optimal ratios of relevant to irrelevant events to obtain the highest precision without compromising recall [42]. Without text analytics in a database workflow, irrelevant results drown out relevant ones, increasing the odds that they will be overlooked.

In another example from Table 2, Forster *et al.* [37] enhance data from KEGG and other databases by extracting kinetic data from the literature. Text mining has been fine-tuned for this exercise by Hakenberg *et al.* [43], and they saw a five-fold improvement in precision using support vector machine classification of articles instead of simple keyword searches. Applying NLP to relevant articles allows the extraction of kinetic measurements, such as binding affinities and reaction rates, further enhancing the efficiency with which a researcher can mine literature for additional information.

## FUTURE DIRECTIONS

Systems biologists will continue to need ways of extracting information types from the literature

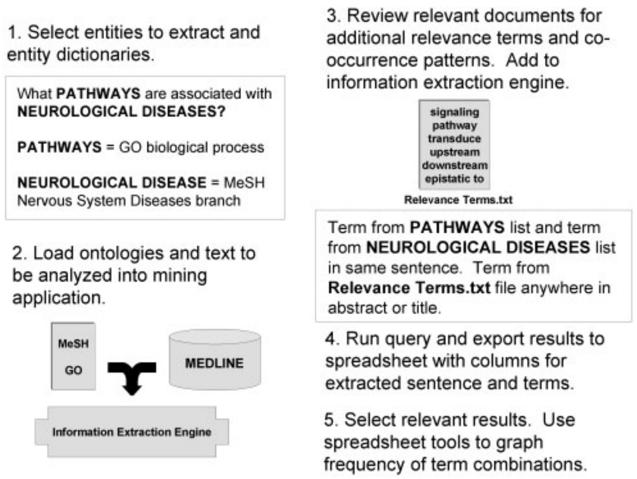
1. Select entities to extract and entity dictionaries.
What PATHWAYS are associated with NEUROLOGICAL DISEASES?
PATHWAYS = GO biological process
NEUROLOGICAL DISEASE = MeSH Nervous System Diseases branch
2. Load ontologies and text to be analyzed into mining application.
MeSH
GO
MEDLINE
Information Extraction Engine
3. Review relevant documents for additional relevance terms and co-occurrence patterns. Add to information extraction engine.
signaling
pathway
transduce
upstream
downstream
epistatic to
Relevance Terms.txt
Term from PATHWAYS list and term from NEUROLOGICAL DISEASES list in same sentence. Term from Relevance Terms.txt file anywhere in abstract or title.
4. Run query and export results to spreadsheet with columns for extracted sentence and terms.
5. Select relevant results. Use spreadsheet tools to graph frequency of term combinations.

**Figure 1:** Workflow for custom literature curation. Problems suitable for text mining are those in which search terms, like 'pathways', are placeholders for a list of related terms. Step 1 involves finding or building terminologies pertinent to the query. Step 2 uses an information extraction engine to search a corpus, such as the current version of MEDLINE (available from the National Library of Medicine) with the selected terminologies. Step 3 refines the query by adding additional terms pulled from relevant articles. Patterns of co-occurrence are also noted to determine how to combine terminologies at the phrase, sentence or abstract level. Step 4 produces results that are efficiently reviewed and prepare data for quantification and visualization in Step 5.

efficiently and comprehensively. Smaller research groups do not have the resources to assemble custom databases of high quality structured information. To improve literature access, one can make better use of existing resources, or develop new resources of wide utility. A conspicuously underutilized resource in systems biology is MeSH, whose coverage overlaps with GO and Online Mendelian Inheritance in Man (OMIM). It is straightforward computationally to link annotation from databases via unique citation identifiers. Using an example from Entrez Gene (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene, last viewed 1 June 2006), the record for LRRN6A cites one article in the Bibliography field. Three interacting proteins imported from the BIND database link to two additional articles. The three PubMed IDs corresponding to the articles are associated with 121 MeSH terms and substances. The UMLS MeSH to GO Mapper [44] yields three GO terms not present in the Entrez Gene record. The absence of MeSH in systems biology studies may reflect the extra steps required to reach MeSH terms from more familiar annotation. Alternatively, it may reflect magnification of errors generated by mapping from one terminology to another. MeSH annotation is extensive and relatively convenient, and will likely show up in future studies.

There will always be biological problems too domain-specific to be adequately supported by communal annotation projects. Half of the references in Table 2 incorporated some form of custom curation, either to create databases *de novo* [45], or to add information to existing databases [37,46]. Building a custom database for integration with other high-throughput data sets is becoming easier for the classically trained biologist. Figure 1 provides a protocol for isolating relevant information with features known to increase efficiency and reduce error, such as presenting results in a user-friendly format, highlighting search terms and isolating relevant text [47]. Resources like the ones shown in Table 1 are readily available for setting up a curation effort. Corpora (the text collection to be mined) like MEDLINE or full text from PubMed Central can be licensed and installed locally. Information extraction systems that transform free text into structured data are commercially available and sufficiently sophisticated to accommodate biomedical complexity [47,48]. The optimal system allows a domain expert to perform iterative searches with minimal programmatic expertise.

## CONCLUSIONS

The scientific literature must be in a computationally accessible format to be useful for systems biology studies.

High-throughput technologies have helped spur the development of databases to house PPI, knockout phenotype, subcellular localization, MF and BP data curated from the literature. These information types can be linked to specific genes and proteins, which act as intermediaries between literature-derived annotation and pan-genomic or proteomic data. Custom curation is frequently employed to meet the needs of systems biologists. Text analytics speeds creation of custom annotation by as much as an order of magnitude [47], thereby lowering the barrier to accessing the wealth of information in the scientific literature.

**Key Point**

- Systems biology uses the published literature to build or validate modeled protein networks.
- Protein-protein interaction (PPI) databases, the Gene Ontology, and manual curation are frequently used for systematic biological studies.
- Text analytics are indirectly used in systems biology to create databases or perform quality control of ontologies,
- Using text mining to structure literature into databases increases curation efficiency and content accuracy.
- Automated entity and relationship extraction can be thought of as the high-throughput data set that connects to domain knowledge.

## *References*

1. Scherf M, Epple A, Werner T. The next generation of literature analysis: integration of genomic analysis into text mining. *Brief Bioinform* 2005;**6**:287–97.
2. Tanabe L, Wilbur WJ. Tagging gene and protein names in biomedical text. *Bioinformatics* 2002;**18**:1124–32.
3. Xia Y, Yu H, Jansen R, *et al*. Analyzing cellular biochemistry in terms of molecular networks. *Annu Rev Biochem* 2004;**73**:1051–87.
4. Cohen AM, Hersh WR. A survey of current work in biomedical text mining. *Brief Bioinform* 2005;**6**:57–71.
5. Erhardt RA, Schneider R, Blaschke C. Status of text-mining techniques applied to biomedical text. *Drug Discov Today* 2006;**11**:315–25.
6. Henry CM. Systems biology. *Chem Eng News* 2003;**81**: 45–55.
7. Ideker T, Winslow LR, Lauffenburger AD. Bioengineering and systems biology. *Ann Biomed Eng* 2006; **34**:257–64.
8. Calvo S, Jain M, Xie X, *et al*. Systematic identification of human mitochondrial disease genes through integrative genomics. *Nat Genet* 2006; Epub 2 April 2006.
9. Hwang D, Smith JJ, Leslie DM, *et al*. A data integration methodology for systems biology: experimental verification. *Proc Natl Acad Sci USA* 2005; Epub 21 November 2005.
10. Choe SE, Boutros M, Michelson AM, *et al*. Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset. *Genome Biol* 2005; Epub 28 January 2005.
11. Jansen R, Gerstein M. Analyzing protein function on a genomic scale: the importance of gold-standard positives and negatives for network prediction. *Curr Opin Microbiol* 2004;**7**:535–45.
12. Suomela BP, Andrade MA. Ranking the whole MEDLINE database according to a large training set using text indexing. *BMC Bioinform* 2005;**6**:75.
13. Kim JJ, Zhang Z, Park JC, *et al*. BioContrasts: extracting and exploiting protein-protein contrastive relations from biomedical literature. *Bioinformatics* 2006; Epub 20 December 2005.
14. Masseroli M, Kilicoglu H, Lang FM, *et al*. Argument-predicate distance as a filter for enhancing precision in extracting predications on the genetic etiology of disease. *BMC Bioinformatics* 2006;**7**:291.
15. Survey of Ontologies in Bioinformatics. In: Baclawski K, Niu T, (eds). *Ontologies for Bioinformatics*. MIT Press, 2005.
16. Bodenreider O. Lexical, Terminological, and Ontological Resources for Biological Text Mining. In: Ananiadou S, McNaught J, (eds). *Text Mining for Biology and Biomedicine*. Boston: Artech House, 2006.
17. Srinivasan P, Libbus B. Mining MEDLINE for implicit links between dietary substances and diseases. *Bioinformatics* 2004; **4**:I290–6.
18. Jenssen TK, Laegreid A, Komorowski J, *et al*. A literature network of human genes for high-throughput analysis of gene expression. *Nat Genet* 2001;**28**:21–8.
19. Motschall E, Falck-Ytter Y. Searching the MEDLINE literature database through PubMed: a short guide. *Onkologie* 2005; Epub 19 August 2005.
20. Chang AA, Heskett KM, Davidson TM. Searching the literature using medical subject headings versus text word with PubMed. *Laryngoscope* 2006;**116**:336–40.
21. Dolan ME, Ni L, Camon E, *et al*. A procedure for assessing GO annotation consistency. *Bioinformatics* 2005;**21**: I136–43.
22. The Gene Ontology (GO) project in 2006. *Nucleic Acids Res* 2006;**34**:D322–6.
23. Ashburner M, Ball CA, Blake JA, *et al*. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000;**25**:25–9.
24. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 2004;**32**:D267–70.
25. Sehgal AK, Srinivasan P. Retrieval with gene queries. *BMC Bioinform* 2006;**7**:220.
26. Chen L, Liu H, Friedman C. Gene name ambiguity of eukaryotic nomenclatures. *Bioinformatics* 2005; Epub 27 August 2004.
27. Tamames J, Valencia A. The success (or not) of HUGO nomenclature. *Genome Biol* 2006;**7**:402.
28. Shi L, Campagne F. Building a protein name dictionary from full text: a machine learning term extraction approach. *BMC Bioinform* 2005;**6**:88.
29. Wren JD, Bekeredjian R, Stewart JA, *et al*. Knowledge discovery by automated identification and ranking of implicit relationships. *Bioinformatics* 2004;**20**:389–98.

30. Hunter L, Cohen KB. Biomedical language processing: what's beyond PubMed? *Mol Cell* 2006;**21**:589–94.
31. Blaschke C, Leon EA, Krallinger M, *et al*. Evaluation of BioCreAtIvE assessment of task 2. *BMC Bioinform* 2005; Epub 24 May 2005.
32. Wu CH, Apweiler R, Bairoch A, *et al*. The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res* 2006;**34**:D187–91.
33. Wheeler DL, Barrett T, Benson DA, *et al*. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2006;**34**:D173–80.
34. Gandhi TK, Zhong J, Mathivanan S, *et al*. Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nat Genet* 2006;**38**: 285–93.
35. Rual JF, Venkatesan K, Hao T, *et al*. Towards a proteome-scale map of the human protein-protein interaction network. *Nature* 2005; Epub 28 September 2005.
36. Carter GW, Rupp S, Fink GR, *et al*. Disentangling information flow in the Ras-cAMP signaling network. *Genome Res* 2006; Epub 13 March 2006.
37. Forster J, Famili I, Fu P, *et al*. Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network. *Genome Res* 2003;**13**:244–53.
38. Tewari M, Hu PJ, Ahn JS, *et al*. Systematic interactome mapping and genetic perturbation analysis of a C. elegans TGF-beta signaling network. *Mol Cell* 2004;**13**:469–82.
39. Janes KA, Gaudet S, Albeck JG, *et al*. The response of human epithelial cells to TNF involves an inducible autocrine cascade. *Cell* 2006;**124**:1225–39.
40. Gunsalus KC, Ge H, Schetter AJ, *et al*. Predictive models of molecular machines involved in Caenorhabditis elegans early embryogenesis. *Nature* 2005;**436**:861–5.
41. Donaldson I, Martin J, de Bruijn B, *et al*. PreBIND and Textomy—mining the biomedical literature for protein-protein interactions using a support vector machine. *BMC Bioinformatics* 2003; Epub 27 March 2003.
42. Wolfe JM, Horowitz TS, Kenner NM. Cognitive psychology: rare items often missed in visual searches. *Nature* 2005; **435**:439–40.
43. Hakenberg J, Schmeier S, Kowald A, *et al*. Finding kinetic parameters using text mining. *Omics* 2004;**8**:131–52.
44. Montoya L. *ConceptDB: A software system used to establish a link between PubMed and GO*. In: Gene Ontology Users Meeting Stanford University 2004.
45. Gilchrist M, Thorsson V, Li B, *et al*. Systems biology approaches identify ATF3 as a negative regulator of Toll-like receptor 4. *Nature* 2006;**441**:173–8.
46. Gavin AC, Aloy P, Grandi P, *et al*. Proteome survey reveals modularity of the yeast cell machinery. *Nature* 2006; Epub 22 January 2006.
47. Milward D, Bjreland M, Hayes W, *et al*. Ontology-based interactive information extraction from scientific abstracts. *Comp Funct Genom* 2005;**6**:67–71.
48. Banville DL. Mining chemical structural information from the drug literature. *Drug Discov Today* 2006;**11**:35–42.
49. Tatusov RL, Fedorova ND, Jackson JD, *et al*. The COG database: an updated version includes eukaryotes. *BMC Bioinform* 2003; Epub 11 September 2003.
50. Hamosh A, Scott AF, Amberger JS, *et al*. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* 2005;**33**: D514–7.
51. Ruepp A, Doudieu ON, van den Oever J, *et al*. The Mouse Functional Genome Database (MfunGD): functional annotation of proteins in the light of their cellular context. *Nucleic Acids Res* 2006;**34**:D568–71.
52. Workman CT, Mak HC, McCuine S, *et al*. A systems approach to mapping DNA damage response pathways. *Science* 2006;**312**:1054–9.
53. Brandman O, Ferrell JE,Jr, Li R, *et al*. Interlinked fast and slow positive feedback loops drive reliable cell decisions. *Science* 2005;**310**:496–8.
54. Colman-Lerner A, Gordon A, Serra E, *et al*. Regulated cell-to-cell variation in a cell-fate decision system. *Nature* 2005; Epub 18 September 2005.
55. Xie X, Lu J, Kulbokas EJ, *et al*. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature* 2005; Epub 27 Febuary 2005.
56. Stiffler MA, Grantcharova VP, Sevecka M, *et al*. Uncovering quantitative protein interaction networks for mouse PDZ domains using protein microarrays. *J Am Chem Soc* 2006;**128**:5913–22.
57. Gaudet S, Janes KA, Albeck JG, *et al*. A compendium of signals and responses triggered by prodeath and prosurvival cytokines. *Mol Cell Proteomics* 2005; Epub 18 July 2005.
58. Park SG, Lee T, Kang HY, *et al*. The influence of the signal dynamics of activated form of IKK on NF-kappaB and anti-apoptotic gene expressions: a systems biology approach. *FEBS Lett* 2006; Epub 9 January 2006.
59. Estrada B, Choe SE, Gisselbrecht SS, *et al*. An Integrated Strategy for Analyzing the Unique Developmental Programs of Different Myoblast Subtypes. *PLoS Genet* 2006;**2**:e16.
60. Bouwmeester T, Bauch A, Ruffner H, *et al*. A physical and functional map of the human TNF-alpha/NF-kappa B signal transduction pathway. *Nat Cell Biol* 2004; Epub 25 January 2004.
61. Begley TJ, Rosenbach AS, Ideker T, *et al*. Damage recovery pathways in Saccharomyces cerevisiae revealed by genomic phenotyping and interactome mapping. *Mol Cancer Res* 2002;**1**:103–12.