BRIEFINGS IN BIOINFORMATICS. VOL 9. NO 6. 479–492
Advance Access publication September 27, 2008
doi:10.1093/bib/bbn035

Literature mining in support of drug discovery

Pankaj Agarwal and David B. Searls

Submitted: 7th May 2008; Received (in revised form): 31st July 2008

Abstract

The drug discovery enterprise provides strong drivers for data integration. While attention in this arena has tended to focus on integration of primary data from omics and other large platform technologies contributing to drug discovery and development, the scientific literature remains a major source of information valuable to pharmaceutical enterprises, and therefore tools for mining such data and integrating it with other sources are of vital interest and economic impact. This review provides a brief overview of approaches to literature mining as they relate to drug discovery, and offers an illustrative case study of a 'lightweight' approach we have implemented within an industrial context.

Keywords: *bibliomics; drug discovery; PubMed; biomedical text mining; MeSH*

INTRODUCTION

Despite the fact that the pharmaceutical and biotech industries have invested heavily in the identification of potential new targets for drug discovery by means of novel platform technologies and large-scale screens such as genetic association studies, it is probably the case that the majority of new targets still derive from novel biological discoveries first appearing in the scientific literature from academic sources. In fact, the importance to drug discovery of 'one-off' biological insights arising from basic research was a reality even before the sequencing of the human genome [1]. Today, a novel target is seldom 'discovered' *de novo* but rather functionally characterized in some new way or associated with a disease process. Beyond target identification, new discoveries about a target already in a pharmaceutical pipeline can have immediate relevance and business impact, the more so as the target and associated compounds progress up to and through full development. In addition, the identification of biomarkers that may prove useful in signaling compound action or disease progression has become an important adjunct to drug discovery [2]. For these reasons, literature awareness has traditionally been a constant preoccupation among therapeutic area experts within pharmaceutical companies and other such enterprises. Major scientific journals and in particular therapeutic area specialty journals are avidly followed within industry, and various approaches to 'literature mining' have attracted great interest.

With the reliably exponential growth in the scientific literature over many decades, keeping current is increasingly challenging for any scientist [3]. In the pharmaceutical and biotech industries, there are additional pressures that early recognition of a target opportunity may provide a business advantage, and more significantly, that new information about the target of a late-stage asset may have huge economic impact. Thus, it is not surprising that searches of biomedical journals were the most frequent queries arising in a mid-size pharmaceutical company in a recent study, although what might be less typical of academia is that these were followed closely by searches of competitive intelligence resources and patents [4]. The leading subjects of queries were drugs, diseases and genes (including proteins), with similar frequencies, again perhaps reflecting a different emphasis from academia.

Corresponding author. Dr Pankaj Agarwal, GlaxoSmithKline R&D, 709 Swedeland Road, UW2230, King of Prussia, PA 19406. Tel: (610) 270 5910; Fax: (610) 270 5580; E-mail: Pankaj.Agarwal@gsk.com

Pankaj Agarwal is Director, Computational Biology at GlaxoSmithKline Pharmaceuticals. He has a PhD in Computer Science from the Courant Institute, New York University. His primary interest is in developing systematic solutions for drug discovery.

David Searls is Senior Vice-President, Computational Biology at GlaxoSmithKline Pharmaceuticals. He has degrees from MIT, Johns Hopkins, and Penn, and his research interests include linguistic analysis of macromolecules and data integration.

However, the sheer volume of publications is not the end of the challenges involved in adequately following the literature. It is increasingly understood that narrowly focusing on a pharmaceutical target in isolation can lead to problems, and that it is necessary to consider the action of drugs and their targets in an overall pathway and systems context, beginning with the direct interactors of the target but extending to all manner of relations, direct and indirect [5]. This opens up the scope of literature mining considerably, since biological networks are notoriously densely connected and exhibit 'small world' properties that suggest the need for encyclopedic coverage far beyond the capabilities of individual scientists to manage with realistic investment of time [6].

In the past, such networks have been of interest in drug discovery primarily from the perspective of well-characterized biological pathways, for instance in performing what is termed 'pathway expansion'—a process by which a putative target that is implicated in disease but is found not to be tractable to intervention ('druggable,' in the current terminology) leads to the consideration of its upstream or downstream interactors as candidate targets. Wider networks of interaction are increasingly of interest, however, for insights they may offer into mechanisms of action and target-related side effects of drugs [7], and even the possibility of developing multi-target drugs [8, 9].

More generally, literature-derived networks of interactions, both direct and indirect, are a major component of many approaches to systems biology, with all its associated complexities [10–13]. In this way, knowledge from the scientific literature promises to contribute to drug discovery in new, indirect and perhaps unforeseen ways [13, 14]. For example, the use of indirect interactions to establish disease relevance may be especially important for target identification [15], while the extraction of both qualitative and quantitative information may support mathematical modeling efforts that in turn relate to drug discovery [16, 17]. Text mining has been a key contributor to the reconstruction of many global connectivity maps, for instance that of human metabolism [18]. Interaction data are available not only in the public domain [19–21], but also in extensively curated form from vendors such as Ingenuity (Redwood City, CA, USA) and GeneGo (St Joseph, MI, USA).

Biomedical text mining has been extensively reviewed [22–24], and this article will not attempt to cover the broad topic in any detail. Rather, we will briefly review technical approaches and specialized tools that relate especially to concerns arising in drug discovery (which are broad enough, to be sure). While a number of text-mining tools are available both in the public domain and as commercial products, we will not offer an exhaustive review of these either, though several will be mentioned for specific features. For the most complete understanding of the issues, a hands-on approach to text mining has been advocated [25]. In that spirit, we will finish by describing a particular implementation of our own that embodies many of the techniques found in various other systems, as a demonstration of how to create tools within an industrial context that are suited to particular domains or purposes related to drug discovery.

TECHNOLOGY OVERVIEW

The basic tools of text search and retrieval have been available for decades and are familiar to researchers in a variety of domain-specific implementations. Through various keyword mechanisms and other forms of indexing or document classification, as well as straightforward text search, sets of documents (generally, literature citations and abstracts) can be retrieved with these tools, generally with such additional refinements as Boolean combinations of search terms, iterative refinement of searches, etc. For most biologists, use of the PubMed resource sponsored by the National Center for Biotechnology Information (NCBI), at the National Library of Medicine of the US National Institutes of Health, is second nature. Commercial alternatives include Scopus (Elsevier, Amsterdam, the Netherlands) and Web of Science (Thomson Scientific, Philadelphia, PA, USA), while a notable and novel approach is offered by Google Scholar, which deploys advanced algorithms for enhanced text retrieval from less structured sources on the Worldwide Web, constituting a useful complement to PubMed searches [26, 27]. The computer science involved in creating such algorithms, for example through more sophisticated indexing methods and ordering of results for relevance, is a well-established field of study. The standard metrics for the field assess the relative quality of sets of documents returned in terms of their relevance to the user's intended search criteria, and comprise *recall*, the fraction of relevant documents returned, and *precision*, the fraction of documents returned that are relevant.

Effective use of keywords is an invaluable aid in text retrieval, but these are compromised by inconsistent use across sources and especially by ambiguity due to differing naming conventions. These issues can be addressed by enforcement of standardized or *controlled vocabularies* in a domain (such as SNOMED in medicine [28]) or by the compilation of *synonym lists* that allow for on-the-fly disambiguation in search. As will be seen, the latter are especially important in recognizing gene names in text.

At the other end of the spectrum from text retrieval methods are those of natural language processing (NLP), a branch of artificial intelligence that goes further than simple *lexical* (word) recognition to interpret text through an understanding of *syntax* (grammar), *semantics* (meaning) and still other layers of analysis [29]. Again this is an established field of computational research, especially with regard to applications in the biomedical arena. *Tagging* of text entities such as gene names (sometimes called *entity recognition* [30–32]) has been aided by NLP technologies [33], as well as the recognition of higher order concepts expressed in a variety of ways, although searches that involve NLP parsing of text are generally more computationally intensive than text retrieval methods. NLP methods may recognize not only entities but relations, such as protein–protein or gene–disease, and at its most sophisticated these extend to *semantic role labeling*, which delves still further into linguistic constructs of biological processes to extract information on location, manner, timing and the like [34, 35]. Vendors in this technically challenging arena are legion, and range from tool sets such as Linguamatics (Cambridge, UK), ClearForest (Reuters, Waltham, MA, USA) and AeroText (Lockheed Martin, Gaithersburg, MD, USA), to integrated text analytics and data packages such as VantagePoint (Norcross, GA, USA) and Thompson Data Analyzer (Thompson Reuters, Philadelphia, PA and London, UK).

Another contribution from artificial intelligence is a current emphasis on the use of *ontologies*, essentially to organize indexed terms into meaningful hierarchies that capture domain knowledge. Ontologies can be seen as establishing relationships between terms (such as part–whole or generalization–specialization) that might not otherwise be recognized as related at a lexical level, and in this sense they can be thought of as a higher order of synonym list, though more sophisticated features may include capacities for establishing defaults, handling exceptions and *description logics* that support automated reasoning on the ontology [36, 37]. PubMed employs the MeSH (Medical Subject Headings) hierarchy to aid in search through disambiguation of topics [38], and newer integration tools have made heavy use of the GO (Gene Ontology) hierarchies to classify genes and gene products [39]. The US National Library of Medicine's Unified Medical Language System (UMLS), which provides an ontology in a form called a *semantic network* that references its extensive compilation of biomedical controlled vocabularies, is oriented to support of NLP applications [40]. By extending the semantics of queries, ontologies can greatly enhance (and speed) text search when a corpus of text is marked up with concepts from that ontology, as is done in systems such as GoPubMed [41] and Textpresso [42, 43].

Text analysis has also made heavy use of disciplines such as signal processing and machine learning. In these approaches, documents are generally represented as sparse vectors indicating how often given terms occur in those documents, for example, gene names found in Medline abstracts [44]. In a technique called *latent semantic indexing* (LSI), these vectors are arranged in a matrix, so that columns comprise documents and rows list terms [45]. LSI uses a standard linear algebra technique called *singular value decomposition* on such matrices, to establish a joint 'concept space' that can readily be queried to determine the similarities of terms or documents to each other, or to find terms and/or documents related to an arbitrary textual string. Recently, LSI has been used to associate PubMed abstracts to GO ontology terms, in the SEGOPubmed system [46].

Machine learning provides other tools for text mining, especially in the area of clustering of documents into similar groups based on term content. *Naïve Bayes classifiers*, which are most notably used to filter e-mail spam, have also been applied to searching Medline for articles most relevant to a given set of articles, as in the MScanner system [47], as well as in more domain-specific applications like the Immune Epitope Database [48]. *Support vector machines* (SVMs), another popular machine-learning methodology, have found wide use in biomedical document classification [43, 49, 50]. SVMs find hyperplanes that provide the widest separation between sets of vectors of high dimension, in a manner well suited to the sparse vectors of terms that represent documents; they have also proven useful in tackling the NLP problems of named entity

recognition [51], word sense disambiguation [52, 53] and semantic role labeling [54].

Thus, combinations of these basic methods and many related ones too numerous to mention are often used to create systems that are well-suited to supporting the biomedical researcher, and in some cases the drug discovery enterprise specifically [4, 55, 56]. Such systems may allow for individual queries of bibliographic databases in some enhanced way [57, 58], or they may allow for clustering, categorizing and summarizing sets of documents [59], which may be particularly helpful in maintaining a knowledge base in a particular therapeutic area.

DRUG DISCOVERY APPLICATIONS

While generic literature-mining techniques have often been used to derive drug-related annotation from literature as just one of a variety of applications, some workers have focused on domain-specific problems in greater depth. Kolarik *et al.* [60] utilized NLP techniques to extract likely annotation terms related to drugs from textual descriptions in DrugBank, and by applying these lexico-semantic patterns to sources such as Medline they find that they are able to automatically extend and update such resources with novel descriptions of pharmacologic effects of drugs. Similar but more specific text-extraction methods have also been applied to such tasks as deriving information about drug–drug interactions [61] and interactions of compounds with drug metabolizing enzymes [62], while other NLP efforts have focused on exploiting taxonomic organization of chemicals and drugs in text mining [63].

Less attention has been paid to extraction of chemical annotation from literature than that related to genes and gene products of late, perhaps because the IUPAC nomenclature and resources such as Index Chemicus and Chemical Abstracts Service (CAS) were well-established at an early date [64], and the lexico-syntactic issues are more clear-cut in the case of chemistry. For instance, it was recognized nearly a half-century ago that a systematic chemical name could be algorithmically converted to a molecular formula and structure [65], a capability for which molecular geneticists might well long. Mining of chemical entities from the literature has been extended with the collection of related attributes, as well as the use of NLP techniques to tackle more sophisticated tasks such as building databases of reactions and structure–activity relationships; both the academic foundations and commercial implementations in this arena have been well reviewed by Banville [64].

Medicinal chemists may be challenged by the need to search the literature not only for chemical compounds but also for text related to biological systems. One particularly imaginative approach to this has been the extension of LSI to chemical structures, by creating a matrix of molecules versus chemical descriptors [66]. This not only allows for LSI-style query of a molecule/descriptor concept space, but it can also be conjoined with conventional textual LSI to jointly search the literature with compounds, descriptors and text terms, in a technique called Text Influenced Molecular Indexing [67]. The fact that machine-learning methodologies such as SVMs have been used extensively for classification based on physicochemical properties of both small molecules [68, 69] and proteins (including proteins of particular relevance to drug discovery, such as common target classes and enzymes related to drug absorption and metabolism [70–72]) suggests the likely fruitfulness of further such approaches to text mining conjoined with molecular search and classification.

The interaction of chemical compounds and gene products is, of course, at the heart of the drug discovery enterprise. Many efforts have concentrated on the co-occurrence in text of gene identifiers; Zhu *et al.* [73], on the other hand, used co-occurrence to discover implicit 'chemical compound–gene' relations in the literature. An earlier experimental system, EDGAR, used a more sophisticated NLP approach to extract both genes and drugs, as well as relations between them [74]. In one example of a useful classification application, a statistical approach was used to identify articles describing gene–drug interactions from MedLine abstracts, identifying nearly 5000 articles deemed relevant to pharmacogenetics with high precision [75]. Co-occurrences of drugs and diseases in the literature as well as in clinical notes have been used to validate semantic interpretations of MedLine abstracts [76] and to assess the strengths of such associations [77]. Overall, however, work in this area has not progressed remarkably considering the importance of the task to the drug discovery enterprise.

On the other hand, network approaches that consolidate relationships among different object types (gene, disease, etc.) are increasingly including drugs among the multi-way interactions depicted [78]. As a rule, these varieties of interaction mining

should not be expected to perform any better than the best-studied of the binary text-mining tasks, those of gene-gene interactions, and specifically physical protein–protein interactions; while such technology has shown improvement [79–83], it still falls short of what can be obtained by manual curation, which however may itself benefit by a computational assist from such systems [84]. What is likely to be of greatest utility with regard to integrating chemical and biological knowledge is the PubChem resource at NCBI, which links many millions of compounds and substances to data from high-throughput screening as well as to PubMed literature and Entrez gene data [85]. Such multi-way networks not only comprise a form of data integration in themselves, *qua* data structure, but also constitute a useful visualization and browsing modality, as an adjunct to query [24].

Beyond the foundational activities of entity recognition and relationship extraction from literature, hypothesis generation has been rightly identified as a literature-based activity of importance in biomedicine and in particular drug discovery [23]. It is interesting that the early work of D. R. Swanson is often held up as the prototype of inferential analysis from multiple sources, for instance transitive closure of relations such as 'influences' (X influences Y from one source, Y influences Z from another, and it is therefore inferred that X influences Z) [86]. Swanson's seminal paper used analogical reasoning across sources to hypothesize that fish oil might be therapeutic in Raynaud's syndrome [87], and a follow-up study suggested a relationship between migraines and magnesium [88], both of which were later supported by clinical and experimental findings. Such computational hypothesis generation is likely to be of increasing interest in the pharmaceutical industry, for example in drug repositioning (by which new applications for existing compounds are identified), as has been demonstrated in the identification of novel indications for thalidomide [89] and curcumin [90]. Among a variety of platform approaches to identifying repositioning opportunities, the IDMap system is notable for combining text mining with chemical structure information [91].

CASE STUDY: GLAXOSMITHKLINE

We present a case study in the use (and reuse) of various tools and technologies for literature mining at GlaxoSmithKline (GSK). These are presented not because they are unique or particularly sophisticated approaches, but rather because they are typical of many such systems that have been deployed to the web or implemented locally. The main point here is to demonstrate how generic methods and resources can be used to meet the needs of any given enterprise with as much customization as is deemed necessary.

At GSK, we make heavy use of PubMed and Entrez Programming Utilities (eUtils) from NCBI (http://www.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html). Given that drug discovery primarily draws on genes with established nomenclature, we rely on known gene synonyms and descriptions to tag human, mouse and rat genes in article abstracts. Although NCBI maintains a much-used web-based query service, we find it desirable to download and maintain a local corpus of Medline. For each article, we extract the MeSH terms, including substance names that have been indexed for the article. We prefer to use the manually curated MeSH terms rather than parsing text based on synonyms from Unified Medical Language System (UMLS) [92] or other ontologies simply because manual curation, when comprehensive and systematic, remains the gold standard.

Each article's title and abstract is also scanned for all high-quality gene names. The Gene name list is built by integrating names and descriptions from multiple fields within EntrezGene, HUGO, and UniProt. The EntrezGene data was downloaded from ftp://ftp.ncbi.nih.gov/gene/DATA/. All gene names and descriptions were extracted from the gene_info.gz file. The file gene2accession.gz was used to map UniProt accessions to EntrezGene. HUGO approved gene names and symbols were extracted from the 'All Data text' downloaded from http://www.genenames.org/data/gdlw_index.html. All the gene names and descriptions in the fields: GN, ID, and DE were extracted from UniProt downloaded from ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/uniprot_sprot.dat.gz.

All human, mouse, and rat gene synonyms are mapped onto the orthologous human EntrezGene. This yields a total of 313K synonyms (from a December 2007 gene synonym build), though only ~74K of the synonyms were found in PubMed abstracts. In addition, we also add identifiers and other internal names that aid in the query interface. Gene synonyms that are ambiguous (i.e., refer to multiple human EntrezGenes) are flagged.

Gene names that correspond to common English words or those that are likely to be other medical terms or abbreviations (for example, most three-letter names) are assigned a downgraded quality and not used with default parameter settings. This eliminates about 8K synonyms with non-zero count. Moreover, we also have a process for downgrading gene names that are manually identified as being not useful. This helps improve specificity over time.

From each PubMed article, the following features are extracted: PubMed identifier, year of publication, title, author list, affiliation, MeSH terms (with flag indicating if major) and substance names. This yields ~200 million article-to-MeSH term or substance-name mappings. In addition, gene names are flagged by scanning title and abstract. Moreover, two data files are used from EntrezGene to augment Gene to PubMed mappings (Table 1): Gene2pubmed (from ftp://ftp.ncbi.nih.gov/gene/DATA/gene2pubmed.gz) and GeneRIF (from ftp://ftp.ncbi.nih.gov/gene/GeneRIF/generifs_basic.gz). GeneRIF (Gene Reference into Function) provides a quality functional annotation that may extend beyond the genes mentioned in the abstract [93].

A number of excellent and well-validated entity recognition systems have been suggested from PubGene [94] to iHoP [95]. We chose to implement our own for reasons of flexibility, availability and history; however, if initiating a new one careful attention should be paid to the best systems from evaluations such as BioCreAtIvE [96]. The GeneRIF data may be used as an extensive gold standard to compute the recall for our scheme for matching a PubMed article with an EntrezGene. GeneRIF contained 259 739 unique gene to PubMed mappings. This is based on the human gene mappings, plus, the human ortholog of the mouse and rat genes mappings as of 5 June 2008. Orthology was determined using Homologene [97]. Of the 260K mappings, 217 020 (84%) were also found independently by our system. (For this comparison, we ignored both GeneRIF and Pubmed2Gene mappings in our system.) In total, the system contains ~7.1 million PubMed-to-human-gene mappings covering ~3.3 M articles and ~17K human genes.

Table 1: Number of unique gene to PubMed mappings for each species extracted from the GeneRIF and the Gene2pubmed files provided by Entrezgene

	GeneRIF	Pubmed2Gene	Entity recognition (M)
Human	183 405	420 467	5.5
Mouse	65 039	538 339	5.0
Rat	26 781	61 880	3.8
Human ortholog	259 739	810 735	7.1

The Entity recognition column has the number of gene to PubMed mappings found across all PubMed using the species-specific subset of the ~300K gene synonyms. The human orthologs numbers are from combining human genes with mouse and rat genes mapped to the human ortholog from Homologene [97].

A major advantage of integrating PubMed with gene names is that for any particular disease it enables us to generate a set of associated genes. We use eUtils (from NCBI) to retrieve the PubMed identifiers corresponding to a particular disease. Given these PubMed identifiers, we can then locally generate lists of all genes mentioned in those articles. This list is then prioritized based on Fisher Exact *P*-values for the association of this gene with the disease [98]. This *P*-value computation takes into account both the specificity of the gene name and the disease in PubMed. It would be useful to have objective estimates for this precision at the disease–gene association level, the numbers for which are both superior and more relevant than the precision for each gene-to-PubMed association. In other words, the Fisher exact test can help with prioritizing the gene–disease level associations, and in our experience associations with a $P < 0.05$ have fairly high precision.

In a drug discovery context, the value of categorizing literature by gene, disease or compound is manifold. We can immediately get a ranked list of genes associated with a disease, see Figure 1 for results based on the PubMed query 'Asthma [majr]' [The suffix *[majr]* indicates that the MeSH term should be a major topic of the article. The suffix *[mh]* includes any article that is indexed with that MeSH term (not necessarily a major topic). See PubMed help (http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=helppubmed.section.pubmedhelp.Search_Field_Descrip) for a complete description]. This list can then be followed up in a top-down fashion until one finds the first

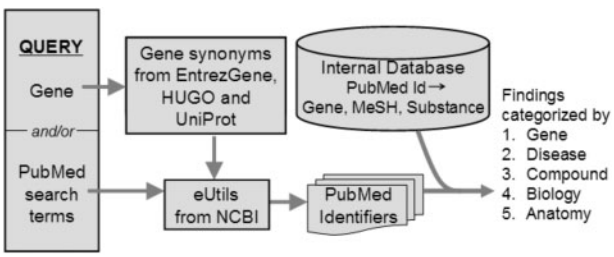

Figure 1: Simplified workflow of the literature mining system at GlaxoSmithKline. See text for a description.

'novel' association, which can then be further evaluated, and possibly rejected (if due to an error in gene name mapping, or due to the use of the gene name in a different context than the disease within the abstract or full-text). The *P*-value is computed independently for each gene name. This makes the system transparent and also helps account for gene names that may be ambiguous. However, a case could also be made for combining the synonyms to enhance statistical power.

The list of genes with associated *P*-values (as in Figure 2) can be used to generate a network

No.	EntrezGene	Gene description	PubMed references	Gene	Observed PubMed count	Expected PubMed count	P-Value
1	EG.3567	interleukin 5 (colony-stimulating factor, eosinophil)	Titles	IL-5 A,S	525	94	4.2e-215
				interleukin-5 A,S	350	61	2.6e-146
				EG.3567 † A,S	11	1.2	2.6e-08
				interleukin 5 A,S	17	5.3	3.1e-05
				EG.16191 † A,S	3	2.2	0.38
2	EG.3596	interleukin 13	Titles	IL-13 A,S	368	51	3.2e-188
				EG.3596 † A,S	22	2.8	1.7e-13
				IL13 A,S	16	2.1	5e-10
				EG.16163 † A,S	15	2.4	2.3e-08
				interleukin 13 A,S	8	2.3	0.0022
3	EG.6037	ribonuclease, RNase A family, 3 (eosinophil cationic protein)	Titles	eosinophil cationic protein A,S	244	23	4.1e-165
				RNASE3 A,S	19	0.83	5.8e-21
				EG.6037 † A,S	9	0.61	9.4e-09
4	EG.6356	chemokine (C-C motif) ligand 11	Titles	eotaxin A,S	184	17	4.8e-124
				CCL11 A,S	130	13	2.5e-86
				EG.6356 † A,S	8	0.87	2.6e-06
				EG.20292 † A,S	1	0.67	0.49
5	EG.80332	ADAM metallopeptidase domain 33	Titles	ADAM33 A,S	69	1.4	4.7e-103
				EG.80332 † A,S	24	0.49	3e-36
				a disintegrin and metalloprotease 33 A,S	8	0.15	1.6e-13

Figure 2: List of genes associated with the most recent 20 000 articles from PubMed based on the query 'Asthma[majr]' (only top few genes with the best Fisher *P*-values shown). All the gene names found in the text are shown. Daggers indicate that the gene association is due to a GeneRIF or a gene2pubmed entry.

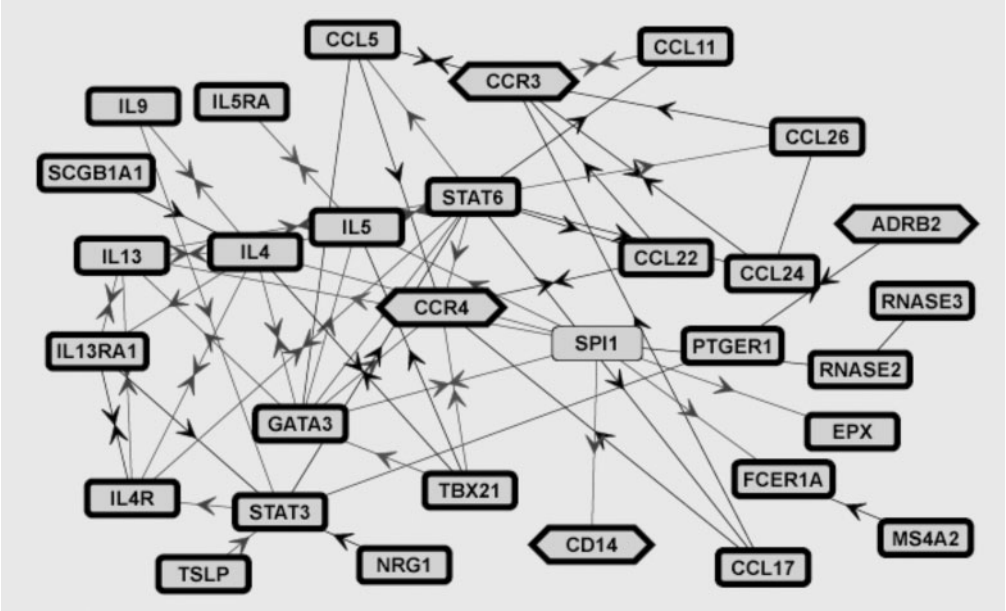

Figure 3: A protein interaction network enriched in genes associated with the query: 'Asthma[majr]'. The subnetwork edges are based on direct interactions between proteins that were extracted manually from literature. This is a maximally scoring subgraph enriched for genes with the best *P*-values [99].

No.	EntrezGene	Gene description	PubMed references	Gene	Observed PubMed count	Expected PubMed count	P-Value
1	EG.80332	ADAM metallopeptidase domain 33	Titles	ADAM33	69	1.4	4.7e-103
				EG.80332 †	24	0.49	3e-36
				a disintegrin and metalloprotease 33	8	0.15	1.6e-13
				EG.110751 †	3	0.12	0.00019
2	EG.4056	leukotriene C4 synthase	Titles	LTC4S	19	0.71	1.8e-22
				EG.4056 †	16	0.67	3.2e-18
				leukotriene C4 synthase	21	1.7	6.5e-17
				LTC4 synthase	12	1.5	4.9e-08
3	EG.57628	dipeptidyl-peptidase 10	Titles	DPP10	12	0.39	1.3e-15
				dipeptidyl peptidase 10	3	0.08	4.6e-05
				EG.57628 †	2	0.15	0.009
				EG.269109 †	1	0.027	0.026
4	EG.240	arachidonate 5-lipoxygenase	Titles	ALOX5	12	0.45	1.3e-14
				EG.240 †	9	1.4	1.4e-05
				arachidonate 5-lipoxygenase	44	29	0.0056
				5-LO	15	8.8	0.034
5	EG.8288	eosinophil peroxidase	Titles	eosinophil peroxidase	37	9.7	1.3e-11
6	EG.128	alcohol dehydrogenase 5 (class III), chi polypeptide	Titles	S-nitrosoglutathione reductase	6	0.31	4.7e-07
				GSNOR	2	0.13	0.0075
				EG.11532 †	1	0.23	0.2
				EG.128 †	1	0.32	0.28
7	EG.241	arachidonate 5-lipoxygenase-activating protein	Titles	ALOX5AP	7	0.56	1.3e-06
				EG.241 †	5	0.64	0.00045

Figure 4: Partial list of small molecule *tractable* targets mentioned in recent literature associated with the PubMed query 'Asthma[majr]'.

view of the disease (Figure 3) by integrating it with protein–protein interaction data derived from either public databases or commercial vendors, as we have described previously [99]. This can help focus attention on the key genes associated with the disease.

A potential list of small molecule drug targets is obtained (Figure 4) by intersecting the resulting gene list by a druggable genome list of proteins that are considered tractable for small molecule compounds. (The druggable genome list used is based on Russ and Lampel [100]; however, other proprietary lists may be easily substituted.) In addition, we can get a list of potential biomarkers for any disease by refining the query to 'Asthma [majr] Biological Marker [mh]' (Figure 5). These results can then be evaluated by scanning the papers that suggest the connection, as there are likely to be a few false positives based on publications that discuss some aspect of a biomarker, but the gene name may be mentioned in a different context. (Of course, an advantage of eUtils is that we can readily process any PubMed query by letting NCBI resolve it.)

We can also retrieve and categorize the results of any query based on a gene. We perform this task by generating all the aliases of that gene and querying eUtils with a composite 'OR' query involving all the

No.	EntrezGene	Gene description	PubMed references	Gene	Observed PubMed count	Expected PubMed count	P-Value
1	EG.6037	ribonuclease, RNase A family, 3 (eosinophil cationic protein)	Titles	eosinophil cationic protein A,S	137	2.3	8.4e-191
				RNASE3 A,S	6	0.084	3.6e-10
				EG.6037 † A,S	3	0.063	3.7e-05
2	EG.3567	interleukin 5 (colony-stimulating factor, eosinophil)	Titles	IL-5 A,S	162	9.6	5.7e-138
				interleukin-5 A,S	119	6.3	3.5e-107
				interleukin 5 A,S	7	0.54	1.5e-06
				EG.3567 † A,S	4	0.12	7e-06
				EG.16191 † A,S	1	0.23	0.2
3	EG.3596	interleukin 13	Titles	IL-13 A,S	102	5.2	1.5e-93
				EG.3596 † A,S	10	0.29	6.2e-13
				EG.16163 † A,S	4	0.24	0.00011
				IL13 A,S	3	0.21	0.0014
4	EG.3565	interleukin 4	Titles	IL-4 A,S	178	30	7e-79
				EG.3565 † A,S	9	0.62	1.9e-08
				interleukin 4 A,S	10	2.3	0.00013
				EG.16189 † A,S	5	0.91	0.0025
5	EG.6356	chemokine (C-C motif) ligand 11	Titles	eotaxin A,S	61	1.8	4.3e-71
				CCL11 A,S	54	1.3	1.3e-67
				EG.6356 † A,S	2	0.089	0.0036
6	EG.3383	intercellular adhesion molecule 1 (CD54), human rhinovirus receptor	Titles	Icam-1 A,S	119	14	4.9e-67
				CD54 A,S	16	2.6	2.1e-08

Figure 5: Partial list of genes mentioned significantly often in abstracts that result from the query 'Asthma[majr] Biological Marker[mh]'.

No.	MeSH	EntrezGene	Gene description	PubMed references	Gene name in text	Observed PubMed count	Expected PubMed count	P-value
1	Insulin Resistance [MeSH]	EG.3643	insulin receptor	Titles	INSR A,S	16	1.1	2.7e-14
		EG.3643	insulin receptor	Titles	EG.16337 † A,S	1	1.1	1
		EG.3643	insulin receptor	Titles	Insulin receptor precursor A,S	3	0.19	0.00093
		EG.3643	insulin receptor	Titles	EG.3643 † A,S	7	2.4	0.01
2	Polycystic Ovary Syndrome [MeSH]	EG.3643	insulin receptor	Titles	INSR A,S	6	0.56	2.4e-05
		EG.3643	insulin receptor	Titles	EG.3643 † A,S	5	1.2	0.0082
3	Myotonic Dystrophy [MeSH]	EG.3643	insulin receptor	Titles	INSR A,S	3	0.3	0.0036
4	Fetal Nutrition Disorders [MeSH]	EG.3643	insulin receptor	Titles	INSR A,S	1	0.005	0.005
5	Diabetes Mellitus, Type 2 [MeSH]	EG.3643	insulin receptor	Titles	INSR A,S	10	3.9	0.006
		EG.3643	insulin receptor	Titles	EG.3643 † A,S	5	8.5	1
6	Hyperphagia [MeSH]	EG.3643	insulin receptor	Titles	Insulin receptor precursor A,S	1	0.019	0.019
7	Adenomatous Polyps [MeSH]	EG.3643	insulin receptor	Titles	INSR A,S	1	0.047	0.046
		EG.3643	insulin receptor	Titles	EG.3643 † A,S	1	0.1	0.098

Figure 6: Diseases mentioned as major MeSH terms in articles that mention the gene: INSR (insulin receptor) or an alias.

gene names. The resulting PubMed identifiers can then be categorized by the diseases mentioned as major MeSH terms in those articles (Figure 6). Thus, the literature on any gene can be summarized by the diseases associated with that gene. The results output is ordered by the 'unexpectedness' of a disease association, which is computed using a Fisher's exact test [98]. Instead of diseases, we can also generate 'Anatomical terms', 'Biological terms' or 'Chemical substances' associated with a query. This takes advantage of the MeSH tree organization, substance name, and the semantic network from the UMLS [92].

A gene-based query can be combined with a set of terms (such as, IC50 OR XC50 OR pKi OR pKd OR agonist OR antagonist) that suggest that the abstract may mention a tool compound (for example, see Figure 7). The resulting abstracts are then categorized by the MeSH terms corresponding to chemical substances and other supplementary substance names mentioned in those articles. This is often an initial step in identifying known compounds that may be used as tool compounds for target validation.

Queries can also be combined using both gene and disease terms. Thus, one can query using both 'insulin' and 'Diabetes Mellitus [majr]' to find other genes associated with insulin in the context of diabetes. This simple framework enables fairly powerful queries that handle genes correctly, based on extensive practical experience.

CONCLUSIONS

Scientists engaged in drug discovery who wish to perform effective literature search might first do well to ensure that they are fully exploiting the powerful search capabilities offered online by PubMed. While the basic service is well-known and heavily used, in our experience it is rare to see the average scientist making the most effective use of Boolean combinations, filters of various kinds and in particular the rich set of MeSH headings by which articles are indexed. Such searches are often usefully complemented by Google Scholar, which draws on a wider variety of sources and rank-order results based on citations.

No.	MeSH	PubMed references	Observed PubMed count	Expected PubMed count	P-Value
1	Hydroxymethylglutaryl-CoA Reductase Inhibitors [MeSH] [PubChem]	Titles	83	1	2.4e-138
2	Hydroxymethylglutaryl CoA Reductases [MeSH] [PubChem]	Titles	37	0.27	3.9e-67
3	Simvastatin [MeSH] [PubChem]	Titles	33	0.33	2.4e-55
4	Lovastatin [MeSH] [PubChem]	Titles	31	0.34	8.3e-51
5	Pravastatin [MeSH] [PubChem]	Titles	27	0.23	1.6e-47
6	Heptanoic Acids [MeSH] [PubChem]	Titles	20	0.27	3.2e-31
7	Mevalonic Acid [MeSH] [PubChem]	Titles	16	0.25	4e-24
8	Anticholesteremic Agents [MeSH] [PubChem]	Titles	22	0.93	1.7e-23
9	atorvastatin [MeSH] [PubChem]	Titles	15	0.23	6.1e-23
10	cerivastatin [MeSH] [PubChem]	Titles	10	0.045	5.5e-21
11	Cholesterol [MeSH] [PubChem]	Titles	45	8.5	7.7e-21
12	Pyrroles [MeSH] [PubChem]	Titles	17	0.98	3.2e-16
13	fluvastatin [MeSH] [PubChem]	Titles	9	0.084	3.8e-16

Figure 7: Partial list of potential tool compounds associated with HMGCR (HMG-CoA reductase).

At the other extreme, bioinformatics or information sciences professionals in a pharmaceutical context may make effective use of new NLP technologies, either downloaded from the public domain or obtained from vendors, to perform highly sophisticated and fit-for-purpose queries, classifications, etc. These will no doubt produce superior results (albeit at an obvious cost), particularly in domains that have not been as effectively indexed or which have other peculiarities calling for syntactic or semantic analysis in the course of search. However, with regard to increasingly important interaction and pathway data, even the most sophisticated NLP approaches are not yet competitive with manual curation, where the onus again reverts to the expert scientist, perhaps aided by the tools above.

A cost-effective alternative that lies between the two extremes, and which will often suffice for specialized searches such as may arise in a pharmaceutical context, is to utilize PubMed and adapt NCBI eUtils as described in the previous section. By downloading the PubMed abstracts (or major subsets), local scripts may be run that perform any necessary compute-intensive operations as well as application-specific operations, while still relying on the NCBI apparatus for intermittent search and ancillary data retrieval. This has the advantage that it is possible to maintain updated databases with relative ease and avoid supporting an extensive code base. This 'lightweight' approach to text mining has proven robust and sustainable in our hands, and is an attractive alternative to 'IT-heavy' solutions with a much greater degree of integration with platform data and other resources. While we typically combine the results from this system with other bibliographic, patent and competitive intelligence sources (not described in this review), such activities are done in a modular way, often with the appropriate vendor tools in standalone fashion, and sometimes by other groups entirely.

This approach may not attain the ideal of comprehensive data integration across all sources for all purposes, but the tools described have proven effective as a first step toward exploiting current literature with a relatively small infrastructure investment. It may be closer in spirit to what has been called a 'Google for bioinformatics,' not so much an authoritative integrated database but rather an enriched literature search with an effort at meaningful rank-ordering of results.

Acknowledgements

The authors would like to thank Vinod Kumar, Liwen Liu, Tom White, Dilip Rajagopalan, William Reisdorf, Karen Kabnick and Ron Liu for their contributions.

References

1. Chin-Dusting J, Mizrahi J, Jennings G, *et al*. Outlook: finding improved medicines: the role of academic-industrial collaboration. *Nat Rev Drug Discov* 2005;**4**:891–7.
2. LaBaer J. Mining the literature and large datasets. *Nat Biotechnol* 2003;**21**:976–7.
3. Hunter L, Cohen KB. Biomedical language processing: what's beyond PubMed? *Mol Cell* 2006;**21**:589–94.

4. Roberts PM, Hayes WS. Information needs and the role of text mining in drug development. *Pac Symp Biocomput* 2008; 592–603.
5. Searls DB. Pharmacophylogenomics: genes, evolution and drug targets. *Nat Rev Drug Discov* 2003;**2**:613–23.
6. Yildirim MA, Goh KI, Cusick ME, *et al*. Drug-target network. *Nat Biotechnol* 2007;**25**:1119–26.
7. Wang Y, Chiu JF, He QY. Proteomics approach to illustrate drug action mechanisms. *Curr Drug Discov Technol* 2006;**3**:199–209.
8. Nacher JC, Schwartz JM. A global view of drug-therapy interactions. *BMC Pharmacol* 2008;**8**:5.
9. Csermely P, Agoston V, Pongor S. The efficiency of multi-target drugs: the network approach might help drug design. *Trends Pharmacol Sci* 2005;**26**:178–82.
10. Peng Y, Zhang X. Integrative data mining in systems biology: from text to network mining. *Artif Intell Med* 2007;**41**:83–6.
11. Roberts PM. Mining literature for systems biology. *Brief Bioinform* 2006;**7**:399–406.
12. Ananiadou S, Kell DB, Tsujii J. Text mining and its potential applications in systems biology. *Trends Biotechnol* 2006;**24**:571–9.
13. Cho CR, Labow M, Reinhardt M, *et al*. The application of systems biology to drug discovery. *Curr Opin Chem Biol* 2006;**10**:294–302.
14. Materi W, Wishart DS. Computational systems biology in drug discovery and development: methods and applications. *Drug Discov Today* 2007;**12**:295–303.
15. Loging W, Harland L, Williams-Jones B. High-throughput electronic biology: mining information for drug discovery. *Nat Rev Drug Discov* 2007;**6**:220–30.
16. Fattore M, Arrigo P. Knowledge discovery and system biology in molecular medicine: an application on neuro-degenerative diseases. *In Silico Biol* 2005;**5**:199–208.
17. Hakenberg J, Schmeier S, Kowald A, *et al*. Finding kinetic parameters using text mining. *Omics* 2004;**8**:131–52.
18. Duarte NC, Becker SA, Jamshidi N, *et al*. Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc Natl Acad Sci USA* 2007; **104**:1777–82.
19. Breitkreutz BJ, Stark C, Reguly T, *et al*. The BioGRID interaction database: 2008 update. *Nucleic Acids Res* 2008;**36**: D637–40.
20. Chatr-aryamontri A, Ceol A, Palazzi LM, *et al*. MINT: the Molecular INTeraction database. *Nucleic Acids Res* 2007;**35**: D572–4.
21. Han K, Park B, Kim H, *et al*. HPID: the human protein interaction database. *Bioinformatics* 2004;**20**:2466–70.
22. Zweigenbaum P, Demner-Fushman D, Yu H, *et al*. Frontiers of biomedical text mining: current progress. *Brief Bioinform* 2007;**8**:358–75.
23. Cohen AM, Hersh WR. A survey of current work in biomedical text mining. *Brief Bioinform* 2005;**6**:57–71.
24. Jensen LJ, Saric J, Bork P. Literature mining for the biologist: from information retrieval to biological discovery. *Nat Rev Genet* 2006;**7**:119–29.
25. Cohen KB, Hunter L. Getting started in text mining. *PLoS Comput Biol* 2008;**4**:e20.
26. Shultz M. Comparing test searches in PubMed and google scholar. *J Med Libr Assoc* 2007;**95**:442–5.
27. Falagas ME, Pitsouni EI, Malietzis GA, *et al*. Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses. *Faseb J* 2008;**22**:338–42.
28. de Bruijn LM, Hasman A, Arends JW. Automatic SNOMED classification – a corpus-based method. *Comput Methods Programs Biomed* 1997;**54**:115–22.
29. Chang JT, Altman RB. Promises of text processing: natural language processing meets AI. *Drug Discov Today* 2002;**7**: 992–3.
30. Leser U, Hakenberg J. What makes a gene name? Named entity recognition in the biomedical literature. *Brief Bioinform* 2005;**6**:357–69.
31. Leaman R, Gonzalez G. BANNER: an executable survey of advances in biomedical named entity recognition. *Pac Symp Biocomput* 2008;652–63.
32. Tsai RT, Wu SH, Chou WC, *et al*. Various criteria in the evaluation of biomedical named entity recognition. *BMC Bioinformatics* 2006;**7**:92.
33. Sandler T, Schein AI, Ungar LH. Automatic term list generation for entity tagging. *Bioinformatics* 2006;**22**: 651–7.
34. Tsai RT, Chou WC, Su YS, *et al*. BIOSMILE: a semantic role labeling system for biomedical verbs using a maximum-entropy model with automatically generated template features. *BMC Bioinformatics* 2007;**8**:325.
35. Kogan Y, Collier N, Pakhomov S, *et al*. Towards semantic role labeling & IE in the medical literature. *AMIA Annu Symp Proc* 2005;410–4.
36. Moreira DA, Musen MA. OBO to OWL: a protege OWL tab to read/save OBO ontologies. *Bioinformatics* 2007;**23**: 1868–70.
37. Rector A. Defaults, context, and knowledge: alternatives for OWL-indexed knowledge bases. *Pac Symp Biocomput* 2004; 226–37.
38. Srinivasan P, Hristovski D. Distilling conceptual connections from MeSH co-occurrences. *Medinfo* 2004;**11**:808–12.
39. Lomax J. Get ready to GO! A biologist's guide to the Gene Ontology. *Brief Bioinform* 2005;**6**:298–304.
40. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 2004;**32**:D267–70.
41. Doms A, Schroeder M. GoPubMed: exploring PubMed with the Gene Ontology. *Nucleic Acids Res* 2005;**33**:W783–6.
42. Muller HM, Kenny EE, Sternberg PW. Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biol* 2004;**2**:e309.
43. Chen D, Muller HM, Sternberg PW. Automatic document classification of biological literature. *BMC Bioinformatics* 2006;**7**:370.
44. Homayouni R, Heinrich K, Wei L, *et al*. Gene clustering by latent semantic indexing of MEDLINE abstracts. *Bioinformatics* 2005;**21**:104–15.
45. Deerwester S, Dumais S, Furnas GW, *et al*. Indexing by latent semantic indexing. *J Am Soc Info Sci* 1990;**41**: 391–407.
46. Vanteru BC, Shaik JS, Yeasin M. Semantically linking and browsing PubMed abstracts with gene ontology. *BMC Genomics* 2008;**9**(Suppl 1):S10.
47. Poulter GL, Rubin DL, Altman RB, *et al*. MScanner: a classifier for retrieving Medline citations. *BMC Bioinformatics* 2008;**9**:108.

48. Wang P, Morgan AA, Zhang Q, *et al*. Automating document classification for the immune epitope database. *BMC Bioinformatics* 2007;**8**:269.
49. Polavarapu N, Navathe SB, Ramnarayanan R, *et al*. Investigation into biomedical literature classification using support vector machines. *Proc IEEE Comput Syst Bioinform Conf* 2005;366–74.
50. Cohen AM. An effective general purpose approach for automated biomedical document classification. *AMIA Annu Symp Proc* 2006;161–5.
51. Zhou GD. Recognizing names in biomedical texts using mutual information independence model and SVM plus sigmoid. *Int J Med Inform* 2006;**75**:456–67.
52. Xu H, Markatou M, Dimova R, *et al*. Machine learning and word sense disambiguation in the biomedical domain: design and evaluation issues. *BMC Bioinformatics* 2006;**7**:334.
53. Pahikkala T, Ginter F, Boberg J, *et al*. Contextual weighting for support vector machines in literature mining: an application to gene versus protein name disambiguation. *BMC Bioinformatics* 2005;**6**:157.
54. Bethard S, Lu Z, Martin JH, *et al*. Semantic role labeling for protein transport predicates. *BMC Bioinformatics* 2008;**9**:277.
55. Erhardt RA, Schneider R, Blaschke C. Status of text-mining techniques applied to biomedical text. *Drug Discov Today* 2006;**11**:315–25.
56. Krallinger M, Erhardt RA, Valencia A. Text-mining approaches in molecular biology and biomedicine. *Drug Discov Today* 2005;**10**:439–45.
57. Hale R. Text mining: getting more value from literature resources. *Drug Discov Today* 2005;**10**:377–9.
58. Perez-Iratxeta C, Bork P, Andrade MA. Exploring MEDLINE abstracts with XplorMed. *Drugs Today* 2002; **38**:381–9.
59. Mack R, Hehenberger M. Text-based knowledge discovery: search and mining of life-sciences documents. *Drug Discov Today* 2002;**7**:S89–98.
60. Kolarik C, Hofmann-Apitius M, Zimmermann M, *et al*. Identification of new drug classification terms in textual resources. *Bioinformatics* 2007;**23**:i264–72.
61. Mille F, Degoulet P, Jaulent MC. Modeling and acquisition of drug-drug interaction knowledge. *Medinfo* 2007;**12**:900–4.
62. Feng C, Yamashita F, Hashida M. Automated extraction of information from the literature on chemical-CYP3A4 interactions. *J Chem Inf Model* 2007;**47**:2449–55.
63. Rindflesch TC, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. *J Biomed Inform* 2003;**36**:462–77.
64. Banville DL. Mining chemical structural information from the drug literature. *Drug Discov Today* 2006;**11**:35–42.
65. Garfield E. An algorithm for translating chemical names to molecular formulas. *J Chem Doc* 1962;**2**:177–9.
66. Hull RD, Singh SB, Nachbar RB, *et al*. Latent semantic structure indexing (LaSSI) for defining chemical similarity. *J Med Chem* 2001;**44**:1177–84.
67. Singh SB, Hull RD, Fluder EM. Text influenced molecular indexing (TIMI): a literature database mining approach that handles text and chemistry. *J Chem Inf Comput Sci* 2003;**43**: 743–52.
68. Hughes LD, Palmer DS, Nigsch F, *et al*. Why are some properties more difficult to predict than others? A study of QSPR models of solubility, melting point, and Log P. *J Chem Inf Model* 2008;**48**:220–32.
69. Zhang L, Zhu H, Oprea TI, *et al*. QSAR modeling of the blood-brain barrier permeability for diverse organic compounds. *Pharm Res* 2008;**25**.
70. Cai C, Xiao H, Yuan Q, *et al*. Function prediction for DNA-/RNA-binding proteins, GPCRs, and drug ADME-associated proteins by SVM. *Protein Pept Lett* 2008;**15**:463–8.
71. Bhasin M, Raghava GP. GPCRpred: an SVM-based method for prediction of families and subfamilies of G-protein coupled receptors. *Nucleic Acids Res* 2004;**32**: W383–9.
72. Cai CZ, Wang WL, Sun LZ, *et al*. Protein function classification via support vector machine approach. *Math Biosci* 2003;**185**:111–22.
73. Zhu S, Okuno Y, Tsujimoto G, *et al*. A probabilistic model for mining implicit 'chemical compound-gene' relations from literature. *Bioinformatics* 2005;**21**(Suppl 2):ii245–51.
74. Rindflesch TC, Tanabe L, Weinstein JN, *et al*. EDGAR: extraction of drugs, genes and relations from the biomedical literature. *Pac Symp Biocomput* 2000;517–28.
75. Rubin DL, Thorn CF, Klein TE, *et al*. A statistical approach to scanning the biomedical literature for pharmacogenetics knowledge. *J Am Med Inform Assoc* 2005;**12**:121–9.
76. Rindflesch TC, Pakhomov SV, Fiszman M, *et al*. Medical facts to support inferencing in natural language processing. *AMIA Annu Symp Proc* 2005;634–8.
77. Chen ES, Hripcsak G, Xu H, *et al*. Automated acquisition of disease drug knowledge from biomedical and clinical documents: an initial study. *J Am Med Inform Assoc* 2008;**15**:87–98.
78. Chen H, Sharp BM. Content-rich biological network constructed by mining PubMed abstracts. *BMC Bioinformatics* 2004;**5**:147.
79. Zhou D, He Y. Extracting interactions between proteins from the literature. *J Biomed Inform* 2008;**41**:393–407.
80. Hunter L, Lu Z, Firby J, *et al*. OpenDMAP: an open-source, ontology-driven concept analysis engine, with applications to capturing knowledge regarding protein transport, protein interactions and cell-specific gene expression. *BMC Bioinformatics* 2008;**9**:78.
81. Rinaldi F, Schneider G, Kaljurand K, *et al*. Mining of relations between proteins over biomedical scientific literature using a deep-linguistic approach. *Artif Intell Med* 2007;**39**:127–36.
82. Jose H, Vadivukarasi T, Devakumar J. Extraction of protein interaction data: a comparative analysis of methods in use. *EURASIP J Bioinform Syst Biol* 2007;53096.
83. Kim H, Park H, Drake BL. Extracting unrecognized gene relationships from the biomedical literature via matrix factorizations. *BMC Bioinformatics* 2007;**8**(Suppl 9):S6.
84. Alex B, Grover C, Haddow B, *et al*. Assisted curation: does text mining really help? *Pac Symp Biocomput* 2008; 556–67.
85. Southan C, Varkonyi P, Muresan S. Complementarity between public and commercial databases: new opportunities in medicinal chemistry informatics. *Curr Top Med Chem* 2007;**7**:1502–8.
86. Narayanasamy V, Mukhopadhyay S, Palakal M, *et al*. TransMiner: mining transitive associations among biological objects from text. *J Biomed Sci* 2004;**11**:864–73.

87. Swanson DR. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspect Biol Med* 1986;**30**:7–18.

88. Swanson DR. Migraine and magnesium: eleven neglected connections. *Perspect Biol Med* 1988;**31**:526–57.

89. Weeber M, Vos R, Klein H, *et al*. Generating hypotheses by discovering implicit associations in the literature: a case report of a search for new potential therapeutic uses for thalidomide. *J Am Med Inform Assoc* 2003;**10**:252–9.

90. Srinivasan P, Libbus B. Mining MEDLINE for implicit links between dietary substances and diseases. *Bioinformatics* 2004;**20**(Suppl 1):i290–6.

91. Ha S, Seo YJ, Kwon MS, *et al*. IDMap: facilitating the detection of potential leads with therapeutic targets. *Bioinformatics* 2008;**24**:1413–5.

92. McCray A. An upper level ontology for the biomedical domain. *Comp Funct Genom* 2003;**4**:80–4.

93. Mitchell JA, Aronson AR, Mork JG, *et al*. Gene indexing: characterization and analysis of NLM's GeneRIFs. *AMIA Annu Symp Proc* 2003;460–4.

94. 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.

95. Hoffmann R, Valencia A. Implementing the iHOP concept for navigation of biomedical literature. *Bioinformatics* 2005; **21**(Suppl 2):ii252–8.

96. Hirschman L, Yeh A, Blaschke C, *et al*. Overview of BioCreAtIvE: critical assessment of information extraction for biology. *BMC Bioinformatics* 2005;**6**(Suppl 1):S1.

97. Wheeler DL, Barrett T, Benson DA, *et al*. Database resources of the national center for biotechnology information. *Nucleic Acids Res* 2008;**36**:D13–21.

98. Fisher RA. On the interpretation of $\chi 2$ from contingency tables, and the calculation of P. *J R Stat Soc* 1922;**85**: 87–94.

99. Rajagopalan D, Agarwal P. Inferring pathways from gene lists using a literature-derived network of biological relationships. *Bioinformatics* 2005;**21**:788–93.

100. Russ AP, Lampel S. The druggable genome: an update. *Drug Discov Today* 2005;**10**:1607–10.