GENOMICS AND NATURAL LANGUAGE PROCESSING

Mark D. Yandell and William H. Majoros‡*

The Human Genome and MEDLINE are both the foci of intense data-mining efforts worldwide. The biomedical literature has much to say about sequence, but it also seems that sequence can tell us much about the biomedical literature. Biological natural language processing is an emerging field of research that seeks to explore systematically the relationships between genes, sequences and the biomedical literature as a basis for a new generation of data-mining tools.

CORPUS
A collection of documents that are used for searching or data mining.

BLAST, FASTA, ClustalW, HMM and PHYLIP — these bioinformatics algorithms are now a part of every molecular biologist's toolkit. DNA sequencing and data mining have become almost as central to biology as transcription and translation are to life. For most biologists, data mining is synonymous with sequence analysis — and understandably so, given the vast number of powerful sequence-analysis tools that are available. Some questions, however, cannot be answered by sequence analysis alone. As every biologist knows, there is much more to a gene than its sequence. For example, genes interact with other genes, they have complex temporal and spatial expression patterns, most have phenotypes and, in some cases, they are involved in disease. The question is: where should we go for such information about a sequence? One obvious destination is the bench; another is MEDLINE.

Most of what is known about genes and genomes is to be found in the biomedical literature. The Human Genome has been called the 'book of life', but surely MEDLINE is also a worthy contender for this title. At the last count, MEDLINE contained more than 11 million titles. Like the Human Genome, a CORPUS of this size can be explored and managed only by computational means. Today, the computational exploration and management of large text repositories are usually accomplished by using search engines and databases that are based on a suite of text processing, indexing and search tools — referred to collectively as natural language processing (NLP) technologies. Exploring and managing the biomedical literature using these technologies, however, presents some interesting challenges — primarily because of the relationships between biomedical texts and biological sequences. More than ever before, biomedical texts can be linked explicitly to the sequences of the genes they discuss, and the role of NLP technologies in biology is expanding and changing to reflect this fact. Anyone who has ever tried to read the MEDLINE abstracts that are associated with a BLAST report will appreciate that understanding the complex relationships that exist between genes, sequences and texts is a daunting task.

The flood of sequence information produced by the rapid advances in genomics is helping to provide new ways of exploring texts, and is blurring the traditional lines that separate bioinformatics and NLP. Information about genes is not only found in papers about genes, but also resides in the DNA, RNA and protein sequences that are associated with genes. The fact that so many texts and sequences are now available electronically naturally raises the question of how best to combine these two resources. For some time now, ENTREZ[1], a literature-search service, has provided a means of exploring this unique aspect of the biomedical literature, as navigating between sequences and texts often sheds new light on genes and their functions.

How best to exploit the synergies that exist between genes, sequences and texts is still an open question — to which there is not a single answer — and the diversity of research in this area reflects the open-ended nature of the problem. Some researchers are focusing on texts as a

**Howard Hughes Medical Institute, Department of Molecular and Cell Biology, Room 545, LSA Building No. 3200, University of California, Berkeley, California 94720-3200, USA. ‡The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, Maryland 20850, USA. Correspondence to M.D.Y.*
e-mail: myandell@fruitfly.org

doi:10.1038/nrg861

means of discovering information about protein interactions, and are wrestling with how best to adapt traditional NLP technologies to this task. Others, taking a more sequence-centred approach, are exploring the use of texts as a means of improving sequence-retrieval algorithms, and as an aid to sequence annotation. Although the research in this area is diverse, it shares a common substrate and a common goal: to use the relationships between genes, sequences and texts as the basis for a new generation of analysis tools and methodologies that combine bioinformatics and NLP technologies in a synergistic fashion. For brevity, we refer to these approaches as 'biological natural language processing', or bio-NLP.

Box 1 | Information retrieval using document vectors

In the vector-space model, each document in a corpus is represented as a list or weighted 'vector' of the words (or phrases) it contains. A portion of a document and its associated weighted vector are shown in (**a**) and (**b**), respectively. Each word that occurs in the list (**b**) has an associated weight, which is intended to represent the relative importance of that word in determining the theme of the document. This weight is usually some function of the frequency of the term in the document (term frequency, TF), so that terms that occur more often in the document will be given higher weights. To prevent common words with less semantic value from dominating the vector, term weights are typically normalized by IDF (the inverse document frequency of the term), which varies inversely with the frequency of the term in the corpus as a whole. Many formulations of the TF × IDF weighting scheme exist. One of these[5] is (1 + logTF) log(N/DF) for a corpus of size N, where DF is document frequency.

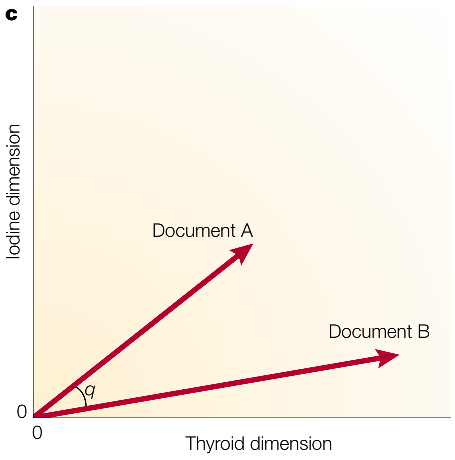
The individual weights in a vector collectively determine the direction that the vector will point in 'word space', wherein each dimension of the space corresponds to a single word or phrase in the document (**c**). In this way, the direction of the vector encodes the content of the underlying document. In **c**, two documents are shown in a two-dimensional space (for simplicity). The angle (q) between the vectors can be used to measure the similarity between the contents of the underlying documents.

a

Grave's disease is an autoimmune disease of the thyroid gland that affects the ability of the patient to retain and metabolize iodine

b

Term	Weight
Grave's	5.6
Thyroid	2.7
Gland	2.5
Iodine	2.3
Autoimmune	1.3
Disease	1.2
Patient	1.0
Metabolize	0.9
Retain	0.9
Affects	0.5
Ability	0.2
Is	0.0
An	0.0
Of	0.0
The	0.0



General aspects of NLP

Most biologists have more than a passing knowledge of bioinformatics. NLP, however, is probably new territory for most. There are three fundamental aspects to NLP: information retrieval, semantics and information extraction. Information retrieval refers to the recovery of documents from a database on the basis of their pertinence to a user's query. This is probably the most familiar aspect of NLP: anyone who has used PubMed to find documents about a given topic has made use of information-retrieval technologies. Traditional information-retrieval methodologies, however, are often frustrated by the complex terminologies that are used to describe genes and their functions. Making sense of such terminology is the goal of semantics, and it is usually accomplished by means of ontologies — that is, classification systems that relate concepts to one another. Google users might have noticed the 'category' heading that is placed at the top of each search-results page; listed beside this heading are the semantic classifications of the search results. This is an example of how ontologies can be used to classify documents and terms, and it is one of the strong points of this popular online search engine. Information extraction — the third basic component of NLP — is the extraction of ideas and concepts from a text, which is a process that is founded on effective semantic classification.

Information retrieval

Information retrieval is the process of returning documents in response to a search from a database that meet specific criteria. The term is used to describe two related tasks: document-based and query-based retrieval. Document-based retrieval scenarios usually take the form of "show me more documents like this one"; whereas query-based retrieval approaches attempt to recover documents that contain some combination of user-specified search terms, or keywords.

Keyword searches comprise the simplest form of information retrieval. Documents are often indexed on every term they contain, and keyword searches simply return all documents that contain at least one query term. Not surprisingly, this approach often returns irrelevant documents. So, most search engines go one step further and attempt to establish a measure of document relevance to a query. The most popular way of doing this is by means of the 'vector-space model' of information retrieval.

The vector-space model. Just as it is possible to quantify sequence similarity, it is also possible to quantify document similarity. In the vector-space model, each document is represented as a vector or list of weighted terms, based on the contents of the document (BOX 1a,b). The query is also 'vectorized', and the relevance of each document in the corpus that is queried is then assessed on the basis of the number of terms shared between a document and the query. Logically, the more often a query term occurs in a document, the more likely that the document is relevant;

ACCURACY
This frequently used term also has a formal definition. The accuracy of an algorithm is often defined as 2 × precision × recall/(precision + recall).

PSI-BLAST
A variation of BLAST that uses profiles that are based on sequence multiple-alignments to improve the sensitivity of protein database searches.

SWISS-PROT
A well-annotated database of protein sequences.

however, the more often a term occurs in the corpus as a whole, the less likely that the term is a reliable indicator of document relevance. These simple assumptions motivate most of the techniques that are used to weight terms in vector-based retrieval methods. The most common term-weighting strategy is TF × IDF, where TF is a measure of intra-document term frequency and IDF (inverse document frequency) is the inverse of the corpus-wide frequency of that term. Document–query similarities are calculated using these vectors of weighted terms. The most common procedure for assessing the similarity between documents is to calculate the angle between the document vectors (BOX 1c) — an approach that is used by ENTREZ[2] — but there are many others[3–5]. Initially, document-similarity metrics might seem to be more complex than sequence-similarity metrics, but this is not really the case: when all is said and done, most are simply scaled measures of the fraction of the words shared between two documents.

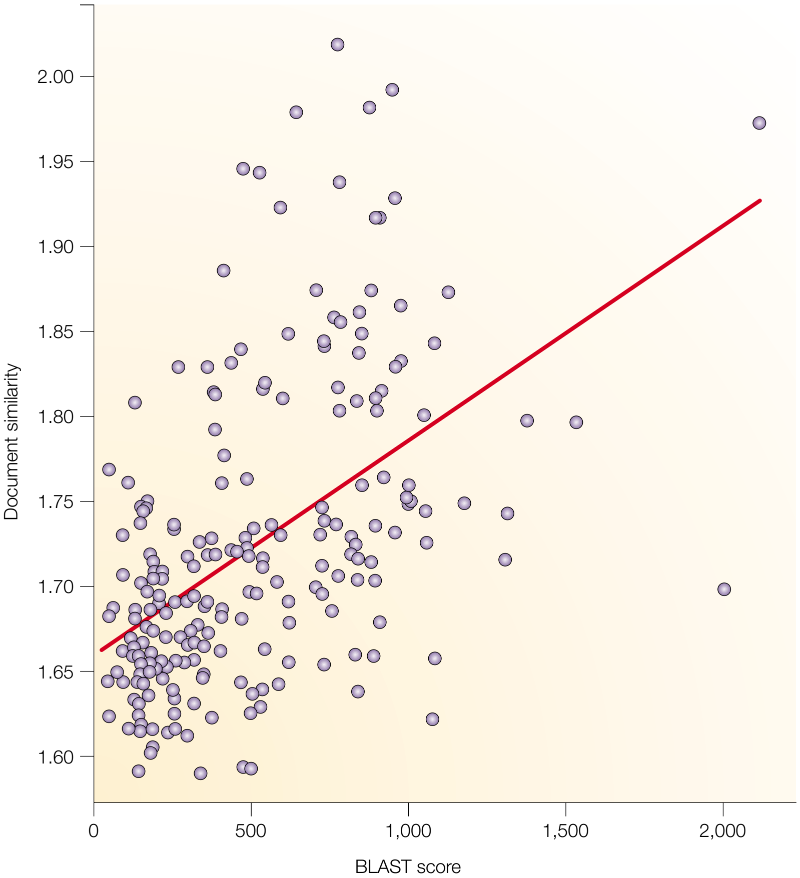


Figure 1 | **Correlation between sequence similarity and document similarity.** Linguistic similarity between documents often correlates with the similarity between sequences that are associated with those documents. A randomly chosen zinc-finger protein (gi:1022788) was used in a BLAST search against GenBank's 'non-redundant' protein database to recover a set of similar protein sequences. MEDLINE identifiers were obtained for these proteins from GenBank, and document-similarity scores were computed (BOX 1). The linear correlation coefficient between similarity of documents and similarity of the sequences that are associated with those documents was highly significant ($r = 0.51$, $P << 0.001$). The best-fitting line of regression is shown superimposed on the data points (slope = 1.3×10^{-4}, intercept = 1.67).

Measures of success. The success of an information-retrieval algorithm is usually measured in terms of precision and recall. These measures are identical to the specificity and sensitivity measures, respectively, that are used to benchmark gene-finder performance[6]. Recall is perfect if a query returns every pertinent document. Similarly, the precision of the algorithm is perfect if only these pertinent documents — and no others — are returned. There is usually some trade-off between the two measures, such that precision can be increased at the expense of recall or vice versa. It is difficult to generalize and say exactly what constitutes good precision and recall, as only a human reader can assess the pertinence of search results to a query. Note that, in this regard, information-retrieval algorithms differ from sequence-retrieval algorithms such as BLAST[7,8], which has very high precision and recall. The lower ACCURACY of information-retrieval algorithms has important implications, as users can never be certain that their search has really recovered all pertinent documents, nor that all of the documents recovered are pertinent; these are cautions to bear in mind when carrying out literature searches.

Better sequence retrieval and annotation

Most journals encourage authors to submit sequences to GenBank before publication. This has proven to be productive because it means that many papers are explicitly linked to the sequences that they describe. This is a unique feature of the biomedical literature, and it has some interesting ramifications. Because sequence similarity often implies similarity of biological function, tools such as BLAST and ENTREZ can be used to explore the biomedical literature through sequence–text links. These links comprise more than a mere navigational convenience, however, as there is often a significant correlation between sequence and document similarity (FIG. 1). In principle, this correlation can be used to improve the accuracy both of document- and of sequence-based retrieval algorithms.

Sequence retrieval. One approach to combining sequence and textual information is to develop sequence-retrieval algorithms that also incorporate textual information. For example, SAWTED[9] is a modified version of PSI-BLAST[8]. It combines vector-space similarity scores based on SWISS-PROT[10] protein-sequence annotations with BLAST similarity scores, to improve precision and recall. In similar work, Chang and colleagues[11] have described a variation of PSI-BLAST that uses literature information linked to SWISS-PROT records, together with MEDLINE cross-references and MeSH headings (a document classification system; see BOX 2 for a list of online resources). Their algorithm excludes from the successive rounds of a PSI-BLAST search those sequences for which related literature (based on their vector-space similarity scores) is least concordant with the literature associated with the original query sequence. This approach increased PSI-BLAST precision from 0.84 to 0.95 (where 1 is 100% precisely) and had a negligible effect on recall, which

Box 2 | Online resources for biological natural language processing

BIND. . . . . http://www.bind.ca
The Biomolecular Interaction Network Database: an excellent resource for exploring protein–protein interactions, with a very extensive set of links to related resources.

Bio-Ontology pages. . . . . http://img.cs.man.ac.uk/stevens/ontology.html
A good 'jumping-off' point for those interested in finding out more about the role of ontologies in biology.

BLAST and related programs . . . . http://www.ncbi.nlm.nih.gov/Sitemap/index.html#BLAST
The BLAST pages at the National Center for Biotechnology Information (NCBI) do much more than simply provide a means to carry out a BLAST search. They are also an excellent source of information about BLAST and related programs such as PSI-BLAST.

DIP . . . . http://dip.doe-mbi.ucla.edu
Database of Interacting Proteins: a popular database that contains protein–protein interaction data.

ENTREZ. . . . . http://www.ncbi.nlm.nih.gov/Entrez
An excellent starting point from which to explore links between sequences, texts and more.

Foundations of Statistical Natural Language Processing: . . . . . http://nlp.stanford.edu/fsnlp
The companion web site for the book by the same name and an indispensable resource for anyone who is interested in natural language processing.

GO . . . . http://www.geneontology.org
Gene Ontology: a tool for the unification of biology, and a very intuitive and usable ontology of genes.

INTERACT . . . . http://www.bioinf.man.ac.uk/resources/interact.shtml
A database of protein–protein interactions that is now part of the UMBER (University of Manchester Bioinformatics Education Research) project. It has interesting three-dimensional views of protein–protein interactions.

MeSH . . . . http://www.nlm.nih.gov/mesh/meshhome.html
The Medical Subject Heading system that is used by the National library of Medicine to classify subjects that occur in the biomedical literature.

MetaMap . . . . http://ii.nlm.nih.gov/MTI/mmi.shtml
An informative page detailing how this indexing methodology works.

PubGene. . . . . http://www.pubgene.uio.no
The web server for the work discussed in Jenssen *et al.*[41] (see main text), with nice tie-ins to Gene Ontology and other resources.

PubMed/MEDLINE . . . . http://www.ncbi.nlm.nih.gov/entrez/query.fcgi
The central online source for biomedical literature searching and reference collection.

SAWTED . . . . http://www.bmm.icnet.uk/~sawted
Structure Assignment With Text Description: carries out text-assisted BLAST searches for distant homologues. It also has an excellent diagram that explains how the procedure works.

UMLS (Metathesarus, Specialist Lexicon and related tools) . . . http://www.nlm.nih.gov/research/umls
The Unified Medical Language System — the National Library of Medicine's biological ontology.

WordNet. . . . . http://www.cogsci.princeton.edu/~wn
A lexical database for the English language, and a great site to explore for anyone who is interested in the English language.

shows that literature similarities can be combined with sequence similarities to improve the specificity of sequence-retrieval algorithms. Although both groups report that the use of text information improved PSI-BLAST performance, one shortcoming of both studies is that text similarities were combined with BLAST similarity scores in a heuristic fashion. The broader applicability of such hybrid approaches would benefit from a statistically rigorous (and as-yet-unformulated) means of combining sequence and text similarities.

The correlation between sequence and text similarities has also been used to facilitate the functional annotation of proteins. Two groups have explored the use of sequence-associated texts to classify protein sequences according to their subcellular locations. Eisenhaber and Bork[12] developed an algorithm, called Meta_A(nnotator), which classifies proteins on the basis of terms contained in their SWISS-PROT annotations, whereas Stapley *et al.*[13] classified *Saccharomyces cerevisiae* proteins on the basis of information contained in their associated MEDLINE abstracts. Both approaches did well: Eisenhaber and Bork report that their *AD HOC* RULE-BASED APPROACH was able to provisionally classify 88% of the protein sequences contained in SWISS-PROT release 34, whereas only 22% of the proteins in this release could be provisionally assigned a cellular location on the basis of simple keyword searches of their associated annotations for terms such as 'membrane' and 'extracellular matrix'. Stapley *et al.* evaluated the effectiveness of their own approach, and reported that the accuracy of assignment to various cellular locations differed greatly, ranging from 0.30 for extracellular or secreted proteins to 0.80 for proteins that are involved in nuclear organization. Interestingly, Stapley *et al.* found that the inclusion of sequence information in the form of amino-acid composition considerably improved assignment accuracy for most compartments, once again illustrating the utility of hybrid sequence–text approaches. Both groups also point out that one of the benefits of automated approaches is that they provide a high-throughput means of identifying errors and inconsistencies in manual sequence annotations. This is

AD HOC RULE-BASED APPROACHES
These are approaches for identifying terms in a text that belong to a particular semantic class. Gene names in *Caenorhabditis elegans*, for example, are denoted with three letters followed by a dash and a number — for example, '*dbl-1*'. So, this approach to identify *C. elegans* genes might consist of searching a text for regular expression of three letters, a dash and a number. Such approaches do not work equally well for identifying all genes and generally are not very precise.

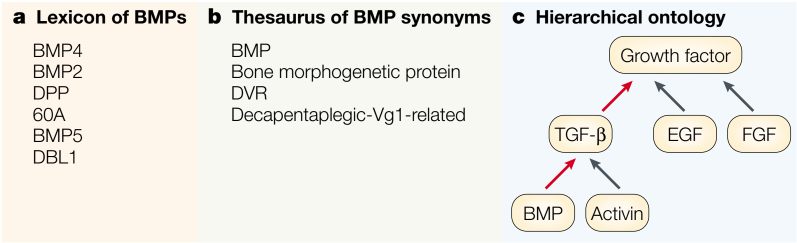


Figure 2 | **Semantic classification and definition of terms using a lexicon, thesaurus and a hierarchical ontology.** Lexica, thesauri and ontologies are used to semantically classify and define terms that occur in a text. At its simplest, a lexicon is merely a list of terms that belong to the same semantic class: bone morphogenetic protein 4 (BMP4) and decapentaplegic (DPP), for example, both belong to the semantic class of 'BMP'. A thesaurus provides a listing of the synonyms for a term, or semantic class, and hierarchical ontologies are used to 'define' the terms that are contained in a lexicon and thesaurus. The definition of a term is produced by tracing the path from a term to the root of the ontology (**c**, path shown in red). The simple ontology shown in **c**, for example, defines BMP as "a TGF-β growth factor". Definitions apply to all members of a semantic class and their synonyms, and can be used as a basis for logical inference: "DPP is a DVR, a DVR is a BMP, a BMP is a TGF-β, and a TGF-β is a growth factor"; therefore, it can be inferred that "DPP is a growth factor", even if no document explicitly states this fact. Note that the ontology shown in **c** is a particular type known as an isa-hierarchy; other types of ontology exist, not all of which are suitable for definition[19,55]. DBL1, decapentaplegic–BMP-like 1; DVR, decapentaplegic-Vg1-related; EGF, epidermal growth factor; FGF, fibroblast growth factor; TGF-β, transforming growth factor-β.

a salient point: the use of bio-NLP tools to flag inconsistent annotations for manual review holds much promise for improving sequence annotation.

Sequence annotations. Sequences are routinely clustered into families on the basis of sequence-similarity measures; in the same way, text-similarity measures can be used to cluster documents. Iliopoulos and colleagues[14], for example, clustered similar MEDLINE documents, and then annotated these document clusters by attaching keywords to them using an ODDS RATIO. Relationships between sequences and texts add a new twist to sequence and document clustering, as sequence homology can be used to assemble groups of documents and vice versa. Several groups have used sequence–text links as a means of facilitating protein-sequence annotation. Andrade and Valencia[15] automatically annotated protein families with statistically salient keywords and sentences that were extracted from documents associated with their protein members. Similarly, Renner and Azodi[16] have described a procedure that annotates clusters of expressed sequence tags (ESTs) that correspond to previously annotated sequences on the basis of the similarities of their associated MEDLINE abstracts. In related work, Raychaudhuri *et al.*[17] automatically assigned genes a function in the GENE ONTOLOGY[18] on the basis of information contained in MEDLINE abstracts; they reported an accuracy of 0.72, showing that text and ontologies can be used together to make automatic predictions of protein function.

In evaluating the performance of these approaches, it is helpful to distinguish between term and concept identification. Methods that annotate sequences solely on the basis of term frequency are problematic: the terms most overrepresented among the documents that are associated with a gene are, by definition, the most representative terms; whether or not they convey useful information to a biologist is another matter. An important strength of the work of Raychaudhuri *et al.*[17] lies in its use of the Gene Ontology[18]. Annotating genes with concepts drawn from an ONTOLOGY ensures consistency and future utility, as these concepts convey useful semantic information (see below) that can be unambiguously communicated to both software and its users.

Semantics

MEDLINE is a corpus of unrivalled complexity. At last count, the 11,450,302 abstracts that comprise MEDLINE contained 1,130,223,622 words, 1,614,538 of which were distinct. By comparison, Darwin's *Origin of Species* contains 207,158 words, of which only 18,285 are distinct. Although size and complexity might not be any indication of profundity, obviously the biomedical literature is rife with jargon, synonyms and equivocal terms, the identification and definition of which are broadly termed 'semantics'.

Biomedical jargon has a massive impact on the verisimilitude of vector-based similarity measures. In general, keyword and vector-space retrieval technologies are semantically blind: they return related documents solely on the basis of shared — perfectly identical — terms, regardless of the 'meaning' of those terms. In practice, this means that, given a document or query that contains the terms 'polymerase chain reaction' and '*PAX6*' (a conserved family of homeodomain transcription factors), naive information-retrieval algorithms will return documents that contain one or both of these terms, but not necessarily documents that contain the terms 'PCR', '*eyeless*' (the *Drosophila melanogaster PAX6* ORTHOLOGUE) or 'aniridia' — a form of hereditary blindness that is caused by heterozygous mutations in *PAX6*. Obviously, this is far from ideal. One way to improve the recall of information-retrieval algorithms, therefore, is to use ontologies to extract semantic information from texts.

Ontologies. NLP technologies often use structured lists of terms called ontologies[19] (FIG. 2) to establish the semantic function of a word in a document. The simplest form of ontology is a lexicon or a list of terms that belong to a particular class. A list of gene names, for example, can be used to identify documents that discuss genes; likewise, a lexicon of diseases can be used to identify documents that discuss disease. Lexica usually consist of specialized terms and (optionally) their definitions, but this is not always the case; some are more general. WordNet[20], a much-used ontology, contains extensive semantic information about English words that goes far beyond that provided by typical dictionaries.

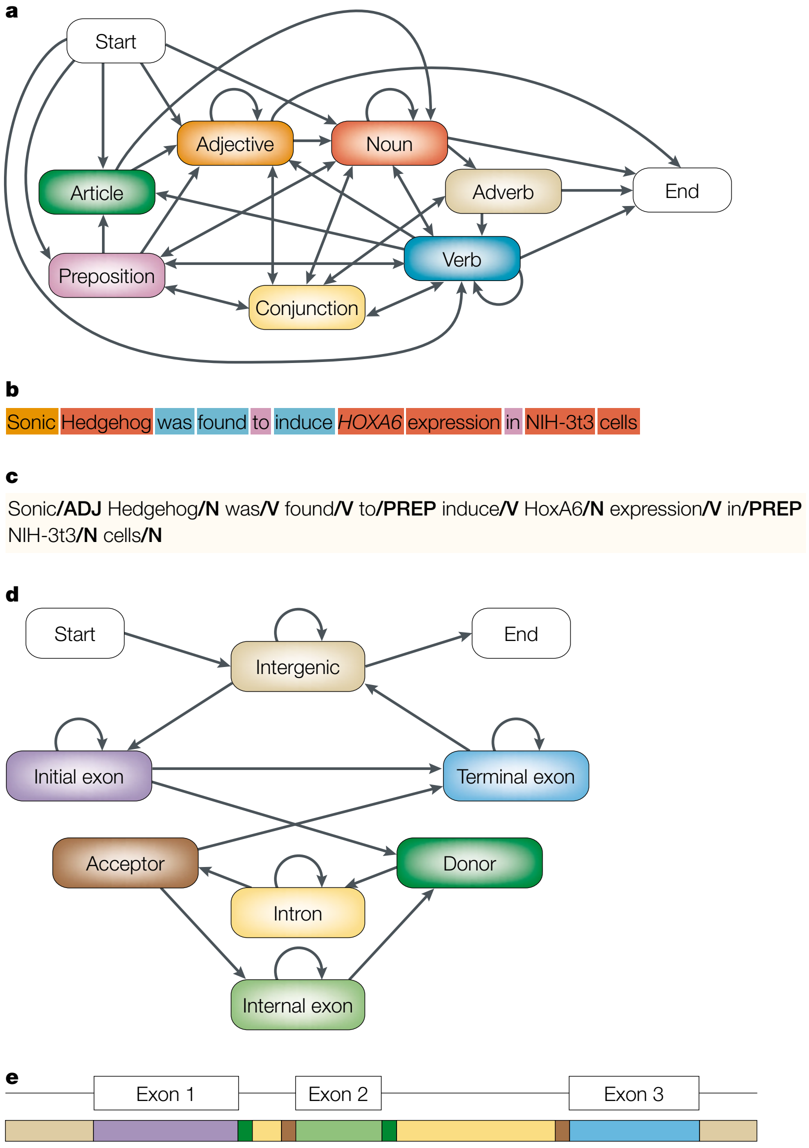
A collection of terms and their synonyms is called a thesaurus. Synonym information is of immense utility for NLP. A thesaurus of genes, for example, can be used to classify a gene's name and its symbol as synonyms, thereby improving the verisimilitude of document-similarity measures.

ODDS RATIO
The ratio between the observed frequency at which an event occurred and the expected frequency of that event given some statistical model. A term that occurs more frequently in a text, or collection of texts, than would be expected based on its frequency in a corpus will therefore have an odds ratio >1.

GENE ONTOLOGY
(GO). A hierarchical organization of concepts (ontology) with three organizing principles: molecular function, the tasks done by individual gene products, an example of which is 'transcription factor'; biological process, broad biological goals, such as mitosis, that are accomplished by ordered assemblies of molecular functions; cellular component, subcellular structures, locations and macromolecular complexes (examples include the nucleus and the telomere).

ONTOLOGY
A hierarchical organization of concepts, typically used to denote 'more-general-than' and/or 'part-of' relationships.

ORTHOLOGUES
Homologous genes that originated through speciation (for example, human β-globin and mouse β-globin).



Sonic Hedgehog was found to induce *HOXA6* expression in NIH-3t3 cells

Sonic/**ADJ** Hedgehog/**N** was/**V** found/**V** to/**PREP** induce/**V** HoxA6/**N** expression/**V** in/**PREP** NIH-3t3/**N** cells/**N**

Figure 3 | **HMMs are used for part-of-speech tagging, as well as for gene prediction.** **a**–**c** | In natural language processing (NLP), the grammar (**a**) of a language can be modelled to determine the grammatical function or 'part-of-speech' of each word in a sentence. For example, the most likely path through the model sentence shown in **b** is used to assign or 'tag' the words of the sentence with their part-of-speech (noun, adjective and so on; **c**). This process uses statistical algorithms called hidden Markov models (HMMs), which are used both in NLP and in bioinformatics. In general, HMMs can be thought of as a probabilistic model of an abstract 'source' that emits a sequence of symbols; in other words, as the 'author' of a sentence or gene. To produce a sequence of symbols, the source is modelled as passing stochastically through a finite sequence of discrete states (shown as coloured rectangles), beginning at the 'start state' and ending at an 'end state'. On entering a new state, a single symbol is produced according to the emission probabilities that are associated with that state. The order in which states might be visited is constrained by the allowable transitions of the model (shown as arrows), and by the probabilities that are associated with those transitions (not shown, for clarity). In this way, the source is modelled as emitting different sequences with different probabilities. **d**,**e** | HMMs are also used in gene prediction. An HMM model such as that represented in **d** can be used to calculate the probability that a given base of DNA resides in any of seven illustrated 'states': 'initial exon', 'terminal exon', 'internal exon', 'intron', 'donor', 'acceptor' or 'intergenic' sequence. These probabilities are then used to infer the structure of the gene (**e**). HMM models (**a**,**d**) and their associated transition and emission probabilities are inferred from training examples — that is, from known sentences and genes. Once determined, these probabilities are used to determine the most likely path through the model that would produce the given sentence or gene. ADJ, adjective; N, noun; PREP, preposition; V, verb.

Although lexica and thesauri are technically ontologies, the term ontology generally implies a hierarchical organization of terms in which concepts higher up in the hierarchy (hypernyms) are more general than those lower down (hyponyms). One popular NLP use for hierarchical ontologies is query expansion — the use of an ontology to append some combination of synonyms, hypernyms and hyponyms to a user's query to retrieve documents that use related terms to describe the same ideas. Given a query term 'growth factor', for example, a hierarchical ontology can be used to 'expand' the user's query to include the hyponyms 'TGF-β' (transforming growth factor-β), 'EGF' (epidermal growth factor) and 'FGF' (fibroblast growth factor), thereby allowing the retrieval of documents that contain these relevant terms (FIG. 2).

Concept identification in texts. The Unified Medical Language System (UMLS) Metathesarus[21] and its related resources, MeSH and the Specialist Lexicon (BOX 2) comprise a widely used system of biomedical ontologies[22]. These resources contain information about many aspects of the biomedical domain, such as diseases, tissues and drugs. One limitation of using ontologies for NLP is the inherent difference between the controlled vocabulary used in an ontology to describe a concept and the terms actually used by authors to describe that concept in text. An author might refer to a disease as 'type II diabetes mellitus', whereas an ontology might describe the same concept as 'diabetes, type II, mellitus'. Although this might seem a minor annoyance, in practice it presents a significant hurdle for software used to search texts for concepts that are contained in an ontology. Several groups have explored this problem with regard to UMLS[23–25]. The MetaMap algorithm[26,27], for example, uses a PART-OF-SPEECH TAGGER[5,28] (FIG. 3) to identify noun phrases in text, and then attempts to map them to a similar UMLS concept. One indication of the difficulties that are associated with accurate concept identification in free text is that the use of UMLS for query expansion can significantly reduce the precision of document retrieval, with little improvement in recall[29]. It is not clear from the study whether this was due to problems with mapping concepts in free text to UMLS, or to inconsistencies in the UMLS ontology itself. There are internal errors and omissions[30] in UMLS and two groups have recently encountered problems when attempting to use UMLS for query expansion and knowledge mining[23,31].

Gene identification in texts. Genes and proteins are concepts, too; however, identifying them in text is proving to be difficult. Gene-naming conventions differ markedly between organisms, and there exists no single authoritative ontology that defines the terminologies used to name and describe genes, gene families and their products. One obvious solution is to compile such a lexicon manually, but this is time consuming, as well as error prone. Several groups have therefore attempted to identify gene and protein names automatically, for example, using general *ad hoc* rule-based approaches[32–34].

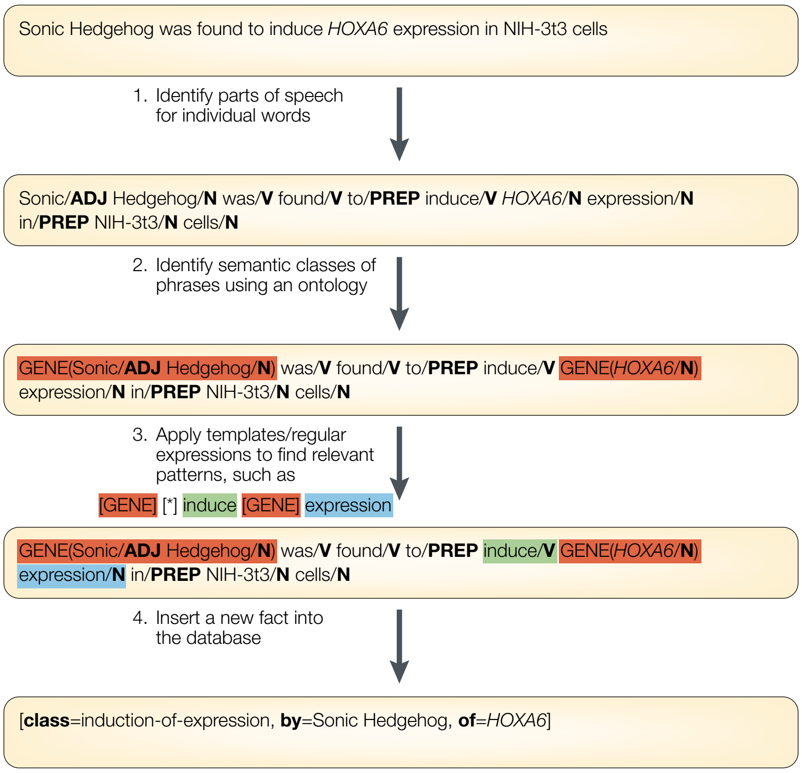

Sonic Hedgehog was found to induce HOXA6 expression in NIH-3t3 cells
1. Identify parts of speech for individual words
Sonic/ADJ Hedgehog/N was/V found/V to/PREP induce/V HOXA6/N expression/N in/PREP NIH-3t3/N cells/N
2. Identify semantic classes of phrases using an ontology
GENE(Sonic/ADJ Hedgehog/N) was/V found/V to/PREP induce/V GENE(HOXA6/N) expression/N in/PREP NIH-3t3/N cells/N
3. Apply templates/regular expressions to find relevant patterns, such as
[GENE] [*] induce [GENE] expression
GENE(Sonic/ADJ Hedgehog/N) was/V found/V to/PREP induce/V GENE(HOXA6/N) expression/N in/PREP NIH-3t3/N cells/N
4. Insert a new fact into the database
[class=induction-of-expression, by=Sonic Hedgehog, of=HOXA6]

Figure 4 | **Information extraction.** The extraction of information typically entails the use of patterns or templates to extract structured information from unstructured text. One common approach is illustrated here. First, individual words are tagged with their parts of speech, such as nouns and verbs (step 1; see also FIG. 3). On the basis of their parts of speech, these words can then be grouped into phrases and resolved to specific concepts in a biological ontology (step 2; see also FIG. 2). In this example, 'Sonic Hedgehog' (*shh*) and '*HOXA6*' (highlighted in red) have been identified as genes. Next, manually crafted templates, or regular expressions are applied to identify sentences that contain key syntactic patterns (step 3). These patterns are then inserted into a database of extracted facts (step 4), where they can be used as the basis for logical inference (see FIG. 2), or simply indexed for querying. Manual or semi-automatic curation and filtering are often necessary to remove errors. ADJ, adjective; HOXA6, homeobox A6; N, noun; PREP, preposition; V, verb.

PART-OF-SPEECH TAGGER
An algorithm that identifies the nouns, verbs and other functional word classes among the words that comprise a sentence.

PARALOGUES
Homologous genes that originated by gene duplication (for example, human β-globin and human α-globin).

Although rule-based approaches work well for identifying genes and proteins from organisms such as *S. cerevisiae*, where these terms strictly adhere to a controlled vocabulary[35], others have found it necessary to supplement rule-based approaches with a lexicon of gene names[36,37]. In related work, Hatzivassiloglou and colleagues[38] developed an algorithm to determine automatically (or to 'disambiguate') whether the occurrence of a gene name in a text refers to the gene itself, its mRNA or its protein product.

Ontologies, homologies and texts. Although automatic tools for identifying gene names in free text might facilitate some bio-NLP applications, accurate retrieval and extraction of information necessitates cross-indexing documents that reference one gene to another. Synonym lists are useful in this respect, but they neglect an equally essential relationship between genes — that is, homology. Sequences come with their own ready-made ontology in the form of their homologous relationships to one another. In principle, this information can be used to build gene thesauri on the basis of homology, as well as synonymy. To return for a moment to an earlier example, homology provides a natural means of recovering documents about *eyeless* from the query 'PAX6'. In principle, information about homologous, PARALOGOUS and orthologous relationships between genes could be used to improve the precision and recall of information-retrieval algorithms.

Information extraction

The key to understanding a text is knowing what the words mean. Likewise, if software is to interpret a text, it must also be able to identify words with respect to their semantic classes, definitions and syntactic functions. Information extraction is the process of using this information to extract meaning from a text[5,39,40]. Towards this end, information extraction normally combines part-of-speech tagging (FIG. 3), ontologies (FIG. 2) and REGULAR EXPRESSIONS (FIG. 4) to produce a structured, machine-readable file that contains essential information gleaned during the extraction phase; ideally, these files are structured in such a way as to facilitate logical inference and information retrieval.

Identifying interactions. Recently, protein–protein interactions have been the focus of many bio-NLP information-extraction efforts. MEDLINE abstracts often contain important information about protein–protein interactions. Jenssen *et al.*[41] constructed a protein–abstract index and then used it to identify possible protein–protein interactions on the basis of the co-occurrence of gene and protein names in abstracts (TABLE 1). They also supplemented this index by further cross-indexing these proteins to MeSH and Gene Ontology terms. Others, in an attempt to classify better the nature of the co-occurrence[40,42,43], have used information-extraction algorithms to identify co-occurring protein names in the context of 'interaction' verbs (such as 'represses', 'phosphorylates' and 'binds') within individual abstracts (FIG. 4). Wong[37] and Blaschke and colleagues[44] use similar methodologies, but attempt to circumvent problems that are associated with the identification of gene and protein names by providing an interface that asks the user to enter keywords.

Table 1 | **Co-occurrence detection**

Gene A	Gene B	#A	#B	#(A,B)
BMP4	*BMP2*	324	618	50
BMP4	*HLP3*	324	128	11
BMP4	*TSG*	324	242	3
BMP4	*DPP*	324	685	4

Many types of association between biological entities can be gleaned from text simply by noting how often words and phrases that belong to a particular semantic class occur together in a corpus (in this case, MEDLINE Abstracts). The co-occurrence of terms in the same sentence or the same document often implies real biological relationships between the named entities[41]. The nature of these relationships can be explored further using ontologies. The ontology shown in FIG.2, for example, can be used to identify the terms BMP4, BMP2, and DPP as 'growth factors'. BMP, bone morphogenetic protein; DPP, decapentaplegic HLP3, human placental lactogen 3; TSG, twisted gastrulation.

Some researchers have taken a broader approach to information extraction in an attempt to do more than simply identify protein–protein interactions. Stapley and Benoit[35] tallied the number of co-occurrences of every pair of *S. cerevisiae* genes in MEDLINE abstracts and used this data to calculate what they term a BioBibliometric distance between genes, such that the rarer the co-occurrence of two genes in the literature database, the larger the distance between them will be. Such text-based similarity measures are especially interesting as they provide a means to assess the similarity of genes, independently of sequence homology. Leroy and Chen[45] chose to explore the prepositional ('by' and 'of') relationships of genes in an attempt to extract a wider assortment of information than that provided by 'interaction' verbs alone; Hahn *et al.*[31] have attempted to go still further and describe a very general approach for biomedical information extraction; and Rindflesch and colleagues[46] describe an attempt to extract from texts information about the relationships between genes, drugs and cells.

Caveats. The basic motivation behind all of these approaches is that the co-occurrence of gene and protein names in abstracts implies a biological relationship. There are, of course, many cases in which co-occurrences are not indicative of interaction. Negation is one trivial reason. Although it is relatively rare, information-extraction algorithms need to take negation into account as well, lest co-occurrences in the form of "A was found *not* to interact with B" be misconstrued[47]. 'Anaphora' resolution, the process of determining what a pronoun refers to, must be handled as well[5,48], as informative phrases often take the form of "*It* was found to bind DNA".

It must also be recognized that genes and proteins participate in many kinds of relationships that are beyond the merely physical. Mutations in two genes might cause similar phenotypes; their expression patterns might overlap; the two genes might lie near one another in the genome; or they might be orthologues, paralogues or simply homologues of one another. The use of ontologies such as Gene Ontology to characterize better the reasons for co-occurrence is one promising avenue for further research.

Co-occurrence, of course, is not the only source of interaction data. Yeast two-hybrid and other experimental procedures often generate direct experimental information about gene and protein interactions. Similarly, many papers explicitly describe genetic interactions and such information can be extracted by manual curation of full-text articles. The INTERACT[49], DIP (Database of Interacting Proteins)[50] and BIND (Biomolecular Interaction Network Database)[51] databases consist of interaction data obtained through experimental procedures. Databases such as these provide an important means by which to benchmark the performance of information-extraction approaches to interaction discovery, and two groups[41,52] have used DIP for exactly this purpose.

Jenssen *et al.*[41], in a test of the precision of their algorithm (PubGene; see BOX 2), found that only 51% of their interactions were recorded in DIP. Further examination revealed that incorrect associations were due primarily to a lack of precision in correctly identifying gene and protein names, and that co-occurrences often described homologous relationships rather than physical interactions. In a similar analysis, Blaschke *et al.*[52] found that approximately two-thirds of their failures were also due to the inconsistent use of gene and protein nomenclature. The degree to which both approaches were frustrated by nomenclature issues underscores the fact that bio-NLP is a field badly in need of better ontologies and automated means to identify these terms in texts.

Both groups also used DIP to evaluate the performance of their algorithms in terms of recall. In this respect, the two approaches differed significantly. Jenssen *et al.* estimate that very few interactions (7.7%) were missed due to lack of co-occurrence in the title or abstract. Blaschke *et al.*, however, report that the absence of interaction information in abstracts decreased recall by 35%. Algorithmic differences are one likely reason for the disparity in recall between the two studies, as Jenssen *et al.* relied on simple co-occurrence, whereas Blaschke *et al.* used a specific set of regular expressions (FIG. 4) to detect and classify interactions.

Future directions

The exponential growth of MEDLINE and GenBank is rapidly transforming bio-NLP from a research endeavour into a practical necessity. Most of the studies discussed in this review are focused on information management, and the development of such tools is a necessary and laudable goal. Nevertheless, if bio-NLP is to achieve its full potential, it will have to move beyond information management and generate specific predictions that pertain to gene function that can be verified at the bench. The synergistic use of sequence and text to extract latent information from the biomedical literature holds much promise in this regard. Realizing this potential, however, will require more and better ontologies, software able to make inferences using sequence and textual information, and access to the full text of articles.

More and better ontologies. A greater diversity of high-quality biomedical ontologies that are designed with NLP applications in mind would do much to strengthen the field. The identification of gene and protein names, for example, was an important factor frustrating much of the research discussed in this review. Accurate semantic classification of terms that occur in the molecular biological literature will require ontologies of genes, drugs, diseases and molecular biology procedures, terminologies used in genetics and in population biology, and ontologies of tissues and phenotypes — to name but a few. As discussed previously, not all ontologies are well suited for use by NLP applications because their controlled vocabularies do not reflect actual usage in text, and methods to adapt

REGULAR EXPRESSION
Computer science parlance for an abstract definition that embodies some common and essential syntactic characteristic that belongs to a set of terms. For example, in the popular PERL programming language, the regular expression '\s* \w+\–\d+\s*' will identify any word in a text that consists of one or more letters (or numbers), followed by a dash, and followed by one or more numbers. This regular expression will identify *Caenorhabditis elegans* gene names.

them for use by bio-NLP applications are needed. When no suitable ontology exists, automated approaches to ontology construction[53] are a promising avenue for further research.

Inference and sequence homology. Inductive reasoning (FIG. 2c) based on information extracted from texts is the ultimate goal of information extraction. Sequence provides a means for inference that is unique to bio-NLP, as sequence homology can be used to uncover latent information in the biomedical literature. If, for example, an abstract reports that two mouse proteins interact, their human orthologues will probably interact as well. Both DIP and BIND (see BOX 2) have begun to use sequence homology to flag potential protein–protein interactions, but there remains much opportunity for significant research in this area. Applications that use relationships between sequences, the domains they contain, phylogenetics and textual information as a means to automatically generate experimentally verifiable predictions of gene expression, function and interactions are a logical next step for the field.

Access to full text. Greater access to full text would do much to help bio-NLP realize its potential for hypothesis generation. Blaschke and colleagues[52] conclude their report with the observation that abstracts often contain insufficient information to characterize protein–protein interactions adequately, although this information is usually present in the body of the article. The number of additional protein–protein interactions that could be found by their approach using full text is unknown, but it seems reasonable to conclude that access to full text would improve the precision and recall of their algorithm. By extension, the restricted information available in abstracts probably imposes an unnecessary handicap on the performance of information-extraction technologies in general. Recently, the Public Library of Science (PLoS)[54] has advocated the construction of a publicly available and machine-readable full-text repository for the scientific literature. Access to full text would do much to help bio-NLP realize its potential. Imagine, for a moment, the possibilities opened by such a resource, what the availability of full text could mean in terms of searchable indices of figures, tables, citations and photographs, and in terms of the accessibility of information contained in 'materials and methods' sections — a full-text repository would fuel substantial advances in bio-NLP. Where would bioinformatics be today without GenBank? Without a public library of science, much scientific knowledge is lost to algorithms and researchers alike.

1. Schuler, G. D., Epstein, J. A., Ohkawa, H. & Kans, J. A. Entrez: molecular biology database and retrieval system. *Methods Enzymol.* **266**, 141–162 (1996)
2. Wilbur, W. J. & Yang, Y. An analysis of statistical term strength and its use in the indexing and retrieval of molecular biology texts. *Comput. Biol. Med.* **26**, 209–222 (1996).
   Describes the vector-space model used by Entrez, the literature-search service maintained by the NCBI.
3. Renner, A. & Aszodi, A. High-throughput functional annotation of novel gene products using document clustering. *Proc. Pacific Symp. Biocomp.* **5**, 54–68 (2000).
4. Shatkay, H., Edwards, S., Wilbur, W. J. & Boguski, M. Genes, themes, and microarrays. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **8**, 317–327 (2000).
5. Manning, C. D. & Schutze, H. S. in *Foundations of Statistical Natural Language Processing* 85 (MIT press, Cambridge, Massachusetts, 1999).
   The indispensable reference for anyone who is interested in statistical natural language processing (NLP).
6. Burset, M. & Guigo, R. Evaluation of gene structure prediction programs. *Genomics* **34**, 353–367 (1996).
7. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
8. Altschul, S. F. *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. **25**, 3389–3402 (1997).
9. MacCallum, R. M., Kelley, L. A. & Sternberg, J. E. SAWTED: structure assignment with text description — enhanced detection of remote homologues with automated SWISS-PROT annotation comparisons. *Bioinformatics* **16**, 125–129 (2000).
10. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999. *Nucleic Acids Res*. **27**, 49–54 (1999).
11. Chang, J. T., Raychaudhuri, S. & Altman, R. B. Including biological literature improves homology search. *Proc. Pacif. Symp. Biocomp.* **5**, 374–383 (2001).
    A quantitative assessment of the utility of combining sequence similarity with document similarity.
12. Eisenhaber, F. & Bork, P. Evaluation of human-readable annotation in biomolecular sequence databases with biological rule libraries. *Bioinformatics* **15**, 528–535 (1999).
13. Stapley, B. J., Kelley, L. A. & Sternberg, M. J. E. Predicting the sub-cellular location of proteins from text using support vector machines. *Proc. Pacif. Symp. Biocomp.* (in the press).
    Describes the use of both text and sequence data to predict subcellular localization.
14. Iliopoulos, I., Enright, A. J. & Ouzounis, C. A. TEXTQUEST: document clustering of Medline abstracts for concept discovery in molecular biology. *Proc. Pacif. Symp. Biocomp.* **6**, 374–383 (2001).
15. Andrade, M. A. & Valencia, A. Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families. *Bioinformatics* **14**, 600–607 (1998).
16. Renner, A. & Aszodi, A. High-throughput functional annotation of novel gene products using document clustering. *Proc. Pacific Symp. Biocomp.* **5**, 54–68 (2000).
17. Raychaudhuri, S., Chang, J. T., Sutphin, P. D. & Altman, R. B. Associating genes with gene ontology codes using a maximum entropy analysis of biomedical literature. *Genome Res.* **12**, 203–214 (2002).
18. The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. *Nature Genet.* **25**, 25–29 (2000).
19. Stevens, R., Goble, C. A. & Bechhofer, S. Ontology-based knowledge representation for bioinformatics. *Brief Bioinform.* **1**, 398–414 (2000).
20. Fellbaum, C. (ed.) *WordNet: an Electronic Lexical Database* (MIT Press, Cambridge, Massachusetts, 1999).
21. Humphreys, B. L., Lindberg, D. A., Schoolman, H. M. & Barnett, G. O. The Unified Medical Language System: an informatics research collaboration. *J. Am. Med. Inform. Assoc.* **5**, 1–11 (1998).
22. Baclawski, K., Cigna, J., Kokar, M. M., Mager, P. & Indurkhya, B. Knowledge representation and indexing using the unified medical language system. *Proc. Pacif. Symp. Biocomp.* **5**, 502–513 (2000).
    A brief introduction to UMLS and related issues.
23. Nadkarni, P., Chen, R. & Brandt, C. UMLS concept indexing for production databases: a feasibility study. *J. Am. Med. Inform. Assoc.* **8**, 80–91 (2001).
    Critically assesses the use of UMLS for concept indexing, and provides a useful discussion of nomenclature issues.
24. Hersh, W. R. & Donohoe, L. C. SAPHIRE International: a tool for cross-language information retrieval. *Proc. 1998 AMIA Annu. Symp.* 673–677 (1998).
25. Maynard D. & Ananiadou S. in *Recent Advances in Computational Terminology* (eds Bourigault, D., Jacquemin, C. & L'Homme, M.-C.) (John Benjamins, Amsterdam, 2000).
26. Aronson, A. R. & Rindflesh, T. C. Query expansion using the UMLS Metathesaurus. *Proc. AMIA Annu. Fall Symp. 1997*, 485–489 (1997).
27. Aronson, A. R. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc. AMIA Annu. Fall Symp. 2001*, 17–21 (2001).
28. Brill, E. A simple rule-based part of speech tagger. *Proc. Third ACL Appl. NLP* (1992).
29. Hersh, W. R., Price, S. & Donohoe, L. Assessing thesaurus-based expansion using the UMLS Metathesaurus. *Proc. AMIA Annu. Fall Symp. 2000*, 344–348 (2000).
30. Bodenreider, O. Circular hierarchical relationships in the UMLS: etiology, diagnosis, treatment, complications and prevention. *Proc. AMIA Annu. Fall Symp. 2001*, 57–61 (2001).
31. Hahn, U., Romacker, M. & Schulz, S. Creating knowledge repositories from biomedical reports: the MEDSYNDICATE text mining system. *Pacif. Symp. Biocomp.* 338–349 (2002)
    Applies sophisticated NLP techniques to the task of information extraction, with excellent results.
32. Proux, D., Rechenmann, F., Julliard, L., Pillet, V. & Jacq, B. Detecting gene symbols and names in biological texts: a first step toward pertinent information. *Proc. Genome Inform. Workshop* **9**, 72–80 (1998).
33. Fukuda, K., Tsunoda, T., Tamura, A. & Takagi, T. Toward information extraction: identifying protein names from biological papers. *Proc. Pacif. Symp. Biocomp.* **3**, 707–718 (1998).
34. Yoshida, M., Fukuda, K. & Takagi, T. PNAD-CSS: a workbench for constructing a protein name abbreviation dictionary. *Bioinformatics* **16**, 169–175 (2000).
35. Stapley, B. J. & Benoit, G. Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in MEDLINE abstracts. *Proc. Pacif. Symp. Biocomp.* **5**, 526–537 (2000).
36. Ng, S.-K. & Wong, M. Toward routine automatic pathway discovery from on-line scientific text abstracts. *Genome Inform.* **10**, 104–112 (1999).
37. Wong, L. PIES, a protein interaction extraction system. *Proc. Pacif. Symp. Biocomp.* **6**, 520–531 (2001).
38. Hatzivassiloglou, V., Duboue, P. & Rzhetsky, A. Disambiguating proteins, genes and RNA in text: a machine learning approach. *Bioinformatics* **17** (Suppl. 1), S97–S106 (2001).

39. Thomas, J., Milward, D., Ouzounis, C., Pulman, S. & Carroll, M. Automatic extraction of protein interactions from scientific abstracts. *Proc. Pacif. Symp. Biocomp.* **5**, 541–551 (2000).
40. Humphreys, K., Demetriou, G. & Gaizauskas, R. Two applications of information extraction to biological science journal articles: enzyme interactions and protein structures. *Proc. Pacif. Symp. Biocomp.* **5**, 502–513 (2000).
41. Jenssen, T.-K., Laegreid, A., Komorowski, J. & Hovig, E. A literature network of human genes for high-throughput analysis of gene expression. *Nature Genet.* **28**, 21–28 (2001).
Describes PubGene — a large-scale information extraction system that uses simple co-occurrence to detect associations between genes.
42. Sekimizu, T., Park, H. S. & Tsujii, J. Identifying the interaction between genes and gene products based on frequently seen verbs in medline abstracts. *Genome Inform.* **9**, 62–71 (1998).
43. Ono, T., Hishigaki, H., Tanigami, A. & Toshihisa, T. Automated extraction of information on protein–protein interactions from the biological literature. *Bioinformatics* **17**, 155–161 (2001).
Shows that information extraction can be carried out with reasonable sensitivity and specificity without using overly sophisticated NLP techniques.
44. Blaschke, C., Andrade, M. A., Ouzounis, C. & Valencia, A. Automatic extraction of biological information from scientific text: protein–protein interactions. *Proc. AAAI Conf. Intell. Syst. Mol. Biol.* **7**, 60–67 (1999).
45. Leroy, G. & Chen, H. Filling preposition–base templates to capture information from medical abstracts. *Proc. Pacif. Symp. Biocomp.* 350–361 (2002).
46. Rindflesch, T. C., Tanabe, L., Weinstein, J. N. & Hunter, L. EDGAR: extraction of drugs, genes and relations from the biomedical literature. *Proc. Pacif. Symp. Biocomp.* **5**, 517–528 (2000).
47. Mutalik, P., Deshpande, A. & Nadkarni, P. Use of general-purpose negation detection to augment concept indexing of medical documents: a quantitative study using the UMLS. *J. Am. Med. Inform. Assoc.* **8**, 598–609 (2001).
48. Pustejovsky, J., Castano, J., Zhang, J., Kotecki, M. & Cochran, B. Robust relational parsing over biomedical literature: extracting inhibit relations. *Proc. Pacif. Symp. Biocomp.* 362–373 (2002).
Describes automatically inferred rules for extracting information using grammar induction techniques.
49. Eilbeck, K., Brass, A., Paton, N. & Hodgman, C. INTERACT: an object oriented protein–protein interaction database. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 87–94 (1999).
50. Xenarios, I., Rice, D. W., Salwinski, L., Baron, M. K. & Marcotte, E. M. DIP: the database of interacting proteins. *Nucleic Acids Res.* **28**, 289–291 (2000).
51. Bader, G. D., Donaldson, I., Wolting, C., Ouellette, B. F., Pawson, T. & Hogue, C. W. BIND — the biomolecular interaction network database. *Nucleic Acids Res.* **29**, 242–245 (2001).
52. Blaschke, C. & Valencia, A. Can bibliographic pointers for known biological data be found automatically? Protein interactions as a case study. *Comp. Funct. Genomics* **2**, 196–206 (2001).
Proposes that biomedical text mining is limited more by inadequate lexica and lack of full-text sources than by data-mining technology. Also includes a useful discussion of nomenclature issues.
53. Hearst, M. A. in *WordNet: an Electronic Lexical Database* (ed. Fellbaum, C.) 131–151 (MIT press, Cambridge, Massachusetts, 1999)
54. Roberts, R. J. PubMed Central: the GenBank of the published literature. *Proc. Natl Acad. Sci. USA* **98**, 381–382 (2001).
55. Cruse, D. A. *Lexical Semantics* (Cambridge University Press, Cambridge, UK, 1986)

Acknowledgements

The authors thank P. Li and S. Lewis for many stimulating discussions on the role of ontologies in biology and natural language processing, G. Marth for many useful comments on the manuscript and R. Mural for professional encouragement.

Online links

FURTHER INFORMATION

GenBank:
http://www.ncbi.nlm.nih.gov/Sitemap/index.html#GenBank

Public Library of Science, PLoS:
http://www.publiclibraryofscience.org

SWISS-PROT: http://www.expasy.ch/sprot/sprot-top.html

Access to this interactive links box is free online.