Literature mining for the biologist: from information retrieval to biological discovery

Lars Juhl Jensen, Jasmin Saric‡ and Peer Bork*§*

Abstract | For the average biologist, hands-on literature mining currently means a keyword search in PubMed. However, methods for extracting biomedical facts from the scientific literature have improved considerably, and the associated tools will probably soon be used in many laboratories to automatically annotate and analyse the growing number of system-wide experimental data sets. Owing to the increasing body of text and the open-access policies of many journals, literature mining is also becoming useful for both hypothesis generation and biological discovery. However, the latter will require the integration of literature and high-throughput data, which should encourage close collaborations between biologists and computational linguists.

The focus of biological research is rapidly shifting from individual genes and proteins to entire biological systems. To make sense of the large-scale data sets that are being generated as a result of this change, biologists must increasingly be able to connect with research fields outside their core competence. This requires the ability to systematically compare large data sets with all the knowledge that is dervied from the published data, which allows the biological relevance of the data set to be interpreted. The information, which is measured in terms of the numbers of articles and journals that are published, is increasing at a considerable rate, so that it is no longer possible for a researcher to keep up-to-date with all the relevant literature manually, even on specialized topics (FIG. 1).

Because of these changes, literature-mining tools are becoming essential to researchers. They enable researchers to identify relevant papers — a process that is known as information retrieval (IR). They also allow entity recognition (ER), in which the biological entities that are mentioned in these papers (for example, genes and proteins) are recognized, and enable specific facts to be pulled out from papers in a process that is called information extraction (IE). IR tools such as PubMed have long been used on a regular basis by most biologists to find papers of interest. By contrast, automatic methods for extracting facts from text (such as IE) have only recently become sufficiently accurate to be useful in practice[1]. It is obvious how both IR and IE can be used for curation efforts; however, they are often dismissed as being useless for discovery purposes as they can only extract what has already been published.

More advanced tools that are based on these methods facilitate systematic searches of the scientific literature for overlooked connections. These so-called text-mining methods (not to be confused with the more general term literature mining) can be used to make novel hypotheses by combining information from multiple papers. However, we believe that the full discovery potential of such tools will only be realized with the advent of data-mining approaches that integrate the literature with other large data sets such as genome sequences, microarray expression studies, or protein–protein interaction screens (FIG. 2).

Here we briefly describe the aim of each field outlined above, give an overview of the methods that are used (BOX 1) and discuss what can currently be achieved. Although our main focus is on text mining and data integration, we first briefly review the most important IR, ER and IE methodologies that are used for text and data mining (see REFS 2–5 for more details on these topics). Throughout, we use the following example sentence: "Mitotic cyclin (Clb2)-bound Cdc28 (Cdk1 homolog) directly phosphorylated Swe1 and this modification served as a priming step to promote subsequent Cdc5-dependent Swe1 hyperphosphorylation and degradation"[6]. Its context is the cell cycle of the yeast *Saccharomyces cerevisiae* and it allows us to demonstrate the powers and pitfalls of current literature-mining approaches.

**European Molecular Biology Laboratory, D-69117 Heidelberg, Germany.*
‡EML Research gGmbH, D-69118 Heidelberg, Germany.
§Max Delbrück Centre for Molecular Medicine, D-13092 Berlin, Germany.
Correspondence to L.J.J. and P.B.
e-mails: jensen@embl.de; bork@embl.de
doi:10.1038/nrg1768

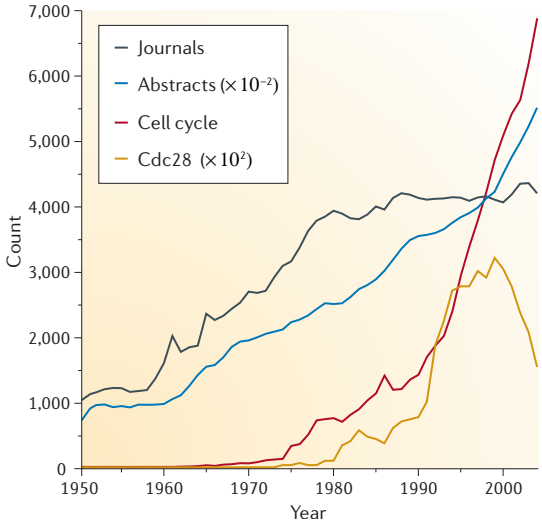


Figure 1 | **Growth of Medline.** The numbers of journals, papers (as represented by Medline abstracts), papers on the cell cycle and papers on Cdc28 that were published each year from 1950 to 2005 are shown. An average for 3 years was calculated for the Cdc28 curve because of much lower numbers. The number of new papers that were published each year continues to increase, especially on certain topics such as the cell cycle, for which it is no longer possible to read all new papers that are published. By contrast, specific proteins that are 'hot' at one point in time tend to lose their popularity later, as exemplified by Cdc28.

Information retrieval: finding the papers

IR systems aim to identify the text segments (be it full articles, abstracts, paragraphs or sentences) that pertain to a certain topic — the example here is the yeast cell cycle. The topic might be either a user-provided query (*ad hoc* IR) or a pre-defined set of papers (text categorization). Ideally, an IR system should recognize our example sentence as being related to the yeast cell cycle, although neither 'yeast' nor 'cell cycle' is explicitly mentioned.

The best-known biomedical IR system, PubMed, is an *ad hoc* system that uses two established IR methodologies — the Boolean model and the vector model. The Boolean model allows the user to retrieve all documents that contain certain combinations of terms by using a logical operation; for example, 'yeast AND cell cycle'. By contrast, the vector model represents each document by a term vector, in which each term is assigned a value according to a frequency-based weighting scheme. These document vectors can subsequently be compared to a query vector that specifies the relative importance of each query term[7]. Alternatively, they can be compared to each other to calculate document similarity, which is used in PubMed by the 'related articles' function[8] and other document-clustering methods[9–11].

The vector representation is also used as input for machine-learning methods, which are trained to discriminate between known relevant (positive) and irrelevant (negative) papers on the basis of their word content[12–18]. Such methods are able to learn complex rules; for example, a method trained to identify sentences that are related to the yeast cell cycle would have learned that the word 'Cdc28' in our example sentence is a strong hint, whereas the words 'Cdk1' and 'Clb2' could also be related to the cell cycles of other organisms.

Ad hoc IR systems such as PubMed generally have more difficulty than text-categorization systems in dealing with the many abbreviations, synonyms and ambiguities in biomedical terminology. However, blind assessments (BOX 2) have shown that most of the lessons that have been learned from IR in other research fields carry over to the biomedical sciences[11,19,20]. These include removing so-called stop words such as 'the' and 'it', which occur in almost every document, and truncating common word endings such as '-ing' and '-s' to allow different forms of the same word to be matched — for example, 'yeast' and 'yeasts'[13]. PubMed and many other good biomedical IR systems also make use of thesauri to automatically expand the query with other related terms[13,19–21]. For example, the Boolean query 'yeast AND cell cycle' might be expanded to '(yeast OR *Saccharomyces cerevisiae*) AND cell cycle'.

Many advanced IR methods, such as MedMiner[22] and Textpresso[23], also use ER methods to better identify documents that mention a certain gene or protein, and an approach known as part-of-speech tagging can be used to determine whether a word such as 'wingless' occurs as a noun or an adjective. Another advance that is expected in the future is tackling the way in which IR results are presented. As many documents might be retrieved by a single query, simply showing them as a long list gives a poor overview. Alternative ways to present and summarize IR results are therefore being explored[24–27].

Even with these improvements, current *ad hoc* IR systems are not able to retrieve our example sentence when they are given the query 'yeast cell cycle'. Instead, this could be achieved by realizing that 'yeast' is a synonym for *S. cerevisiae*, that 'cell cycle' is a Gene Ontology term, that the word 'Cdc28' refers to an *S. cerevisiae* protein and finally, by looking up the Gene Ontology terms that relate to Cdc28 to connect it to the yeast cell cycle. Although this will not be easy, we see this form of query expansion as the next logical step for *ad hoc* IR.

Entity recognition: identifying the substance(s)

The seemingly modest goal of ER is to find the biological entities that are mentioned within a text; in particular, the names of genes and proteins. This task is often divided into two sub-tasks: first, the recognition of words that refer to entities and second, the unique identification of the entities in question. In our example sentence, the terms 'Clb2', 'Cdc28', 'Cdk1', 'Swe1' and 'Cdc5' should therefore all be recognized as gene or protein names and uniquely identified by, for example, their respective *Saccharomyces* Genome Database accession numbers.

Although ER might at first seem neither challenging nor particularly useful, it is possibly the most difficult task in biomedical text mining and is a prerequisite for both IE and advanced IR. The early ER methods relied on manually devised rules that look for typical features of names — such as letters that are followed by numbers, or the ending '-ase' — as well as contextual information

Machine learning
The ability of a machine to learn from experience or extract knowledge from examples in a database. Artificial neural networks and support-vector machines are two commonly used types of machine-learning method.

Gene Ontology
A set of controlled vocabularies that are used to describe the molecular functions of a gene product, the biological processes in which it participates and the cellular components in which it can be found.

from nearby words such as 'gene' and 'receptor'[13,28,29]. As literature collections (corpora) in which gene and protein names have been tagged are now available (BOX 1), most newer systems rely instead on machine-learning algorithms to recognize names on the basis of their characteristic features[29–34].

In contrast to these systems, several dictionary-based methods rely instead on a comprehensive list of synonymous gene names that are matched against the documents using algorithms that allow variation in how the names are written — for example, 'CDC28', 'Cdc28', 'Cdc28p' or 'cdc-28'[13,31,35–41]. Dictionary-based approaches have one crucial advantage over feature-based ones — they not only recognize names but also identify the accession numbers of the genes or proteins to which they refer. Many systems combine dictionary matching with either rule-based or statistical methods to reduce the number of false positives[31,36–38]. The best-performing ER methods in blind assessments rely on careful curation of gene-name lists to remove aliases that cause many false positives[40,41].

The main difficulty in ER arises from the lack of standardization of names. Each gene or protein typically has several names and abbreviations (for example, 'Cdc28' is also known as 'cyclin-dependent kinase 1' or just 'Cdk1'), some of which are also common English words (for example, 'hairy'), biological terms (for example, 'SDS') or names of other genes (for example, 'Cdc2' refers to two completely unrelated genes in budding and fission yeast)[42]. The recent development of methods for resolving ambiguity in gene or protein names is therefore an important advance for ER[41,43,44].

Instead of focusing on this important problem, many methods have attempted to recognize whether a particular occurrence of a name refers to a gene or its protein product[28,30,37]. However, this distinction is not always clear. For example, 'Cdc5-dependent Swe1 hyperphosphorylation' depends on both the Cdc5 protein and also the gene that encodes it. Human annotators only agree with each other in 77% of cases when asked to distinguish between genes, RNAs and proteins that are mentioned in the literature[45]. Fortunately, the ability to discriminate between genes and proteins is of little consequence for downstream IE applications.

Although ER is generally intended as a building block for IR and IE systems, it can also be useful on its own for crosslinking the literature that is related to certain genes. A good example of this is iHOP (Information Hyperlinked over Proteins)[25], which is a web-based tool that allows the user to browse sentences from Medline abstracts on the basis of the biomedical entities that they mention.

Information extraction: formalizing the facts

In contrast to IR systems that identify texts concerning particular topics, IE systems aim to extract pre-defined types of fact — in particular, relationships between biological entities. From our example sequence, an IE system should deduce that Cdc28 binds Clb2, that Swe1 is phosphorylated by the Cdc28–Clb2 complex and that Cdc5 is involved in Swe1 phosphorylation. These facts can subsequently be stored in a database, with the option of being verified by a curator reading the paper in question. Two fundamentally different approaches to extracting relationships from biological texts are currently being used extensively, namely co-occurrence and natural-language processing (NLP).

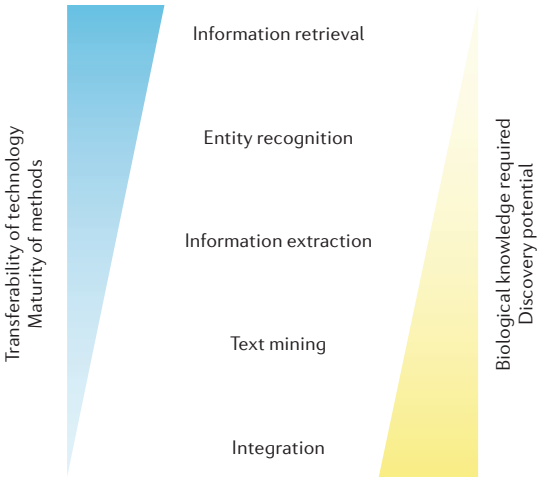


Figure 2 | **The current state of biomedical literature mining.** Biomedical literature mining can be divided into several disciplines that have independent goals but use related tools. These disciplines range from information retrieval to the integration of text with other data sources. Whereas information retrieval, entity recognition and information extraction are established tasks in computational linguistics for which methods from other fields have been transferred to biomedicine, text mining and data integration are still in their infancy and few methods have been put forward. This is because of the extensive biological insight that is required to develop them. However, such tools are also the most rewarding for biologists to help develop as they have the greatest potential for leading to new biological discoveries.

Co-occurrence. The simplest approach to IE is to identify entities that co-occur within abstracts or sentences. As two entities might be mentioned together without being in any way related, most systems use a frequency-based scoring scheme to rank the extracted relationships[14,25,46–58]. If two entities are repeatedly mentioned together, it is likely that they are somehow related, although the type of relationship is not known[49,52]. Co-occurrence methods tend to give better recall but worse precision than NLP methods[56,59], and are well suited as parts of exploratory tools because of their ability to identify relationships of almost any type[53,54].

Co-occurrence methods can also be used to extract relationships of a certain type only, such as physical protein–protein interactions, by combining them with a customized text-categorization system to identify the relevant abstracts or sentences[14,46–50,60]. This set-up is particularly attractive for database curation as the custom-made text-categorization system can also be used on its own and high coverage can be attained[14,48]. However, complex sentences that contain multiple

relationships give rise to additional, erroneous relationships — our example sentence might link Cdc5 to Clb2. This approach is also unable to extract directional relationships (for example, whether Cdc5 is involved in Swe1 phosphorylation or *vice versa*) and has difficulty distinguishing between direct and indirect relationships (for example, whether or not Swe1 is directly phosphorylated by Cdc5).

Box 1 | **Online tools and resources**

There are numerous literature collections (corpora), software modules and web-based applications for biomedical literature mining. Here we list some applications that can be accessed through web interfaces and a small subset of resources that are useful for developers of new literature-mining systems.

Web-based applications

Information retrieval

E-BioSci ..... http://www.e-biosci.org
EBIMed ..... http://www.ebi.ac.uk/Rebholz-srv/ebimed
Google Scholar ..... http://scholar.google.com
GoPubMed ..... http://www.gopubmed.org
MedMiner ..... http://discover.nci.nih.gov/textmining
PubFinder ..... http://www.glycosciences.de/tools/PubFinder
PubMed ..... http://www.pubmed.org
Textpresso ..... http://www.textpresso.org
XplorMed ..... http://www.ogic.ca/projects/xplormed

Entity recognition

iHOP ..... http://www.pdg.cnb.uam.es/UniPub/iHOP

Information extraction

iProLINK ..... http://pir.georgetown.edu/iprolink
JournalMine ..... http://textmine.cu-genome.org
PreBIND ..... http://prebind.bind.ca
PubGene ..... http://www.pubgene.org

Text mining

Arrowsmith ..... http://arrowsmith.psych.uic.edu

Integration

BITOLA ..... http://www.mf.uni-lj.si/bitola
G2D ..... http://www.ogic.ca/projects/g2d_2
ProLinks ..... http://dip.doe-mbi.ucla.edu/pronav
STRING ..... http://string.embl.de

Text collections

Full text corpora

HighWire Press ..... http://highwire.stanford.edu
PubMed Central ..... http://www.pubmedcentral.org

Tagged corpora

FetchProt ..... http://fetchprot.sics.se
GENETAG ..... ftp://ftp.ncbi.nlm.nih.gov/pub/tanabe
GENIA ..... http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA
PennBioIE ..... http://bioie.ldc.upenn.edu
Yapex ..... http://www.sics.se/humle/projects/prothalt

Information-extraction modules

Entity taggers

ABNER ..... http://www.cs.wisc.edu/~bsettles/abner
GAPSCORE ..... http://bionlp.stanford.edu/gapscore

Part-of-speech taggers

Brill Tagger ..... http://www.cs.jhu.edu/~brill
TNT Tagger ..... http://www.coli.uni-saarland.de/~thorsten/tnt
TreeTagger ..... http://www.ims.uni-stuttgart.de/~schmid

Parsers

CASS ..... http://www.vinartus.net/spa
Collins Parser ..... http://people.csail.mit.edu/mcollins
Stanford Parser ..... http://nlp.stanford.edu/software

Natural-language processing. These issues can all be addressed by NLP methods that combine the analysis of syntax and semantics. The text is first 'tokenized' to identify sentence and word boundaries, and a part-of-speech tag (for example, a noun or verb) is assigned to each word. A syntax tree is then derived for each sentence to delineate noun phrases (for example, 'Mitotic cyclin (Clb2)-bound Cdc28 (Cdk1 homologue)') and represent their interrelationships. ER methods and simple dictionaries are subsequently used to semantically tag the relevant biological entities (for example, genes and proteins) and other keywords (for example, activation, repression or phosphorylation). Finally, a rule set is used to extract relationships on the basis of the syntax tree and the semantic labels. Few NLP systems attempt to resolve anaphoric relationships, so most systems are therefore unable to extract relationships that span multiple sentences[61]. This is not as big a limitation as it might seem because most relationships are mentioned within a single sentence[47,59].

Several programs exist for tokenization and part-of-speech tagging of English texts (BOX 1), most of which are easily adapted to biomedical texts by retraining them on a manually tagged corpus such as GENIA or PennBioIE[38,62]. Semantic tagging is more complicated, but it can be greatly simplified using existing ER methods. By contrast, the development of grammar and extraction rules that can correctly parse sentences and extract facts remains challenging.

The idealized work flow described above indicates that syntactic parsing of sentences, and their semantic interpretation are carried out as two separate steps[63–66]. However, most generic English parsers perform poorly if applied directly to biomedical texts because of the technical terminology that they contain and, particularly, the use of long, complex noun phrases. Better results can be obtained by first tagging the noun phrases[65]. However, many biomedical NLP systems have combined the syntactic parser and the semantic extraction rules into a customized partial parser that specifically targets only the relevant parts of sentences and directly extracts the facts[62,67–69]. The main drawback of this approach is that a large number of extraction rules are needed to cover the many slightly different ways of expressing a certain relationship. These rules can either be developed manually[62,67–69] or learned automatically from a corpus[46,70]. Both methods are labour-intensive, as the latter requires the prior manual tagging of a large training corpus.

Applications of information extraction. Most studies using IE have focused on extracting few types of relationship. These include physical protein–protein interactions[14,47–49,65–71] and interactions that involve unspecified molecular mechanisms between proteins[25,49–55,64–68]. Relationships have also been extracted for concepts such as disease names, Gene Ontology terms and nouns in general[46,56–58,60,63]. Recently, NLP methods have been developed for extracting

Box 2 | **The jungle of quality estimates**

As text mining has been pursued mainly by computational linguists, and biologists are just beginning to explore the methods, the classic scenario arises of two research communities that need to communicate in order to learn from each other. This starts with developing a common language and common evaluation standards.

To evaluate a literature-mining method, its output is either compared to a gold standard or is manually inspected by an expert. This yields three important values: correct retrievals/extractions (true positives, TP), type I errors (false positives, FP), and type II errors (false negatives, FN). From these, several other measures can be derived. Within literature mining, the most common are:

- Recall: The fraction of relevant documents that were retrieved ($TP/(TP + FN)$); also known as sensitivity.
- Precision: The fraction of retrieved documents that were relevant ($TP/(TP + FP)$); also known as specificity.
- F-score: The most commonly used measure for ranking information-retrieval, entity-recognition and information-extraction methods. It is defined as the harmonic mean of the recall and the precision ($2 \times recall \times precision/(recall + precision)$). Because the relative importance of recall and precision varies between tasks, the method with the best F-score is not necessarily the best for a given task.

One of the biggest problems with these quality estimates has been largely ignored in the literature-mining community. That is, precision and F-score are not inherent properties of the methods but also depend on the frequency of positive examples in the evaluation corpus[109]. Imagine that an information-extraction method has 90% precision and recall when applied to an evaluation corpus in which 50% of the sentences contain relations that are to be extracted. If this method were applied to another corpus in which only 1% of the sentences contain relevant relations, the precision would drop to 15%.

Evaluation on a non-representative subset of Medline can also affect recall. If a method for extracting protein interactions is tested only on sentences that contain the verb 'to bind', the recall might be overestimated. For example, methods would not be penalized for failing to extract nominalized phrases (phrases without a verb) such as 'Mitotic cyclin (Clb2)-bound Cdc28'.

Even if these pitfalls are avoided, it is hard to compare the merits of different approaches, as they have not been benchmarked against a common reference. This issue has been addressed through several blind assessments and their associated conferences, namely the Text Retrieval Conference (TREC)[19,20], the KDD-Cup[110], and BioCreAtIvE[111]. These efforts have been important in accelerating the development of information-retrieval and entity-recognition methods but have been less successful for information extraction.

information on gene regulation[62], protein phosphorylation[61,62,67,68] and tissue specificity of alternative transcripts[17]. Probably because of the inherent complexity of the task, only a few systems have been designed that are able to extract multiple types of relationship[62,66–68].

Using an NLP-based system that is able to provide information on multiple types of interaction, all the relationships that are mentioned in our example sentence can be correctly extracted[62]. To demonstrate how IE can be used at a larger scale, we have applied this method to all Medline abstracts, extracting more than 5,000 binary relationships (which might each be mentioned multiple times). Of these, 370 interactions are between yeast proteins, and are shown in FIG. 3a as a network together with the interactions that were identified by co-occurrence[54] (see also supplementary information S1 (figure)). This identifies almost 3,000 interactions between these proteins, but only 150 of these are of comparable reliability to those that were obtained by NLP. With the growing interest in systems biology, IE will probably become a mainstream tool for biologists in the near future, as it is one of the only ways to identify diverse types of relationship on a large scale.

Text mining: finding nuggets in the literature

Often used as a catch-all term for computational text analysis, text mining is more strictly defined as "the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources" (M. Hearst, personal communication; see also REF. 72). IE methods do not therefore qualify as text-mining tools themselves, as they can only extract what has already been published. Rather, they form the basis for text mining in the same way that ER forms the basis for IE[72].

Inferring indirect relationships. It might at first seem impossible for a computer to make discoveries on the basis of literature alone; after all, IE is only able to extract the facts that have already been published. The trick is to use facts that have been extracted from several different publications (A leads to B and B leads to C) to infer new, indirect relationships (A leads to C). As the literature is so vast that each researcher can only read a small subset, it might be that no person is aware of all the facts that are required to make this logical inference. This is plausible especially if the facts were published within two disconnected research areas[72,73] or if an overwhelming number of papers are published on a single topic[74].

For almost two decades, Don Swanson has argued along these lines and he used a simple semi-automated method — Arrowsmith — to infer the following new relationships: fish oil can help patients suffering from Raynaud disease[73]; magnesium deficiency has a role in migraine headache[75]; arginine intake has an effect on levels of somatomedin C in the blood[76]; and oestrogen protects against Alzheimer disease[77]. The first two have subsequently been confirmed experimentally[78,79]. However, these early predictions were all made using a 'closed' framework in which the user provides the hypothesis (A is related to C), which is then tested by a computational search for shared, related words (B) that could support the hypothesis. It can therefore be argued that the computer did not actually make the discovery.

Syntax
The orderly manner in which words are put together to form phrases and sentences.

Semantics
The meaning that is implied by words and sentences. If an information-extraction method extracts the right facts from a sentence, it has interpreted the semantics correctly.

Anaphoric relationships
Back-references to previously mentioned entities. A protein that is mentioned in an earlier sentence might, for example, be subsequently be referred to as 'it'.

Corpus
A collection of texts. A corpus might consist of either the raw text only (for example, Medline) or be tagged so that, for example, gene and protein names are labelled (for example, GENIA).

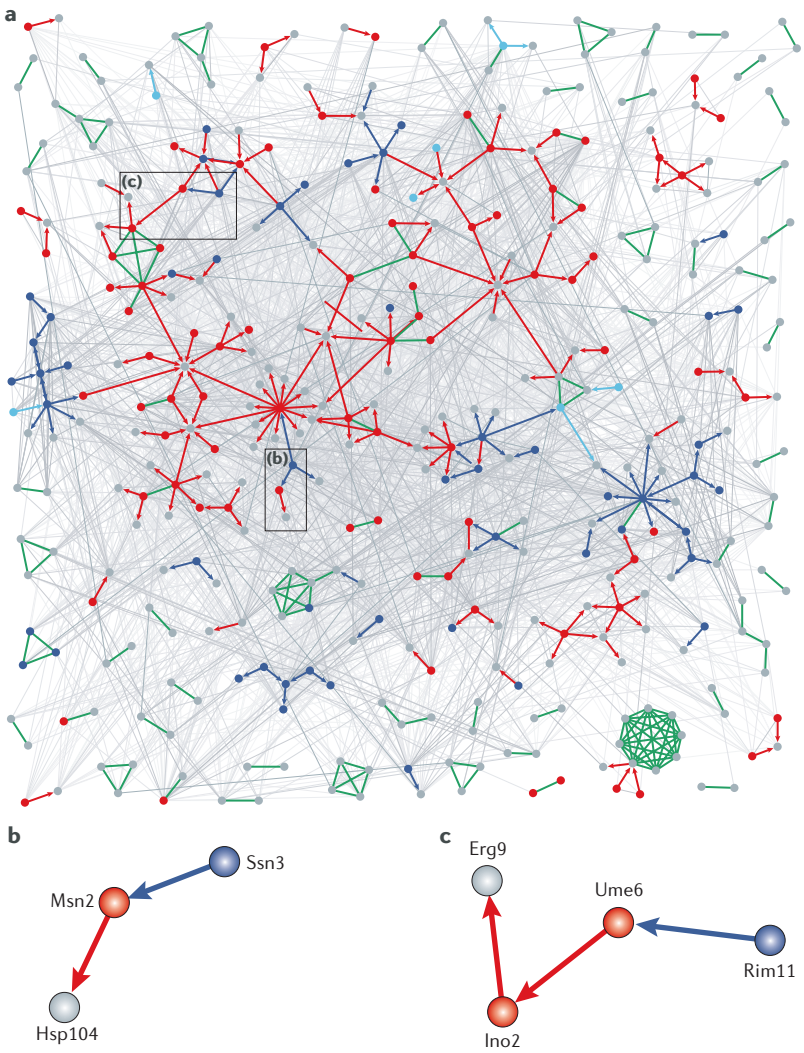


Figure 3 | **A literature-derived network for yeast. a** | A yeast protein network was derived that applied information-extraction approaches to all abstracts that are stored in Medline, using both a statistical co-occurence method[54] and a natural-language-processing (NLP)-based one[62]. Functional associations that were derived from co-occurrence are shown in shades of grey according to the level of confidence that was achieved. The NLP method extracts four types of relationship: stable physical interactions (green), regulation of expression (red), phosphorylation (dark blue) and dephosphorylation (light blue). The proteins (circles) are coloured according to their functional annotation: (co-)regulators of expression (red), kinases and cyclins (dark blue), phosphatases (light blue) and other proteins (grey). A version of this figure that includes all protein names is available in the supplementary information S1 (figure). **b,c** | Examples of unpublished relationships that can be inferred from the network. From the network we can infer that Ssn3 probably influences Hsp104 expression through phosphorylation of Msn2 (**b**). In addition, Ume6 probably regulates Erg9 expression and Rim11 is predicted to regulate the expression of both Ino2 and Erg9 (**c**). None of these hypotheses has been tested experimentally.

The corresponding 'open' discovery problem is more challenging, but also potentially more rewarding, as it starts from only a single entity (A; for example, a disease) and attempts to find indirect, undiscovered relationships to other entities (C; for example, chemicals or genes). Several methods exist that rely on the same strategy: first, identify the terms B that co-occur with A, and second, identify the terms C that co-occur with B but not with A[80–83]. The main problem with this approach is that inferences are made from undirected relationships of unknown types, so that causality cannot be taken for granted. For example, Cdc28 co-occurs with many of its substrates in Medline abstracts, which would cause most existing methods to propose novel but incorrect relationships between unrelated Cdc28 substrates.

To our knowledge, no published studies have made use of NLP-based IE for text mining, although this could ensure that the novel relationships are inferred from causal chains of relationships. A probable reason is that few NLP systems are able to accurately extract a sufficiently large number of directed relationships to allow this approach.

However, the following example demonstrates the feasibility of using NLP-based text mining to discover novel relationships. We used the yeast network of phosphorylation and gene expression that we derived using IE (FIG. 3a) to indirectly link 64 pairs of proteins, in which the two members of each pair do not co-occur in Medline abstracts. Manual inspection of the literature indicates that more than 90% of the inferred relationships are correct. For example, the network indicates that the cyclin-dependent kinase Ssn3 (also known as Srb10) influences expression of the stress-response protein Hsp104 through phosphorylation of Msn2 (FIG. 3b). It is known that Hsp104 expression is activated by Msn2 (REF. 84) and that Msn2 is phosphorylated by Ssn3 (REF. 85). In addition, Ssn3 was recently shown to be a repressor of the general stress response, although whether and how this is mediated by Msn2 phosphorylation remains controversial[86,87]. Therefore, it is plausible that Ssn3 regulates Hsp104 expression but this has not been experimentally verified.

In a second example, it is known that Rim11 phosphorylates Ume6 (REF. 88), which regulates the expression of another transcription factor, Ino2 (REF. 89), and that Ino2 in turn regulates Erg9 expression[90]. It can therefore be inferred that Ume6 is likely to regulate Erg9 expression, and that Rim11 regulates the expression of both of the other two proteins. Remarkably, however, neither of these relationships seem to have been described in the published literature, although they can be inferred using our NLP-based method (FIG. 3c).

Although most are correct, the vast majority of the inferred relationships in our study of yeast interactions also turn out to be well known, despite the proteins never having been mentioned together in any abstract. Without full-text access to all published papers, it is unfortunately impossible to rule out that an inferred relationship has already been published. In addition, some relationships are probably considered to be so trivial that no one has ever published them. To avoid overwhelming the user with trivial hypotheses, text-mining methods need to integrate data sources other than the scientific literature itself — in particular, databases of curated information.

Searching for global trends. An alternative text-mining strategy is to search for global trends within the literature.

Box 3 | **Buzzword hunting**

Medline is a rich resource for mining various facts of social or economical nature[112–115]. Here we present another application that highlights a predictive aspect of text mining that has not yet been exploited: the prediction of research fields that are about to become popular. Beyond its usefulness for science managers and/or grant writers, it demonstrates the power of text mining in discovering global trends.

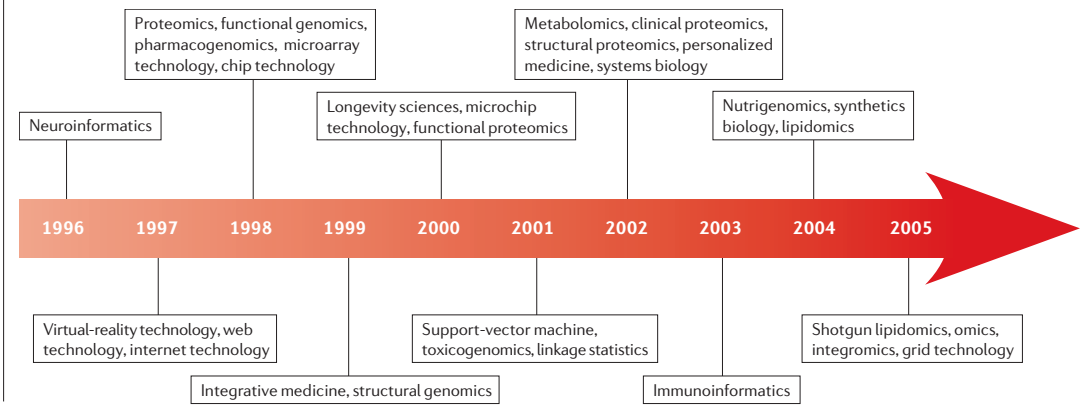
As most buzzwords are names of research areas or technologies, we limited the analysis to terms with endings such as '-ics' or '-ology'. The characteristic behaviour of a buzzword is that it suddenly starts being mentioned frequently, after being mentioned at most a couple of times in earlier years. To formalize this, we define a heuristic buzzword index (BWI) for a term:

$$BWI = \mathrm{Ln}(n) \times \frac{(n + 1)/(N + 1)}{(n^* + 1)/(N^* + 1)}$$

where n is the number of times a term was mentioned in the past year and n^* is the number of times the word was mentioned in the previous 10 years. N and N^* are similarly calculated on the basis of the number of occurrences of the word 'biology'; this normalizes the score with respect to the general growth in biology-related papers.

By running this method on all Medline abstracts that were published up to a certain year and inspecting the terms with BWI > 25 and $5 \leq n < 50$, we tested which buzzwords we would have suggested if running the methods at the end of each year and obtained the results shown in the figure.

Most buzzwords of the past that we are aware of would have been detected at the time that they became popular. The predictive power of the method could be improved by including other text sources or taking into account the fact that many new buzzwords are derived from former ones; for example, proteomics later gave rise to functional proteomics, structural proteomics and clinical proteomics.



Temporal trends can be revealed by simply counting how many times a given term was mentioned each year, and this has been used to analyse changes in which genes are 'hot'[91] (FIG. 1), and can also find emerging buzzwords (BOX 3). Such methods could also be used, for example, to predict future 'hot' proteins, which are commercially attractive targets for the development of antibodies and inhibitors.

An established data-mining method that has not previously been used in text mining is a search for correlated events, as exemplified by Amazon's "Customers who bought this item also bought…" function. In biology, this could be used to discover fundamental properties of regulatory networks. As an example, to test the feasibility of this approach, we compared the lists of yeast proteins shown in FIG. 3 that are described in the literature as being regulated through expression and phosphorylation. The overlap between the two sets is more than fourfold larger than expected by chance ($P < 5 \times 10^{-4}$), indicating that phosphorylation and regulation of expression target the same proteins, as was recently proposed by de Lichtenberg *et al.* from the integration of several large-scale experimental data sets[92].

Similarly, analysis of the relationships in FIG. 3 also reveals that protein kinases preferentially phosphorylate each other ($P < 9 \times 10^{-9}$) and that transcription factors tend to regulate the expression of other transcription factors ($P < 2 \times 10^{-7}$), reflecting the existence of signalling cascades and transcriptional networks, respectively. The individual pieces of information that are required for making other such discoveries are likely to be present in the literature and could be combined using a similar systematic, computational method. A drawback of this methodology is that statistically significant correlations can arise because of study bias; for example, many cell-cycle proteins will have been examined for both phosphorylation and changes in expression. However, this can be overcome by combining IE results with genome-wide experimental data sets, as discussed below, because such data sets are unbiased with respect to how well the proteins have been studied. For example, the correlation between phosphorylation and regulation of expression is confirmed by comparing the IE results

Study bias
Study biases arise because some proteins (or other molecules) are more studied than others. For example, if a protein is known to be phosphorylated, it is also more likely to have been studied in other respects, and is therefore more likely to be known to be regulated by expression, for example.

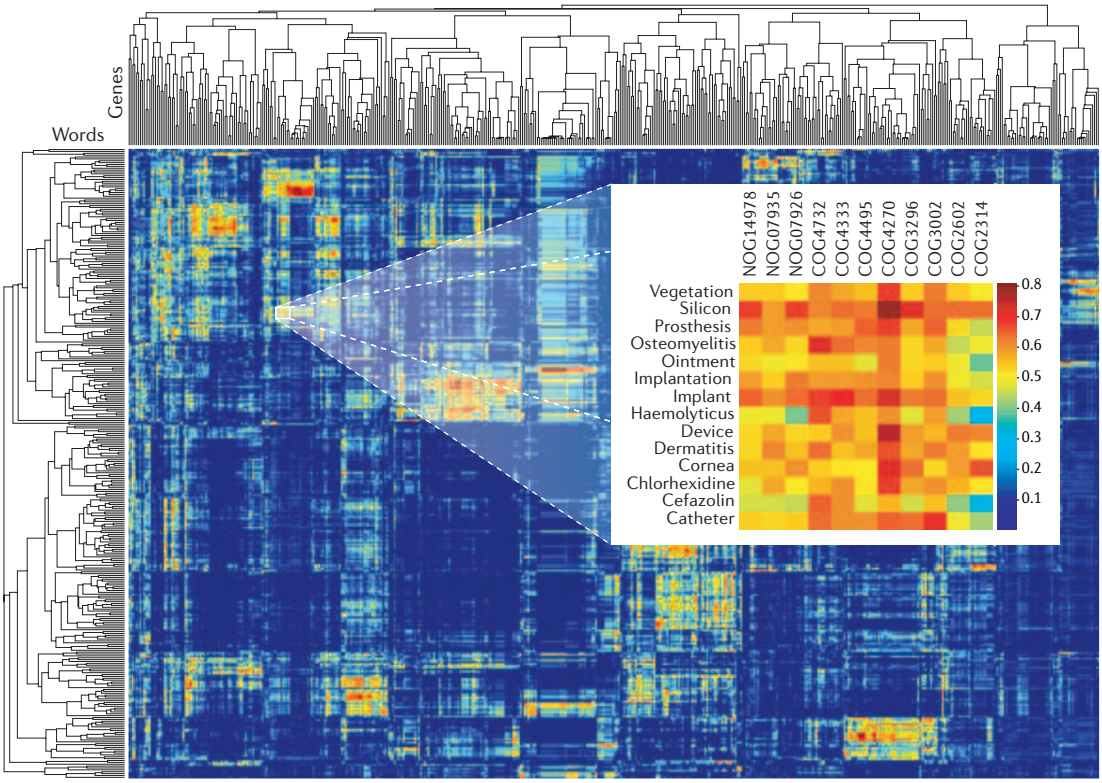


Figure 4 | **Correlating phenotypes with genotypes.** The approach that was used involved the integration of gene occurrence in genomes with keywords that are overrepresented in the literature associated with certain species[105]. Species distributions of keywords that were derived from Medline were compared with the species distributions of genes to calculate how strongly they are associated. The resulting association scores are shown as a heat map. The two trees show the individual clustering of species profiles for genes and keywords. The insert shows a cluster that contains 11 groups of orthologous genes with unknown function (referred to by their clusters-of-orthologous-groups-of-proteins (COG) or non-supervised-orthologous-group (NOG) accession numbers) that are only present in *Staphylococci* species and certain other hospital bacteria. All these genes are strongly associated with words that occur more frequently in abstracts that relate to those species, such as osteomyelitis (a disease that is related to *Staphylococci*), cornea (a part of the eye that can be infected by *Staphylococci*), cefazolin (an antibiotic that is often used against *Staphylococci*) and chlorhexidine (a disinfectant against which *Staphylococci* are resistant). As both genes and words seem to be associated with this species subset, the genes are probably directly or indirectly associated with the corresponding phenotypes. The genes might be directly involved in disease phenotypes or might only be indirectly involved by contributing to the lifestyle. In either case, the specificity of these genes to a limited set of infectious bacteria makes them candidates for drug targets. Figure modified from REF. 105.

on phosphorylation with a set of cell-cycle-regulated genes that has been derived from microarray expression studies[92] ($P < 4 \times 10^{-4}$).

Integration: combining text and biological data

Although text mining can be used to uncover overlooked relationships, data-mining approaches that integrate literature with other data types have greater potential for making biological discoveries. As an example of how this could be achieved, relationships that apply to a particular protein of interest could be extracted from the literature, followed by sequence-similarity searches to transfer these relationships to orthologous proteins[4]. In this way, text-mining methods could be used to make inferences that are based on relationships from multiple species, and therefore connect communities of researchers who work on different model organisms.

To test this approach, we combined the *Drosophila melanogaster* and mouse equivalents of the yeast network shown in FIG. 3 (REF. 62) using sequence-based orthology assignments from the STRING database[54]. Using this method, we uncovered the following indirect relationship: in *D. melanogaster*, Suppressor of Hairless (Su(H)) has been shown to be a direct transcriptional repressor of *single-minded*[93]. As the mouse Single-minded 1 protein is a transcriptional activator of erythropoietin (*Epo*)[94], we make the hypothesis that one or more of the mural Su(H) orthologues downregulate *Epo* expression, although none of them co-occurs with *Epo* in Medline abstracts. The power of such approaches will improve only with both the growth of the literature and an increase in the availability of large-scale data sets.

However, most attempts so far to integrate the literature with biological data have been directed towards the annotation of data that has been obtained from functional-genomics studies, as manual in-depth analysis is not feasible in these studies because of the amount of data that is generated[52,95–101]. Most approaches first use

ER methods or database cross-referencing to retrieve the Medline abstracts that are associated with one or more genes — for example, a protein family or a cluster of genes that are co-expressed in a microarray experiment. Then, these abstracts can be used to identify significant overrepresentation of keywords within the text[95,96] or of annotated MeSH terms (medical subject heading terms), both of which can contribute to characterizing the genes in question[97,98]. Alternatively, the abstracts can be used to evaluate the cluster coherence (a measure of functional similarity for a group of genes)[99–101] or construct a functional association network of the genes from their co-occurrence in abstracts[52,55].

Through their ability to bring together many types of data, networks have the potential to form the basis for text and data integration. There are several web-based tools that provide access to protein networks that are based on both IE and high-throughput experiments[25,53,54]. These have proved valuable as exploratory tools that allow researchers to browse many types of information for a set of proteins of interest. In addition, they are useful for providing a structured overview of other types of high-throughput data. For example, expression data can be mapped onto protein-interaction networks to visualize how the synthesis of protein complexes is regulated at the transcript level[92]. Such networks can also be combined with other types of data to provide insights into the molecular basis of a disease. For example, literature-based protein networks have been integrated with linkage-mapping studies to identify candidate genes for Alzheimer disease within a genomic region on the basis of their interactions with genes that are already known to have a causal role in the disease[102].

The types of network that are described above only include relations at the molecular level; however, the possibility of making discoveries is improved by integrating relationships at multiple levels. This is exemplified by several literature-mining tools that are used to prioritize candidate genes with potential roles in inherited diseases for further study. The first such system, G2D, was published in 2002 (REF. 103). It combines the MeSH annotation in Medline with the Gene Ontology annotation of entries in the NCBI RefSeq database to infer logical chains of connections from disease names, through chemicals and drugs, to molecular functions. Combined with functional annotations that are inferred from sequence similarity, this allows the genes within a mapped region to be ranked on the basis of a score that represents their likelihood of being associated with the disease in question. A second system, BITOLA, relies instead on pure text mining to find candidate genes that are indirectly connected to a given disease, and subsequently filters these on the basis of chromosomal-mapping data about the disease[83]. A third approach identifies co-occurring disease and tissue names in Medline and combines these with tissue-expression annotations from Ensembl to link the tissues to candidate disease genes[58]. Although the original G2D method was limited to Mendelian diseases, these approaches have recently been shown to work for complex genetic diseases[58,104].

Even broader in scope is a recent study that correlates text mining for phenotypic information with gene occurrences across species (genotype information) to infer phenotypic roles for genes of unknown function[105]. Medline was systematically searched for keywords that were associated with each prokaryote for which the genome has been sequenced. The resulting species distributions of keywords were then matched against the species distributions of genes to associate keywords with genes (FIG. 4). The set of keywords that is associated with a group of genes can reveal the phenotypic characteristics that are caused by these genes. For example, genes that are unique to *Staphylococci* species and other hospital bacteria cluster together with descriptive keywords such as 'osteomyelitis' (a disease that is related to *Staphylococci*) and less obvious ones such as 'chlorhexidine' (a disinfectant against which *Staphylococci* are resistant) (FIG. 4). This indicates putative roles for these genes of unknown function and highlights them as possible drug targets. When it was applied globally, the approach recaptured many known genotype–phenotype relationships and also predicted several new ones, such as genes that encode enzymes that are involved in degradation of plant tissues, and genomic determinants for food poisoning[105].

Outlook

The peer-reviewed scientific literature will continue to be a prime resource for accessing worldwide scientific knowledge, and the continuing growth and diversification of that literature will require tremendous systematic and automated efforts to utilize the information that it contains. In the near future, tools for mining this knowledge base will probably have a pivotal role in systems biology. So far, more than 90% of all biomedical literature mining has been based on Medline, mainly because it is freely availably in a convenient format. To realize the full potential of these approaches, future methods should be able to extract information from the full text of papers, including citation information, which could then be cross-referenced between papers. This will require some methodological improvements as not all sections of a paper are equally relevant[106,107] and because some information must be extracted from figures and tables, which current methods are not designed to deal with. However, it is the restricted access to the full text of papers and to citation information, rather than the technology, that is currently the greatest limitation, despite some encouraging open-access initiatives such as PubMed Central and Highwire Press[4,108].

Bridging the gap between biologists and computational linguists will be crucial to the success of biomedical literature mining in general, and to its integration with high-throughput experimental data in particular. The field is currently dominated by researchers who have computational backgrounds; however, only biologists possess the knowledge that is required to properly evaluate methods (BOX 2), to identify specific tasks for which tools are needed and to point out other data sources that it would be valuable to integrate with the literature. To bring more

MeSH terms
A controlled vocabulary that is used for annotating Medline abstracts. Several classes of MeSH term exist, the most relevant for literature mining being 'Chemicals and Drugs' (MeSH-D) and 'Diseases' (MeSH-C).

Linkage mapping
A method for localizing genes that is based on the co-inheritance of genetic markers and phenotypes in families over several generations.

biologists into the field, tool developers need to focus more on designing user interfaces that make the tools accessible to non-specialists. Finally, both sides need to contribute to diversity and novelty within this field, as too many researchers currently use the same few methods to solve the same few tasks. We hope that this review will make more biologists aware of the importance of literature mining, and that it will inspire the development of new tools for making the most of the growing bodies of both scientific literature and experimental data.

1. Rebholz-Schuhmann, D. Facts from text — is text mining ready to deliver. *PLoS Biol.* **3**, e65 (2005).
2. Andrade, M. A. & Bork, P. Automated extraction of information in molecular biology. *FEBS Lett.* **476**, 12–17 (2000).
3. Hirschman, L., Park, J. C., Tsujii, J., Wong, L. & Wu, C. H. Accomplishments and challenges in literature data mining for biology. *Bioinformatics* **18**, 1553–1561 (2002).
4. Yandell, M. D. & Majoros, W. H. Genomics and natural language processing. *Nature Rev. Genet.* **3**, 601–610 (2002).
5. Krallinger, M. & Valencia, A. Text-mining and information-retrieval services for molecular biology. *Genome Biol.* **6**, 224 (2005).
6. Asano, S. *et al.* Concerted mechanism of *swe1/wee1* regulation by multiple kinases in budding yeast. *EMBO J.* **24**, 2194–2204 (2005).
7. 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).
8. Wilbur, W. J. & Coffee, L. The effectiveness of document neighboring in search enhancement. *Inf. Process. Manage.* **30**, 253–266 (1994).
9. Renner, A. & Aszodi, A. High-throughput functional annotation of novel gene products using document clustering. *Pac. Symp. Biocomput.* **5**, 50–68 (2000).
10. Iliopoulos, I. Enright, A. J. & Ouzounis, C. A. Textquest: document clustering of Medline abstracts for concept discovery in molecular biology. *Pac. Symp. Biocomput.* **6**, 384–395 (2001).
11. Glenisson, P., Antal, P., Mathys, J., Moreau, Y. & De Moor, B. Evaluation of the vector space representation in text-based gene clustering. *Pac. Symp. Biocomput.* **8**, 391–402 (2003).
12. Marcotte, E. M., Xenarios, I. & Eisenberg, D. Mining literature for protein–protein interactions. *Bioinformatics* **17**, 359–363 (2001).
13. Bhalotia, G., Nakov, P. I., Schwartz, A. S. & Hearst, M. A. *BioText team report for the TREC 2003 genomics track* [online], <http://trec.nist.gov/pubs/trec12/papers/ucal-berkeley.genomics.pdf> (2003).
14. Donaldson, I. *et al.* PreBIND and Textomy — mining the biomedical literature for protein–protein interactions using a support vector machine. *BMC Bioinformatics* **4**, 11 (2003).
15. Kayaalp, M. *et al. Methods for accurate retrieval of MEDLINE citations in functional genomics* [online], <http://trec.nist.gov/pubs/trec12/papers/nlm.genomics.pdf> (2003).
16. Goetz, T. & von der Lieth, C.-W. PubFinder: a tool for improving retrieval rate of relevant PubMed abstracts. *Nucleic Acids Res.* **33**, W774–W778 (2005).
17. Shah, P. K., Jensen, L. J., Boue, S. & Bork, P. Extraction of transcript diversity from scientific literature. *PLoS Comp. Biol.* **1**, e10 (2005).
18. Suomela, B. P. & Andrade, M. A. Ranking the whole MEDLINE database according to a large training set using text indexing. *BMC Bioinformatics* **6**, 75 (2005).
19. Hersh, W. & Bhuptiraju, R. T. *TREC genomics track overview* [online], <http://trec.nist.gov/pubs/trec12/papers/GENOMICS. OVERVIEW3.pdf> (2003).
20. Hersh, W. R. *et al. TREC 2004 genomics track overview* [online], <http://trec.nist.gov/pubs/trec13/papers/GEO.OVERVIEW.pdf> (2004).
21. Büttcher, S., Clarke, C. L. A. & Cormack, G. V. Domain-specific synonym expansion and validation for biomedical information retrieval [online], <http://trec.nist.gov/pubs/trec13/papers/uwaterloo-clarke.geo.Pdf> (2004).
22. Tanabe, L. *et al.* MedMiner: An internet text-mining tool for biomedical information, with application to gene expression profiling. *Biotechniques* **27**, 1210–1217 (1999).
23. Muller, H. M., Kenny, E. E. & Sternberg, P. W. Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biol.* **2**, e309 (2004).

This paper presents an advanced full-text IR tool that is designed for the *Caenorhabditis elegans* research community.

24. Perez-Iratxeta, C., Bork, P. & Andrade, A. M. XplorMed: a tool for exploring MEDLINE abstracts. *Trends Biochem. Sci.* **26**, 573–575 (2001).
25. Hoffmann, R. & Valencia, A. A gene network for navigating the literature. *Nature Genet.* **36**, 664 (2004).
26. Doms, A. & Schroeder, M. GoPubMed: exploring PubMed with the Gene Ontology. *Nucleic Acids Res.* **33**, W783–W786 (2005).
27. Hoffmann, R. *et al.* Text mining for metabolic pathways, signaling cascades, and protein networks. *Sci. STKE* **283**, pe21 (2005).
28. Fukuda, K., Tamura, A., Tsunoda, T. & Takagi, T. Toward information extraction: identifying protein names from biological papers. *Pac. Symp. Biocomput.* **3**, 707–718 (1998).
29. Tanabe, L. & Wilbur, W. J. Tagging gene and protein names in biomedical text. *Bioinformatics* **18**, 1124–1132 (2002).
30. Coller, N., Nobata, C. & Tsujii, J. Extracting the names of genes and gene products with a hidden Markov model. *Int. Conf. Comput. Linguist.* **18**, 201–207 (2000).
31. Chang, J. T., Schutze, H. & Altman, R. B. GAPSCORE: finding gene and protein names one word at a time. *Bioinformatics* **20**, 216–225 (2004).
32. McDonald, R. & Pereira, F. Identifying gene and protein mentions in text using conditional random fields. *BMC Bioinformatics* **6**, S6 (2005).
33. Settles, B. ABNER: an open source tool for automatically tagging genes, proteins, and other entity names in text. *Bioinformatics* **21**, 3191–3192 (2005).
34. Zhou, G., Shen, D., Zhang, J., Su, J. & Tan, S. Recognition of protein/gene names from text using an ensemble of classifiers. *BMC Bioinformatics* **6**, S7 (2005).
35. Krauthammer, M., Rzhetsky, A., Morozov, P. & Friedman, C. Using BLAST for identifying gene and protein names in journal articles. *Gene* **259**, 245–252 (2000).
36. Leonard, J. E., Colombe, J. B. & Levy, J. L. Finding relevant references to genes and proteins in Medline using a Bayesian approach. *Bioinformatics* **18**, 1515–1522 (2002).
37. Mika, S. & Rost, B. Protein names precisely peeled off free text. *Bioinformatics* **20**, i241–i247 (2004).
38. Finkel, J. *et al.* Exploring the boundaries: gene and protein identification in biomedical text. *BMC Bioinformatics* **6**, S5 (2005).
39. Crim, J., McDonald, R. & Pereira, F. Automatically annotating documents with normalized gene lists. *BMC Bioinformatics* **6**, S13 (2005).
40. Fundel, K., Güttler, D., Zimmer, R. & Apostolakis, J. A simple approach for protein name identification: prospects and limits. *BMC Bioinformatics* **6**, S15 (2005).
41. Hanisch, D., Fundel, K., Mevissen, H. T., Zimmer, R. & Fluck, J. ProMiner: rule-based protein and gene entity recognition. *BMC Bioinformatics* **6**, S14 (2005).

This paper describes a simple biomedical ER system that relies primarily on a carefully curated list of synonyms. It was one of the methods that performed best in the BioCreAtIvE assessment.

42. Chen, L., Liu, H. & Friedman, C. Gene name ambiguity of eukaryotic nomenclatures. *Bioinformatics* **21**, 248–256 (2005).

These authors provide a quantitative overview of the causes of gene-name ambiguity, and suggest how researchers and publishers can help to minimize this problem.

43. Gaudan, S., Kirsch, H. & Rebholz-Schuhmann, D. Resolving abbreviations to their senses in Medline. *Bioinformatics* **21**, 3658–3664 (2005).
44. Schijvenaars, B. J. A. *et al.* Thesaurus-based disambiguation of gene symbols. *BMC Bioinformatics* **6**, 149 (2005).
45. Tanabe, L., Xie, N., Thom, L. H., Matten, W. & Wilbur, W. J. GENETAG: a tagged corpus for gene/protein named entity recognition. *BMC Bioinformatics* **6**, S3 (2005).
46. Craven, M. Kumlien, J. Constructing biological knowledge bases by extracting information from text sources. in *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **7**, 77–86 (1999).
47. Cooper, J. W. & Kershenbaum, A. Discovery of protein–protein interactions using a combination of linguistic, statistical and graphical information. *BMC Bioinformatics* **6**, 143 (2005).
48. Ramani, A. K., Bunescu, R. C., Mooney, R. J. & Marcotte, E. M. Consolidating the set of known human protein–protein interactions in preparation for large-scale mapping of the human interactome. *Genome Biol.* **6**, R40 (2005).
49. Stephens, M., Palakal, M., Mukhopadhyay, S., Raje, R. & Mostafa, J. Detecting gene relations from Medline abstracts. *Pac. Symp. Biocomput.* **6**, 483–495 (2001).
50. Blaschke, C. & Valencia, A. The frame-based module of the SUISEKI information extraction system. *IEEE Intell. Syst.* **17**, 14–20 (2002).
51. Stapley, B. J. & Benoit, G. Biobibliometrics: information retrieval and visualization from co-occurrence of gene names in Medline abstracts. *Pac. Symp. Biocomput.* **5**, 529–540 (2000).
52. Jenssen, T. K., Lægreid, A., Komorowski, J. & Hovig, E. A literature network of human genes for high-throughput analysis of gene expression. *Nature Genet.* **28**, 21–28 (2001).

This paper describes an IE system, PubGene, that is based on simple co-occurrence, and shows how it can be used for the interpretion of microarray expression data.

53. Bowers, P. M. *et al.* Prolinks: a database of protein functional linkages derived from coevolution. *Nucleic Acids Res.* **5**, R35 (2003).
54. von Mering, C. *et al.* STRING: known and predicted protein–protein associations, integrated and transferred across organisms. *Nucleic Acids Res.* **33**, D433–D437 (2005).
55. Schlitt, T. *et al.* From gene networks to gene function. *Genome Res.* **13**, 2568–2576 (2003).
56. Wren, J. D. & Garner, H. R. Shared relationship analysis: ranking set cohesion and commonalities within a literature-derived relationship network. *Bioinformatics* **20**, 191–198 (2004).
57. Alako, B. T. *et al.* CoPub Mapper: mining MEDLINE based on search term co-publication. *BMC Bioinformatics* **6**, 51 (2005).
58. Tiffin, N. *et al.* Integration of text- and data-mining using ontologies successfully selects disease gene candidates. *Nucleic Acids Res.* **33**, 1544–1552 (2005).

This study combines tissue-expression data with disease–tissue relationships that were extracted from the literature to predict candidate disease genes.

59. Ding, J., Berleant, d., Nettleton, D. & Wurtelle, E. Mining Medline: abstracts, sentences, or phrases? *Pac. Symp. Biocomput.* **7**, 326–337 (2002).
60. Ray, S. & Craven, M. Learning statistical models for annotating proteins with function information using biomedical text. *BMC Bioinformatics* **6**, S18 (2005).
61. Narayanaswamy, M., Ravikumar, K. E. & Vijay-Shanker, K. Beyond the clause: extraction of phosphorylation information from Medline abstracts. *Bioinformatics* **21**, i319–i327 (2005).
62. Saric, J., Jensen, L. J., Ouzounova, R., Rojas, I. & Bork, P. Extraction of regulatory gene/protein networks from Medline. *Bioinformatics* 26 July 2005 (doi:10.1093/bioinformatics/bti597).
63. Rindflesch, T. C., Tanabe, L., Weinstein, J. N. & Hunter, L. EDGAR: extraction of drugs, genes and relations from the biomedical literature. *Pac. Symp. Biocomput.* **1**, 517–528 (2000).

64. Proux, D., Rechenmann, F. & Julliard, L. A pragmatic information extraction strategy for gathering data on genetic interactions. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **8**, 179–285 (2000).
65. Yakushiji, A., Tateisi, Y., Miyao, Y. & Tsujii, J. Event extraction from biomedical papers using a full parser. *Pac. Symp. Biocomput.* **6**, 408–419 (2001).
66. Daraselia, N. *et al.* Extracting human protein interactions from MEDLINE using a full-sentence parser. *Bioinformatics* **20**, 604–611 (2004).
67. Friedman, C., Kra, P., Yu, H., Krauthammer, M. & Rzhetsky, A. GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics* **17**, S74–S82 (2001).
68. Rzhetsky, A. *et al.* GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data. *J. Biomed. Inform.* **37**, 43–53 (2004).
This paper is a good introduction to NLP-based IE and to the design of complex IE systems such as GeneWays.
69. Temkin, J. M. & Gilder, M. R. Extraction of protein interaction information from unstructured text using a context-free grammar. *Bioinformatics* **19**, 2046–2053 (2003).
70. Hao, Y., Zhu, X., Huang, M. & Li, M. Discovering patterns to extract protein–protein interactions from the literature: part II. *Bioinformatics* **21**, 3294–3300 (2005).
71. Thomas, J., Milward, D., Ouzounis, C., Pulman, S. & Carroll, M. Automatic extraction of protein interactions from scientific abstracts. *Pac. Symp. Biocomput.* **5**, 707–709 (2000).
72. Hearst, M. A. Untangling text data mining. *Proc. Assoc. Comput. Linguist.*, **37**, 3–10 (1999).
73. Swanson, D. R. Fish oil, Raynaud's Syndrome, and undiscovered public knowledge. *Perspect. Biol. Med.* **30**, 7–18 (1986).
This is the original text-mining paper, which shows how new knowledge can be inferred from the existing literature.
74. Blagosklonny, M. V. & Pardee, A. B. Unearthing the gems. *Nature* **416**, 373 (2002).
75. Swanson, D. R. Migrane and magnesium: eleven neglected connections. *Perspect. Biol. Med.* **31**, 526–557 (1988).
76. Swanson, D. R. Somatomedin C and arginine: implicit connections between mutually isolated literatures. *Perspect. Biol. Med.* **33**, 157–186 (1990).
77. Smalheiser, N. R. & Swanson, D. R. Linking estrogen to Alzheimer's disease: an informatics approach. *Neurology* **47**, 809–810 (1996).
78. Swanson, D. R. Intervening in the life cycle of scientific knowledge. *Library Trends* **41**, 606–631 (1988).
79. Smalheiser, N. R. & Swanson, D. R. Assessing a gap in the biomedical literature: Magnesium deficiency and neurological disease. *Neurosci. Res. Commun.* **15**, 1–9 (1994).
80. Weeber, M. *et al.* Text-based discovery in biomedicine: the architecture of the DAD-system. *Proc. AMIA Symp.* **20**, S903–S907 (2000).
81. Srinivasan, P. & Libbus, B. Mining MEDLINE for implicit links between dietary substances and diseases. *Bioinformatics* **20**, i290–i296 (2004).
82. Wren, J. D. Extending the mutual information measure to rank inferred literature relationships. *BMC Bioinformatics* **5**, 145 (2004).
83. Hristovski, D., Peterlin, B., Mitchell, J. A. & Humphrey, S. M. Using literature-based discovery to identify disease candidate genes. *Int. J. Med. Inform.* **74**, 289–298 (2005).
84. Grably, M. R., Stanhill, A., Tell, O. & Engelberg, D. HSF and Msn2/4p can exclusively or cooperatively activate the yeast *HSP104* gene. *Mol. Microbiol.* **44**, 21–35 (2002).
85. Chi, Y. *et al.* Negative regulation of Gcn4 and Msn2 transcription factors by Srb10 cyclin-dependent kinase. *Genes Dev.* **15**, 1078–1092 (2001).
86. Bose, S., Dutko, J. A. & Zitomer, R. S. Genetic factors that regulate the attenuation of the general stress response of yeast. *Genetics* **169**, 1215–1226 (2005).
87. Lenssen, E. *et al.* The Ccr4–Not complex independently controls both Msn2-dependent transcriptional activation — via a newly identified Glc7/Bud14 type I protein phosphatase module — and TFIID promoter distribution. *Mol. Cell. Biol.* **25**, 488–498 (2005).
88. Xiao, Y. & Mitchell, A. P. Shared roles of yeast glycogen synthase kinase 3 family members in nitrogen-responsive phosphorylation of meiotic regulator Ume6p. *Mol. Cell. Biol.* **20**, 5447–5453 (2000).
89. Eiznhamer, D. A., Ashburner, B. P., Jackson, J. C., Gardenour, K. R. & Lopes, J. M. Expression of the *INO2* regulatory gene of *Saccharomyces cerevisiae* is controlled by positive and negative promoter elements and an upstream open reading frame. *Mol. Microbiol.* **39**, 1395–1405 (2001).
90. Kennedy, M. A., Barbuch, R. & Bard, M. Transcriptional regulation of the squalene synthase gene (*ERG9*) in the yeast *Saccharomyces cerevisiae*. *Biochim. Biophys. Acta* **1445**, 110–122 (1999).
91. Hoffmann, R. & Valencia, A. Life cycles of successful genes. *Trends Genet.* **19**, 79–81 (2003).
92. de Lichtenberg, U., Jensen, L. J., Brunak, S. & Bork, P. Dynamic complex formation during the yeast cell cycle. *Science* **307**, 724–727 (2005).
93. Morel, V. & Schweisguth, F. Repression by Suppressor of Hairless and activation by Notch are required to define a single row of *single-minded* expressing cells in the *Drosophila* embryo. *Genes Dev.* **14**, 377–388 (2000).
94. Woods, S. L. & Witelaw, M. L. Differential activities of Murine Single Minded 1 (SIM1) and SIM2 on a hypoxic response element. *J. Biol. Chem.* **277**, 10236–10243 (2002).
95. 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).
96. Blaschke, C., Oliveros, J. C. & Valencia, A. Mining functional information associated with expression arrays. *Funct. Integr. Genomics* **1**, 256–268 (2001).
97. Masys, D. R. *et al.* Use of keyword hierarchies to interpret gene expression patterns. *Bioinformatics* **17**, 319–326 (2001).
98. Chaussabel, D. & Sher, A. Mining microarray expression data by literature profiling. *Genome Biol.* **3**, research0055.1–research0055.16 (2002).
99. Raychaudhuri, S., Schutze, H. & Altman, R. B. Using text analysis to identify functionally coherent gene groups. *Genome Res.* **12**, 1582–1590 (2002).
100. Raychaudhuri, S., Chang, J. T., Imam, F. & Altman, R. B. The computational analysis of scientific literature to define and recognize gene expression clusters. *Nucleic Acids Res.* **31**, 4553–4560 (2003).
101. Glenisson, P. *et al.* TXTGate: profiling gene groups with text-based information. *Genome Biol.* **5**, R43 (2004).
102. Krauthammer, M., Kaufmann, C. A., Gilliam, T. C. & Rzhetsky, A. Molecular triangulation: bridging linkage and molecular-network information for identifying candidate genes in Alzheimer's disease. *Proc. Natl Acad. Sci. USA* **101**, 15148–15153 (2004).
The study shows how literature-based molecular networks and genetic linkage mapping can be integrated to find candidate disease genes.
103. Perez-Iratxeta, C., Bork, P. & Andrade, M. A. Association of genes to genetically inherited diseases using text mining. *Nature Genet.* **31**, 316–319 (2002).
104. Perez-Iratxeta, C., Wjst, M., Bork, P. & Andrade, M. A. G2D: A tool for mining genes associated to disease. *BMC Genetics* **6**, 45 (2005).
Reference 103 integrates genetic linkage-mapping data with data from the literature to suggest candidate genes for inherited diseases. Reference 104 shows later improvements of the method.
105. Korbel, J. O. *et al.* Systematic association of genes to phenotypes by genome and literature mining. *PLoS Biol.* **3**, e134 (2005).
These authors present a method for linking genotypes to phenotypes by comparing species profiles of genes and literature-derived keywords.
106. Shah, P. K., Perez-Iratxeta, C., Bork, P. & Andrade, M. A. Information extraction from full text scientific articles: Where are the keywords? *BMC Bioinformatics* **4**, 20 (2003).
107. Schuemie, M. J. *et al.* Distribution of information in biomedical abstracts and full-text publications. *Bioinformatics* **20**, 2597–2604 (2004).
108. Dickman, S. Tough mining. *PLoS Biol.* **1**, 144–147 (2005).
109. Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A. F. & Nielsen, H. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* **16**, 412–424 (2000).
110. Yeh, A. S., Hirschman, L. & Morgan, A. A. Evaluation of text data mining for database curation: lessons learned from the KDD Challenge Cup. *Bioinformatics* **19**, i331–i339 (2003).
111. Hirschman, L., Yeh, A., Blaschke, C. & Valencia, A. Overview of BioCreAtIvE: critical assessment of information extraction for biology. *BMC Bioinformatics* **6**, S1 (2005).
112. Krauthammer, M. *et al.* Of truth and pathways: chasing bits of information through myriads of articles. *Bioinformatics* **18**, S249–S257 (2002).
113. Perez-Iratxeta, C. & Andrade, M. A. Worldwide scientific publishing activity. *Science* **297**, 519 (2002).
114. Netzel, R., Perez-Iratxeta, C., Bork, P. & Andrade, M. A. The way we write. *EMBO Rep.* **4**, 446–451 (2003).
115. Newman, M. E. J. Coauthorship networks and patterns of scientific collaboration. *Proc. Natl Acad. Sci. USA* **101**, 5200–5205 (2004).

Acknowledgements
The authors would like to thank T. Doerks and S. Hooper for help with figures, and other group members of P.B.'s group at the European Molecular Biology Laboratory and I. Rojas's group at EML Research for valuable discussions. J.S. is funded by the Klaus Tschira Foundation. This work was supported by grants from the European Community and the German Ministry for Education and Science through Nationales Genomforschungsnetz (NGFN).

Competing interests statement
The authors declare no competing financial interests.

DATABASES
The following terms in this article are linked online to:
Entrez Gene:
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene
Epo | *single-minded*
OMIM:
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM
Alzheimer disease | Raynaud disease
UniProtKB: http://www.expasy.org/uniprot
Cdc5 | Cdc28 | Clb2 | Erg9 | Hsp104 | Ino2 | Msn2 | Rim11 | somatomedin C | Ssn3 | Su(H) | Swe1 | Ume6

FURTHER INFORMATION
An extended bibliography of biological literature-mining papers: http://www.bork.embl.de/Docu/literature_mining/
Arrowsmith: http://arrowsmith.psych.uic.edu
BITOLA: http://www.mf.uni-lj.si/bitola
Ensembl: http://www.ensembl.org
G2D: http://www.ogic.ca/projects/g2d_2
GENIA: http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA
iHOP: http://www.pdg.cnb.uam.es/UniPub/iHOP
HighWire Press: http://highwire.stanford.edu
Medline: http://medline.cos.com
MedMiner: http://discover.nci.nih.gov/textmining
NCBI RefSeq: http://www.ncbi.nlm.nih.gov/RefSeq
PennBioIE: http://bioie.ldc.upenn.edu
PubMed: http://www.pubmed.org
PubMed Central: http://www.pubmedcentral.org
STRING: http://string.embl.de
***Saccharomyces* Genome Database:**
http://www.yeastgenome.org
Textpresso: http://www.textpresso.org

SUPPLEMENTARY INFORMATION
See online article: S1 (figure)
Access to this links box is available online.