COMPUTATIONAL TOOLS

Text-mining solutions for biomedical research: enabling integrative biology

Dietrich Rebholz-Schuhmann[1,2], *Anika Oellrich*[1] *and Robert Hoehndorf*[3,4]

Abstract | In response to the unbridled growth of information in literature and biomedical databases, researchers require efficient means of handling and extracting information. As well as providing background information for research, scientific publications can be processed to transform textual information into database content or complex networks and can be integrated with existing knowledge resources to suggest novel hypotheses. Information extraction and text data analysis can be particularly relevant and helpful in genetics and biomedical research, in which up-to-date information about complex processes involving genes, proteins and phenotypes is crucial. Here we explore the latest advancements in automated literature analysis and its contribution to innovative research approaches.

Hypotheses
Testable statements that, if true, may explain an observed phenomenon.

Knowledge bases
Databases of statements covering a knowledge domain. Often, statements are represented in a form that permits the automated or manual inference of statements that are not explicitly stated using inference rules.

[1]*European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK.*
[2]*Institut für Computerlinguistik, Universität Zürich, Binzmühlestrasse 14, 8050 Zürich, Switzerland.*
[3]*Department of Genetics, University of Cambridge, Downing Street, Cambridge CB2 3EH, UK.*
[4]*Department of Physiology, Development and Neuroscience, University of Cambridge, Downing Street, Cambridge, CB2 3EG, UK.*
Correspondence to D.R.-S.
e-mail: rebholz@ebi.ac.uk

doi:10.1038/nrg3337
Published online
14 November 2012

The scientific literature is the key distribution channel for novel findings and hypotheses from research. As the number of publications continuously grows, retrieving relevant scientific information and identifying connections between pieces of scientific knowledge becomes a challenging task[1–3]. As a consequence, automated literature analysis is now frequently a part of complex biomedical research and often delivers crucial background knowledge[4]. In the future, it is likely that solutions will be developed that produce and test hypotheses against the knowledge bases. Automated analysis of literature complements the reading of scientific literature by individual researchers as it allows rapid access to information contained in large volumes of documents and may increase the reproducibility of literature searches by enabling users to process all documents for a specific result[5].

Nowadays, text mining is successfully being applied to the identification of molecular causes of diseases using facts from databases and literature[6–8]. For example, biomedical researchers have to cope with large sets of genes that have been proposed as candidate genes for a given disease. For most complex diseases, some causal genes have been identified — for example, in the case of type 2 diabetes mellitus (T2DM)[9–11] — but we have to assume that the known set is not yet complete. Furthermore, the functional implications and interactions of the candidate genes are not completely understood, thereby limiting the potential for successfully identifying drugs that may target these genes or the pathways in which they act. Thus, a biologist must scan the entire range of scientific literature available on all implicated genes to gather the known facts on their functions and then must focus their research on the most relevant genes (D.R.-S., unpublished observations).

In addition to providing information about genes, the scientific literature can contribute to both phenotype and genotype data[12]. The contribution of these data sharpens clinical and genetic criteria to define disease categories — for example, distinguishing T2DM from related conditions, such as T1DM and obesity — and to select an appropriate treatment[13–15].

Automatic literature analysis and its integration with biomedical data resources have reached new levels of sophistication. The latest developments in text-mining solutions allow a shift from the analysis of abstracts to the analysis of the full text of papers: from the analysis of gene- and protein-related information to the analysis of information about cells, tissues and whole organisms; and from analyses that are entirely literature-based to analyses that integrate information from the literature with data sets from other domains, including genome-wide association studies (GWASs), gene expression, functional genomics, biochemistry and phenotyping[16]. Software applications have been developed that enable us to visualize knowledge contained in the scientific literature and that provide improved integration of text-mining results with other data resources[17]. For the future, we expect that seamless querying and searching across biomedical knowledge will facilitate the systematic generation and exploration of hypotheses as well as the identification of new research topics and existing controversies[18,19].

In this Review, we assess the current state-of-the-art of text mining, focusing on recent developments that are ready to be integrated into the day-to-day research work

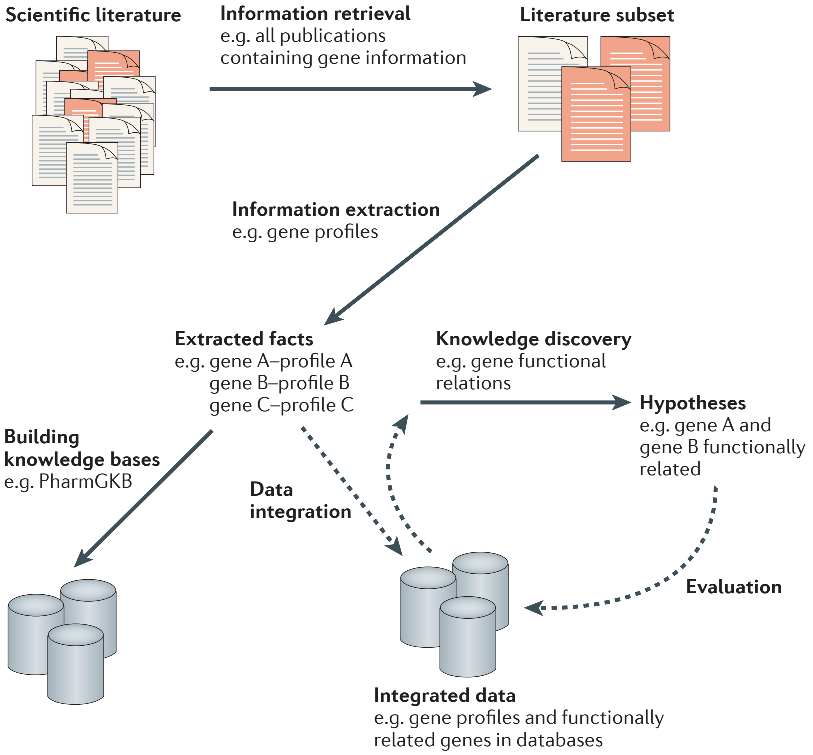


Figure 1 | **Categories of text-mining solutions.** The diagram gives an overview of the different categories of situations in which text mining is applied. Document retrieval is the initial step and leads to the collection of documents for a given query. The other solutions target the identification and evaluation of information that is explicitly stated in the documents.

Facts
Objective and (experimentally) verifiable ways in which the world is structured.

Information retrieval
The process of selecting information or documents from a collection as a result to the submission of a query.

Information extraction
The process of automatically assessing documents, data or knowledge bases to extract statements that are likely to be true given the available information. Information extraction can be based on defined patterns, machine-learning techniques, statistical analyses or automated reasoning.

Knowledge discovery
The process of analysing a set of statements to identify new statements that are true. To discover new knowledge, evidence must have already been gathered in support of the identified statements.

of biological and clinical scientists. We start by providing an insight into the categories of text-mining solutions and continue with an overview of existing tools that can be used as entry points. Furthermore, we outline limitations of the current solutions, provide a summary of challenges and then present an outlook for future work in the field of text mining.

Categories of text-mining solutions

Text mining comprises the discovery and extraction of knowledge from free text[20] and can extend to the generation of new hypotheses by joining the extracted information from several publications[21]. Text-mining solutions can achieve different objectives, depending on the tasks they then have to address. Primarily, we can distinguish four different categories of purposes for text-mining solutions: information retrieval, information extraction, building knowledge bases and knowledge discovery. All of these categories are illustrated in FIG. 1.

In information or document retrieval, the user submits a query to the search engine and receives documents or text passages fitting to the submitted query in return. The documents or text passages are retrieved on the basis of matching keywords contained in the query, matching the query string as such in the documents or matching other data attributed to a scientific publication that form the metadata, such as the title, author names or the journal title.

Information extraction comprises the identification of entities, such as genes or diseases, as well as the identification of complex relationships between those entities, including protein–protein interactions and gene–disease associations. The scientific facts extracted from literature can then either be used to populate databases directly or to assist the work of curation teams. From this, knowledge bases can be built that contain the collected statements together with collected evidence and provenance in the form of references to the scientific literature[5,22]. Specifically, information extraction methods identify statements in literature that may represent facts in the world if true, and evidence and provenance can be used to identify and to resolve possibly contradicting statements.

Knowledge discovery aims at the identification of hidden or as of yet undiscovered knowledge by applying data-mining algorithms to the collection of facts extracted from the literature, potentially including an intermediary step of data integration to other databases[21]. In this category, the text-mining results are used to suggest hypotheses automatically that can then be used to shape or to plan experiments to validate or to disprove the proposed hypotheses.

Retrieving relevant documents

Most researchers regularly use document retrieval to seek background information related to a research question. In the case of the researcher analysing the causes of a disease, the first step is either the retrieval of all documents linked to the mention of the disease or all documents relevant to a selection of genes linked to the disease. These needs can be satisfied by search and retrieval engines (TABLE 1), such as PubMed or UK PubMed Central (UKPMC), as well as by commercial providers of document retrieval engines such as Google Scholar and Elsevier's ScienceDirect[23,24]. All of these have local repositories to hold the documents used for the retrieval.

We can distinguish search engines by the scope of content delivered from the search engine as well as the functionality that is available to the user. PubMed retrieves documents from MEDLINE and thus produces a comprehensive search but delivers only the short summaries of the scientific manuscripts. Recent developments have led to a number of retrieval engines that also process full-text documents from the open-access distribution of the scientific literature and thereby provide access to additional relevant content[16]. PubMed and UKPMC[24] are such engines that are maintained by the US National Center for Biotechnology Information (NCBI) and the European Bioinformatics Institute (EBI), respectively. Also, both of these engines include additional content that becomes freely available from journal publishers only after a predefined embargo period. Publications in other languages or from other sources may be included in the retrieval, such as Chinese biological abstracts, agricultural literature (see the US Department of Agriculture National Agriculture Laboratory website) or patent documents (see the European Patent Office and UKPMC). In

Statements
Declarative sentences that can be said to be either true or false. True statements express facts.

Evidence
The information that has been gathered to demonstrate that the statement is true (that is, it corresponds to a fact); in science, evidence usually contains experimental results.

addition, other search engines, such as PolySearch and the main search engines of the NCBI and the EBI, deliver database content as a part of the retrieval to give the user a comprehensive overview[25].

Some systems offer additional functionality. For example, RefMED adapts the document retrieval to the user's needs by gathering feedback on the 20 best-ranked documents generated by the query[26] and then uses a classifier to optimize the retrieval according to the feedback. In the case of Facta+[27], the user-generated query produces all relevant documents from the index and, in a second retrieval step, all contextual terms for the selected documents. Documents are then further categorized on the basis of the contextual information.

GoPubMed organizes the retrieved documents into a hierarchical structure in which each document is listed under one or several concept labels from the Gene Ontology[28–30] or from the Medical Subject Headings (MeSH) thesaurus. This assignment of categories facilitates navigation through the given set of documents and thus reduces time for browsing the search results. Query expansion is another improvement to the functionality:

Table 1 | **Examples of text-mining tools and resources**

Name	Content	Input	Description	URL
Information retrieval				
PubMed	Abstracts	Standard query	Retrieves abstracts of scientific publications according to user query. Results are provided as a list and can be further filtered with Medical Subject Headings (MeSH) terms and an advanced search functionality	http://www.ncbi.nlm.nih.gov/pubmed
GoPubMed	Abstracts	Standard query	Retrieves publications from MEDLINE and additional functionality by classifying publications according to Gene Ontology concepts to allow improved screening of results	http://www.gopubmed.com/web/gopubmed
RefMED	Any text	Standard query	Allows user to submit feedback and consequently learns how to search PubMed for relevant articles according to feedback provided	http://dm.postech.ac.kr/refmed
UK PubMed Central (UKPMC)	Full text	Standard query	Retrieves full-text documents from PubMed and mines the documents for mentions of genes, drugs and Gene Ontology concepts using the Whatizit infrastructure	http://ukpmc.ac.uk
PolySearch	Abstracts, databases	Standard query	Retrieves information (such as documents and database entries) according to particular patterns of queries. Supports 50 different classes of queries	http://wishart.biology.ualberta.ca/polysearch/index.htm
Information extraction				
Textpresso	Full text	Standard query	Provides extracted statements containing entities of interest on a subset of full text articles. A subset of articles is determined by Textpresso itself: for example, only mouse- or worm-specific articles	http://www.textpresso.org
CoPub	Abstracts	Concepts or identifiers	Retrieves co-occurring biomedical concepts from MEDLINE abstracts. The user specifies a list of concepts or identification numbers and retrieves back an overview about co-occurring concepts divided into categories	http://services.nbic.nl/copub/portal
iHOP	Abstracts	Standard query	Processes MEDLINE abstracts and generates a hyperlinked set of data for protein interactions. iHOP provides interactive functionality for searching genes and related information	http://www.ihop-net.org/UniPub/iHOP
Reflect	Any text	Proteins	Processes documents to highlight proteins and small molecules in the document and to link the entity to reference data resources	http://reflect.embl.de
Open Biomedical Annotator	Any text	Ontologies and configuration parameters	Processes documents to annotate text spans with ontology concepts. Covers all ontologies provided from the BioPortal Web page	http://bioportal.bioontology.org/annotator
Database				
Side Effect Resource (SIDER)			Holds information about the side effects of drugs extracted from drug leaflets and scientific literature	http://sideeffects.embl.de
PharmGKB			Provides information about the influences of genetic variation on drug responses. Information is extracted from scientific literature and is partially curated	http://www.pharmgkb.org
BioCaster			Retrieves disease relevant information from Twitter tweets and shows current hotspots of disease outbreaks on an interactive map	http://born.nii.ac.jp
Transcript Based Isoform Interaction Database (TBIID)			Provides information on human protein isoforms and their differential interactions	http://tbiid.emu.edu.tr
STITCH			Holds known and predicted interactions of small molecules and proteins, partially derived from scientific literature	http://stitch.embl.de

The table gives an overview of data resources and tools that are available to the public. For each category, a selection has been chosen to demonstrate the purpose of that category.

further relevant terms — mainly synonyms — are added to the query to enlarge the retrieval set. For example, when searching for documents about the *Casp1* gene with the query term 'Casp1', documents that contain the string 'caspase 1' instead of 'Casp1' may also be returned.

Further advantages can be achieved by limiting the query and the retrieved results to preselected and pre-identified semantic types. For example, PolySearch expects that the user specifies genes, gene sets or specific biological concepts (for example, as drugs, metabolites or diseases) in the query. It then returns a result set of documents and database entries in which the query term (or terms) is co-mentioned in a sentence of a document or in a database entry with the other specified types of information (that is, the 'query terms' generate a result set of 'database terms')[26]. Another option is the retrieval engine MedEvi, which aims to return only individual sentences or text fragments that match a user's query[31]. This engine identifies combinations of query terms in the sentences of the documents and then aligns the query terms in a tabular view.

Identifying statements from scientific text

Specific biomedical queries cannot always be answered using document retrieval engines. For example, a researcher may wish to retrieve all proteins and their interaction partners involved in a given disease process, possibly including information from all databases in the biomedical domain. Another example might be finding all metabolites that have been mentioned in the literature as a product of a plant or all transcription factors that have been reported in any species and any experimental setting to upregulate the proteins in a specified set. Such queries require more advanced text-mining tools.

Identification of named entities. After documents have been retrieved, text mining can be applied to analyse the content of these documents, in particular to identify statements[32]. These are phrases or syntactic expressions, such as 'Casp3 binds to Fam21', that represent a property, relation or event linked to an entity and can be verified or falsified[33] (BOX 1).

A first step in identifying statements from the literature is to recognize the entities that are referred to within the text (see the figure in BOX 1). In the biomedical domain, for example, text mining has to identify the parts of the text that make reference to proteins and genes (a process called entity recognition) and then has to identify the entry in gene and protein databases, such as UniProt, that corresponds to the particular entity mentioned in the text (a process called entity normalization).

In principle, we can distinguish a dictionary-based approach for entity recognition from a machine-learning approach. In the dictionary-based approach, the entity mentioned in the text is fitted to the best match from the dictionary resource and is then immediately linked to a database entry[2,5]. In the machine-learning approach, the computer program identifies any string mention in the text that resembles an entity mention and requires a secondary analysis to link the entity to the correct database entry. This approach can identify only entities that are referred to in a pre-annotated training corpus: that is, it relies on the quality of the annotations in the corpus. For example, BANNER[34] is a freely available, generic piece of named-entity recognition software that has been trained to identify gene and protein names in the text. Further solutions are available for the identification of species (using, for example, Linnaeus), chemical entities (using, for example, Oscar) and diseases (using, for example, Whatizit)[34–38].

Entity recognition is also provided through user interfaces such as the Open Biomedical Annotator[39]. This tool identifies terms in the text that belong to any of the available ontological resources — for example, in the gene ontology or human phenotype ontology — and returns the annotated text to the user[40]. Another solution is Reflect[41], which analyses the full-text document, selects the entities in the text and links from the entities to related database entries. These tools can be used either to present information directly to researchers or as a processing step to reach data integration or knowledge discovery methods. The search engine CoPub makes extensive use of lexical resources for genes, proteins, Gene Ontology labels, diseases, pathways, drugs and tissues to identify and statistically to qualify the significance of a specific term for a gene or a set of genes[42]. The user receives a set of annotations for their genes of interest. The popular tool Textpresso can be customized by its users — usually a curation team — to include and identify those terms that are specific to the ongoing curation tasks[43].

Entity disambiguation. It remains a challenge of text mining to disambiguate terms with multiple meanings in text: that is, 'polysemous' terms. For every mention of an entity, the correct type of the entity must be identified, as the same term may be used to refer to several types of entities (BOX 1). For example, *Casp1* denotes a gene as well as the corresponding protein and the term 'mice' can represent a species or — with different orthography — the MicE protein. In the case of 'retinoblastoma', the distinction between the gene causing the disease and the disease itself will be difficult to achieve. Similarly, *Streptococcus pneumoniae* could be mistaken as a disease, whereas *Streptococcus pneumoniae 19A serotype* (PMID 22693804) represents a species but contains a part of the first term. However, the distinction between *orb* denoting either a gene or a globe requires input only from the context in which the term appears. More complex disambiguation problems include terms such as 'left breast cancer', in which it remains unclear whether the cancer in the left breast is of a different type from the one in the right breast or whether this is a breast cancer of the left breast only.

Entity normalization follows entity recognition (BOX 1), is also still a challenging task and is mainly required to integrate database and literature content. Solutions for the transformation of gene and protein names into the EntrezGene identifier have been suggested — for example, GENO and GNAT — but no standard solution has been established yet[44,45]. It is, for example, necessary to identify the species in the context of a gene or protein name whenever their names are

Provenance
A reference to literature from which a statement or its supporting evidence were derived.

Terms
Single words or compositions of words with well-defined meanings.

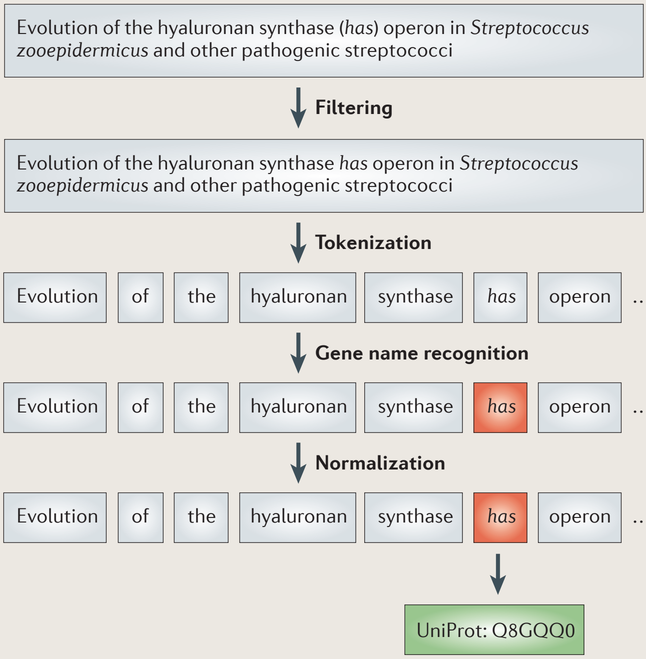
Types
The conceptualization of categories of entities or conceptual instances, represented by a unique identifier, a label and a definition.

Entity recognition
The extraction of text constituents representing a specific type, preferably entities with a name such as a protein.

Entity normalization
The mapping of a named entity or type in the text to a unique identifier, possibly requiring disambiguation and contextual analysis.

Ontology
A representation of a conceptualization of a domain of knowledge, characterizing the classes and relations that exist in the domain. Commonly, ontologies are represented as graph structure that represents a taxonomy.

Box 1 | Text mining: processing steps


Evolution of the hyaluronan synthase (*has*) operon in *Streptococcus zooepidermicus* and other pathogenic streptococci
Filtering
Evolution of the hyaluronan synthase *has* operon in *Streptococcus zooepidermicus* and other pathogenic streptococci
Tokenization
Evolution of the hyaluronan synthase *has* operon …
Gene name recognition
Evolution of the hyaluronan synthase *has* operon …
Normalization
Evolution of the hyaluronan synthase *has* operon …
UniProt: Q8GQQ0

Access to the digital representation of the scientific literature (for example, MEDLINE abstracts or PubMed Central (PMC) full-text documents) requires automatic text processing (key steps are illustrated in the figure). Before any specific text processing is initiated, a number of standard routines are carried out: first, the documents are adapted to a shared data format; second, sections are identified such as the title, the body text and its sections, figure captions, images and the meta-data of the document; and third, the individual sentences are identified.

After pre-processing the document, the following step is 'tokenization' of the document. This is done to identify the constituents of the text (which are called 'tokens'), such as single words, numbers or punctuation. The tokenization step tends to be error-prone if complex terminologies, such as chemical entities, or complex syntax are used within the document (for example, parentheses are a common feature in the syntax for denoting chemical entities). Consecutive tokens can also be combined to determine named entities, such as genes and proteins, chemical entities, drugs, diseases and others. The identification of named entities (that is, named entity recognition) is a complex task, as the identification of the correct boundaries of composed terms and the disambiguation of terms that are used with different meanings are not trivial processes (see the main text for examples)[98].

Contextual information, such as sentence structure, sequence and even discourse structure, is used to disambiguate polysemous terms for entities in the text[99]. In particular, the identification of the syntactical roles (that is, nouns, verbs, adjectives, and so on) of tokens, as determined by automatic part-of-speech tagging, contributes to the disambiguation. For example, the term 'has' may denote either the third-person simple present tense of the verb 'to have', the short form of the protein name 'hyaluronan synthase' or the gene encoding this protein.

A potential next step after the identification of entities is a process called term normalization. In this process, the previously detected entities are linked to preferred terms in shared terminologies or to database identifiers. Therefore, term normalization contributes to the integration of literature with data contained in biomedical resources.

Most information extraction systems specifically identify associations or relations after identifying the entities of interest. The identification of relations is more complex than the other previous steps and can be achieved with different methods. Co-occurrence analysis determines pairs of named entities that are mentioned together in a portion of text such as a sentence. Syntactic parsing analyses the syntactical structure of a sentence to determine the dependencies and then the relations between the entities. Relations comprise the identification of protein–protein interactions and protein interaction networks[60,61], among others.

The underlying mechanisms used for hypothesis generation and knowledge discovery range from basic co-occurrence techniques to complex machine-learning algorithms for identifying meaningful relationships among the extracted scientific facts. Co-occurrence analysis identifies named entities that are mentioned together in a portion of text, such as a sentence, a paragraph, a section or a whole document. Co-occurrences are then analysed using statistical approaches (for example, a hyper-geometric distribution) and methods from information theory to identify important novel related terms. Co-occurrence can be indicative of a biological relation between the identified entities and therefore leads to a novel hypothesis that can be tested.

shared in different species. However, the correct species name is often missing, which creates challenges even for human curators[46].

Identification of relations between named entities. After entities have been recognized in a text, relations that are asserted to hold between the given entities can be identified through text mining. The simplest form of a relation between entities is an association based on the co-occurrence of the entity names in the text: when two entities are frequently mentioned together in a sentence, paragraph, section or document, an association relation between the entities is inferred. This approach was successful in deriving large-scale networks for associated genes and proteins that have been extracted from the complete corpus of MEDLINE abstracts[47].

One prominent example is iHOP[48], a tool with a user interface that provides gene interaction networks extracted from scientific literature. In this solution, the literature search is initiated by defining a gene of interest. A network is then constructed from all scientific publications related to the query-defined gene. By constructing the network, the content of all scientific publications is exploited through abstracting the genes from their specific content and focusing on their associations alone. The user can then browse the results by navigating through the network. However, the network may be incomplete in terms of biology, because not all relations can be found in the literature and some will be missed in the extraction process. Another application that produces network representations from text-mined information is the Diseasome[49]. It provides an interactive map of 22 manually predetermined categories of human disorders and their related genes. The causality of the gene–disease associations is quite diverse, and most of them still have to be confirmed by experimental results or by manual network analysis of the genetic interactions.

More selective approaches for the identification of relations include the extraction of specific types of statements, such as those referring to causal relations between mutations and diseases[50], interactions between genes and proteins[51,52] and relations between environmental features and diseases[53]. To carry out this task, pattern-based approaches are used[54] as well as more sophisticated solutions, such as machine learning, statistical analyses and formal inference[4]. Examples of the use of these approaches include the identification of subcellular locations of proteins[55], the prediction of protein functions on the basis of protein structures[56] and the annotation of mutations and residues of genes and proteins[57].

Solutions have been tailored for biomedical researchers that extract statements from the literature answering specific questions. Features from the context of the gene (for example, tokens or mentions of functions) have been used to contribute to the clustering of genes (for example, from microarray experiments) to identify genes with similar functions[58,59]. The large-scale identification of protein–protein interactions in literature has allowed the creation of protein interaction networks[56,60–62], links between proteins and phosphorylation[63,64] and gene regulatory events[65–67]. Typically, the extracted information is integrated into a database for future exploration.

The availability of biomedical pathways and networks based on large-scale data gathering through text mining offers new opportunities to explain the causality of relationships between biological entities, to provide support for assertions found in the literature using the context provided by the network or automatically to evaluate findings from the literature against database content or formalized knowledge[68]. Biological databases are often focused on one or only a few types of entities — such as proteins in UniProt or drugs and their targets in DrugBank — and therefore it is increasingly important to bring multiple databases and domains together so that their content can be analysed by integrated biomedical analyses[69,70]. The potential of this integration work can be seen in solutions that combine resources in a semantic Web application; this has been done, for example, in neurosciences, Alzheimer's disease and T2DM[71,72] (D.R.-S., unpublished observations).

Building knowledge bases from literature content

As well as using the results of information extraction approaches directly for scientific analyses, they can also be automatically deposited into databases or used to support manual database curation efforts[53,73,74]. A large number of databases now use text mining to gather their data. For example, the Transcript Based Isoform Interaction Database (TBIID) provides statements on protein interactions that have been automatically extracted from MEDLINE abstracts using text classification as well as entity recognition and syntactic parsing[75]. The curated databases for protein interactions either use text classification techniques (such as PRINTS)[76], protein entity recognition (such as MINT)[77] or classification techniques to identify the correct documents (such as BIND)[78]. PharmGKB applies syntactic parsing to identify drug-related information linked to genotype variability[79], and the Side Effect Resource (SIDER)[80] holds information about the side effects of drugs extracted from package inserts as well as scientific publications. BioCaster automatically filters public information streams such as Web pages, news feeds and social media for the mention of diseases to identify disease outbreaks at an early stage[81] and makes all extracted information available in the Biocaster database.

Up-to-date delivery of data into the database and harvesting of most published information require regular re-analysis of the literature to identify updated and novel information. The gain in efficiency through automated literature processing is often compromised by a loss of accuracy in comparison to manually curated databases, and manual database curation can complement text-mining-based information extraction efforts[5]. A very precise text-mining solution — that is, one with a low rate of false-positive results (BOX 2) — allows nearly automatic integration of its accurate results into the curated data resource. By contrast, a text-mining solution with a high recall — that is, one with a low false-negative rate — raises confidence that the results are as comprehensive as possible. The right trade-off between both parameters has to be set on a case-by-case basis. In summary, the combination of automated extraction of information from the literature and manual curation is likely to be irreplaceable in the delivery of database content[68,74].

Text mining for knowledge discovery

Hypothesis generation. Information extraction methods that use text mining alone or in combination with other resources, such as databases, can provide novel insights into existing research questions or might generate novel research questions. For example, predictions from the scientific literature have been used to suggest disease biomarkers, and predictions of protein interactions have been made on the basis of the assumption that two proteins are likely to interact with each other if they share a substantial amount of contextual information[82,83].

A fairly new approach is the combination of text mining and ontologies to generate hypotheses[84]. In this approach, the text-mining tool applies named entity recognition to find terms that have been selected from an ontology (BOX 3). The background knowledge provided

Features
Any constituents of the text — such as tokens, words, complex terms or representations of a concept — that serve as an input to a text-mining solution.

Assertions
Statements that are represented in a formal language to denote the properties or relations of an entity (or concept).

Semantic Web
The extension of the World Wide Web to provide, simultaneously, human- and computer-readable semantics through references to well-defined resources.

Syntactic parsing
Processing of the sentence structure using statistics or grammar rules to produce an electronic representation that delivers logical components (for example, a 'noun phrase'), their roles (for example, the 'subject') and dependencies.

Box 2 | Evaluation of text mining

Different approaches can be applied to verify the validity of the results retrieved with text-mining software. The type of evaluation is dependent on the availability of resources such as databases, labour and time. The most reliable evaluation method is an experimental verification of the results. However, this is not possible for all of the results as it is cost- and time-intensive.

Most evaluation approaches in text mining are based on comparison with a gold-standard corpus (GSC) against which true- and false-positive results, as well as negative results, can be determined. The GSC is a manually (or semi-automatically) extracted and curated reference set: for example, a set of gene–disease associations from a database or manual annotations of abstracts with disease terms.

In the case of manual curation, trained domain experts assess a subset of the results to the best of their knowledge and through consulting scientific literature and other databases. Manual evaluation can only be carried out when results can be judged as true or false by an expert; predictive methods that suggest novel hypotheses cannot easily be evaluated by an expert. As the subset of results to assess is usually chosen arbitrarily, it is assumed to be representative of all the data and thus provides an estimation of the quality of the measured results. The testing of text-mining tools against the GSCs under experimental conditions has led to a number of challenges and competitions, such as BioCreAtIve, BioNLP shared task, Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA), and others, which give better insight into their performances[100].

With the measure of true and false positives based on a text corpus or reference resource, 'precision', 'recall' and 'F measure' are commonly quantitative characteristics. Precision provides a measure of how many of the retrieved documents or entities are correctly retrieved by a method. Recall measures how many of the true positives are actually identified by the method and also missed. The F measure is the harmonic mean of precision and recall. In most cases, precision and recall are negatively co-related: precision can be increased at the cost of the recall and vice versa.

Further means of evaluation include an analysis of the receiver operating characteristic (ROC) curve of a text-mining approach. An ROC curve is a plot of the true-positive rate as a function of the false-positive rate and is widely applied to measure the performance of binary classification methods (for example, a method that distinguishes between gene–disease pairs in which the gene is causally involved in the disease and those pairs in which the gene is not involved in the disease). The area under the curve (AUC) of the ROC curve is a quantitative measure of the performance of the method: a random classification will, asymptotically, result in an AUC of 0.5 (that is, the true-positive rate increases at the same rate as the false-positive rate), whereas a perfect classification will result in an AUC of 1.

by the ontologies — for example, the relations between terms and, in particular, the term hierarchy — can be used to categorize text passages according to an ontological hierarchy or to identify statements that represent instances of ontological knowledge. To generate hypotheses, the results from different resources (for example, literature and databases) can be compared against each other using measures of semantic similarity[85]; the differences between the resources or forms of evidences can be used to improve the ontological knowledge representation. For example, text mining has been used to identify reports of interactions between genes, drugs and phenotypes in the literature; the extracted relations (such as metabolization, inhibition and activation) were mapped to a common ontology, and then drug–drug interactions could be identified using a machine-learning approach[86,87]. Each predicted drug–drug interaction is linked to specific statements that support and explain the hypothetical interaction. Another study identified terms from a phenotype ontology for drug labels and used the results to define a similarity measure between side effects[88]. This work showed that the similarity of drug side effects can be used to hypothesize novel drug targets and drug–drug interactions; 13 predicted drug targets and 20 drug–drug interactions were validated experimentally by *in vitro* binding assays.

Semantic resources
Biomedical ontologies and databases serve as semantic resources, as they define and describe concepts and entities.

Driving integrative research. Together, two processes pave the way towards biomedical resources working in harmony in an openly accessible infrastructure: first, the methods discussed above for linking terms from the literature to semantic resources; and second, the methods for integrating assertions from the literature with annotated content in the life science databases. The development of an integrated infrastructure needs to be supported by semantic Web technology. Such technology would need to enable results from the literature analysis to be distributed — along with provenance information and contextual details — in a computer-readable format. The technology should also ensure that the references required for the semantic Web integrate the extracted data with a growing web of biomedical data[89]. In such an integrated infrastructure, the data extracted through text mining and the data that are directly deposited into research databases would coexist and have equivalent importance.

The development of an integrated infrastructure would have several implications. First, the distinction between statements retrieved from the literature and from other resources would become less visible and less important. Second, new ways of validating statements against existing data would emerge. Specifically, a scientific publication provides information that allows assertions to be weighted according to additional evidence (such as references to other publications), the discourse structure of the manuscript's narrative or whether the assertion is confirmed by other external resources (such as expression-profiling experiments or automatic reasoning against a knowledge base). Third, the scientific publication can be processed with statistical and logical means to identify 'hidden knowledge' that has not been explicitly exposed by the author. This is the case whenever the author has described his or her experiments and results but did not consider all interpretations of the experiments to suggest results that are outside of the researcher's own focus. For example, pain treatment is often assessed in patients with rheumatoid arthritis, and the authors of such a study might focus primarily on the strength of the drug, but the manuscript could also reveal subgroups of patients that differ in their pain response; other researchers might focus on this detail instead.

To achieve integrated biology, interoperability of data resources is required. This can be done through the use of shared ontological resources as has been demonstrated in, for example, Bio2RDF[70,89]. The integration of the literature with such shared resources would use the text-mining solutions that have been discussed in this Review. In particular, named entity normalization means that the literature — including patents and patient records — can be exploited for integration into knowledge bases and for the discovery of new knowledge, such as novel markers or protein–protein

Box 3 | Development and use of ontologies

Ontologies in biology are constructs that are used as hierarchically organized controlled vocabularies. By standardizing the meaning of terms within a domain, ontologies support the integration of databases. Ontologies contain concepts (also called classes, types or terms) of a domain of knowledge, provide stable identifiers for their meaning and additional background knowledge, such as 'is-part-of' relations between anatomical parts of an organism or relations between cell types and their functions.

The BioPortal is an ontology portal that provides access to about 300 ontologies in the biomedical domain. Ongoing efforts that are linked to the development of ontologies include the Open Biological and Biomedical Ontologies (OBO) Foundry[101], which focuses on the development of a set of interoperable ontologies in the biomedical domain based on well-defined principles that are aimed at achieving high-quality ontologies for a wide range of biomedical analyses. It would be an advantage to use all of them in conjunction, but reaching their full interoperability will still require substantial efforts.

Many biomedical ontologies are formalized in the OBO Flatfile Format as well as the Web Ontology Language (OWL)[91]. Software tools to construct ontologies include OBOEdit[102] and the Protege ontology editor[103].

Text-mining solutions such as the National Center for Biomedical Ontology (NCBO) Annotator[104] can exploit biomedical ontologies and identify the labels of the concepts in natural language texts, thereby leading to a tight integration of both the data annotated with ontology concepts and the background information contained in the ontologies.

interactions[71,82,83]. Results reported in the scientific literature — such as a novel association of a genetic variant with a disease — could immediately be integrated into the knowledge infrastructure through the representation of all findings as assertions; these are called 'nanopublications'[90].

Exploiting formal knowledge with reasoning techniques. The transformation of results from text mining and curated databases into formalized representations is being aided by the increasing availability of mature ontologies in multiple biological domains, the development of robust tools for their efficient processing and the use of knowledge representation formalisms such as the Web Ontology Language (OWL)[91]. However, the ontologies — in conjunction with biomedical data resources — have to use stable identifiers across the biomedical domain for full support from OWL.

The analysis of such formalized knowledge is beginning to have an important role in translational biomedical research[92,93]. For example, using a combination of text mining and automated reasoning, phenotype descriptions resulting from mutations in five species were integrated with the phenotypes of human diseases in a single OWL knowledge base[12]. The systematic comparison of phenotypes in animal models with the phenotypes of human diseases has revealed causal mutations with high accuracy[12,94,95]. This demonstrates the importance of bridging domains and scales for translating results from basic research into insights that can improve human health.

The tight integration of text mining with reasoning using formal knowledge provides evidence for generated hypotheses and eliminates unsound hypotheses. On a larger scale, verified statements that are transformed into knowledge bases or ontologies result not only in consistent and validated hypothesis but also in formalized micro-theories for some domains of biology. Such an approach has been implemented as a prototype in the Hanalyzer system, an automated system that 'reads' literature, generates a theory from the information in the text and formulates hypotheses. The application of the Hanalyzer led to the identification of genes relevant to the craniofacial development in mice that have been experimentally confirmed after candidate screening[4]. Such an approach can be taken further by automating the design, execution and interpretation of the experiments. This approach is taken by the Robot Scientist, an autonomous robot that can carry out all steps of the scientific method, including experimentation, analysis, interpretation and reporting of the results[96].

Future challenges

In the future, text mining will need to address several major challenges, including improving literature analysis, exploitation of extracted information, formalization and integration with formalized knowledge bases, inclusion of all areas of scientific investigations in biomedical sciences into the text-mining analysis and exposure of scientists to the results.

Major challenges linked to the processing of the text documents are coordination and co-reference resolution. Coordination is the complex syntactic structure that binds together sentence components (the 'conjuncts') through a coordinator (for example, 'and', 'or' and 'but'). For example, a sentence structure such as '… Dkk-1 and Dkk-2 binding to LRP-5 and LRP-6 …' induces difficulties for automated processing, as the syntactical constructs on either side of both coordinations have to be syntactically and semantically aligned, often leading to alternative and even conflicting interpretations.

A co-reference is a common syntactical structure to refer to an expression of a thing that occurred previously in the text. For example, in the sentence component 'dkk-1 and its binding partners', both 'its' and 'its binding partners' are references to previously mentioned named entities. The co-reference resolution has to identify and match the co-referents to their named entities to determine the correct interpretation of the sentence. Both challenges require a deeper understanding of text content and are especially relevant for longer narratives, such as the full text of documents.

Resolving identities of entities across literature, data and knowledge bases still poses one of the greatest challenges for the integration process. After data integration is achieved, a researcher will face inconsistencies among results from multiple data sources owing to different data representation standards in the primary data. For example, a protein interaction pair that is stored without any further information in different databases could be a part of a protein complex or could be a hormone and its receptor, both of which have different interaction characteristics. Again, formalizing the data across resources requires standardization of statements and assertions using, for example, OWL. The inconsistencies arise from contradictory information and such contradictions impair automated reasoning in the same way as they impair any other scientific analysis.

Automated reasoning
The use of software to derive statements automatically from a knowledge base using inference rules.

Micro-theories
Sets of assertions that share the same topic or that result from the same source. The assertions must be conflict-free within a micro-theory but can contradict other micro-theories.

Hypothesis generation
The selection or creation of hypotheses that can explain a given phenomenon. Commonly, selection criteria regarding relevance, parsimony or consistency with existing knowledge are applied to select the most viable hypotheses for a given phenomenon.

One way to address inconsistencies after data integration is to include the data source, the provenance and evidence for statements. Thus, in the assessment of contradictions and incongruities, information retrieval and extraction, knowledge discovery and hypothesis generation distinguish the information sources and also distinguish explicit discoveries. Inconsistencies can then be systematically addressed and resolved, leading not only to clean repositories of information but also to alternative, inconsistent scientific explanations for phenomena that require further investigation to resolve.

The last type of challenge is the interaction of the biomedical scientists with the huge amount of data (see BOX 4 for a discussion of access to literature); this is particularly important in fields that are experiencing massive increases in data owing to technological developments (such as genomics, transcriptomics and proteomics). The evaluation of hypotheses and research findings is currently carried out by scientists using statistical approaches but with little or no automated support in hypothesis generation and testing. Combining text mining with software for data visualization, integration, evaluation and analysis could aid researchers in coping with big data. Ideally, such a system would integrate the knowledge representation from the literature analysis and scientific databases with data obtained from experimental work and other data sets and would produce statements from the knowledge resources exposing agreements and disagreements with the experimental results.

Conclusion

The availability of the full text of articles, as a consequence of the widespread adoption of open-access publishing, has further increased the amount of information that is available for analysis[24]. We expect that literature databases will become another component in a shared infrastructure of biomedical resources that can be used by scientists and automated systems to identify and to retrieve the most relevant work, to formulate hypothesis, to find supporting and contradicting evidence for hypotheses, to integrate research results into a framework of whole biological systems and to support the translation of research results across domains and into clinical applications.

Box 4 | Access to literature

Text mining not only provides tools to support scientific discovery but also has the potential to improve researchers' abilities to collaborate and disseminate their findings and to provide a means for publishers to enhance the accessibility of a manuscript. A prime example is the Utopia Reader[17], which is a PDF reader that allows the semantic integration of visualization and data analysis tools. Furthermore, the Utopia Reader also links data from the published article to information from text mining and research databases. Text mining and manual annotations have also been used to enrich a research article semantically (see the online annotated version of the *PLoS Neglected Tropical Diseases* article 'Impact of environment and social gradient on *Leptospira* infection in urban slums'). Although most biomedical text-mining tools are designed to identify biomedical entities and relations, journal articles require different types of entities to be recognized, including person names and locations.

Online resources such as CiteULike and Mendeley or specialized client software such as Papers or Paperpile offer possibilities to manage and annotate libraries of scientific manuscripts and to share references with other researchers. Sharing of references, assignment of notes and discussions about manuscripts allow a virtual version of a traditional journal club. Furthermore, users can add meta-data to publications that can be exploited to identify similar research or hubs of publications that are particularly influential within a research community[105].

Imagine a researcher who has sequenced a species or a patient. He or she submits his or her data into the semantic Web, and the required analytical methods for that type of data are automatically used (for example, methods for sequence analysis and delivery of annotations from knowledge databases)[97]. The automatic analysis includes a judgement on the deviation of the genetic code or its annotations from the reference genome and in addition provides hypotheses for the interpretation of the data. Finally, the scientific literature supporting these novel hypotheses are provided for reference. All data will be compiled to provide a rough draft of a manuscript that serves as an input to the scientist's next publication, including all references to data, resources and literature. This is a future vision of text mining, literature analysis and integrative biology.

1. Jensen, L. J., Saric, J. & Bork, P. Literature mining for the biologist: from information retrieval to biological discovery. *Nature Rev. Genet.* **7**, 119–129 (2006).
2. Kim, J. J. & Rebholz-Schuhmann, D. Categorization of services for seeking information in biomedical literature: a typology for improvement of practice. *Brief. Bioinformat.* **9**, 452–465 (2008).
This manuscript exploits assumptions and observations linked to search behaviour from users of Web pages to judge the information-seeking behaviour of scientists. It judges available text-mining tools according to these assumptions.
3. Altman, R. B. *et al.* Text mining for biology—the way forward: opinions from leading scientists. *Genome Biol.* **9** (Suppl. 2), S7 (2008).
4. Leach, S. M. *et al.* Biomedical discovery acceleration, with applications to craniofacial development. *PLoS Comput. Biol.* **5**, e1000215 (2009).
5. Hirschman, L. *et al.* Text mining for the biocuration workflow. *Database* **2012**, bas020 (2012).
6. Perez-Iratxeta, C., Bork, P. & Andrade, M. A. Association of genes to genetically inherited diseases using data mining. *Nature Genet.* **31**, 316–319 (2002).
7. Perez-Iratxeta, C., Wjst, M., Bork, P. & Andrade, M. A. G2d: a tool for mining genes associated with disease. *BMC Genetics* **6**, 45 (2005).
8. Blagosklonny, M. V. & Pardee, A. B. Conceptual biology: unearthing the gems. *Nature* **416**, 373 (2002).
9. Malandrino, N. & Smith, R. J. Personalized medicine in diabetes. *Clin. Chem.* **57**, 231–240 (2011).
10. Herder, C. & Roden, M. Genetics of type 2 diabetes: pathophysiologic and clinical relevance. *Eur. J. Clin. Invest.* **41**, 679–692 (2011).
11. McCarthy, M. I. Progress in defining the molecular basis of type 2 diabetes mellitus through susceptibility-gene identification. *Hum. Mol. Genet.* **13** (Suppl. 1), 33–41 (2004).
12. Hoehndorf, R., Schofield, P. N. & Gkoutos, G. V. PhenomeNET: a whole-phenome approach to disease gene discovery. *Nucleic Acids Res.* **39**, e119 (2011).
The authors describe their approach to the integration of phenotype resources to judge gene–disease associations. The paper demonstrates the potential of phenotype descriptions in the understanding of biological processes.
13. Li, S. *et al.* Genetic predisposition to obesity leads to increased risk of type 2 diabetes. *Diabetologia* **54**, 776–782 (2011).
14. O'Rahilly, S. Human genetics illuminates the paths to metabolic disease. *Nature* **462**, 307–314 (2009).
15. Smith, R. J. *et al.* Individualizing therapies in type 2 diabetes mellitus based on patient characteristics: what we know and what we need to know. *J. Clin. Endocrinol. Metab.* **95**, 1566–1574 (2010).

16. Cohen, K. B., Johnson, H. L., Verspoor, K., Roeder, C. & Hunter, L. E. The structural and content aspects of abstracts versus bodies of full text journal articles are different. *BMC Bioinformatics* **11**, 492 (2010).

17. Attwood, T. K. *et al.* Utopia documents: linking scholarly literature with research data. *Bioinformatics* **26**, i568–i574 (2010).

18. Kim, J. J., Zhang, Z., Park, J. C. & Ng, S. K. BioContrasts: extracting and exploiting protein-protein contrastive relations from biomedical literature. *Bioinformatics* **22**, 597–605 (2006).

19. Rzhetsky, A., Iossifov, I., Loh, J. M. & White, K. P. Microparadigms: chains of collective reasoning in publications about molecular interactions. *Proc. Natl Acad. Sci. USA* **103**, 4940–4945 (2006).

This article explores how authors report on their results and how the collection of reported facts can be traced, compared and evaluated against each other. It gives early indications of what results might be produced if we applied automatic reasoning to the information from scientific literature and other resources.

20. Hearst, M. A. Untangling text data mining. *Proc. 37th Annu. Meeting Assoc. Comput. Linguistics* **1999**, 3–10 (1999).

21. Swanson, D. R. Medical literature as a potential source of new knowledge. *Bull. Med. Libr. Assoc.* **78**, 29–37 (1990).

22. Karamanis, N. *et al.* Natural language processing in aid of FlyBase curators. *BMC Bioinformatics* **9**, 193 (2008).

23. Sayers, E. W. *et al.* Database resources of the national center for biotechnology information. *Nucleic Acids Res.* **40**, D13–D25 (2012).

24. McEntyre, J. R. *et al.* UKPMC: a full text article resource for the life sciences. *Nucleic Acids Res.* **39**, D58–D65 (2011).

25. Cheng, D. *et al.* PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites. *Nucleic Acids Res.* **36**, 399–405 (2008).

26. Yu, H. *et al.* Enabling multi-level relevance feedback on PubMed by integrating rank learning into DBMS. *BMC Bioinformatics* **11** (Suppl. 2), S6 (2010).

27. Tsuruoka, Y., Tsujii, J. & Ananiadou, S. Facta: a text search engine for finding associated biomedical concepts. *Bioinformatics* **24**, 2559–2560 (2008).

28. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nature Genet.* **25**, 25–29 (2000).

29. Consortium, G. O. The gene ontology: enhancements for 2011. *Nucleic Acids Res.* **40**, D559–D564 (2012).

30. Doms, A. & Schroeder, M. GoPubMed: exploring PubMed with the Gene Ontology. *Nucleic Acids Res.* **33**, W783–W786 (2005).

31. Kim, J. J., Pezik, P. & Rebholz-Schuhmann, D. Medevi: retrieving textual evidence of relations between biomedical concepts from MEDLINE. *Bioinformatics* **24**, 1410–1412 (2008).

32. Cohen, K. B. & Hunter, L. Getting started in text mining. *PLoS Comput. Biol.* **4**, e20 (2008).

33. Brachman, R. J. & Levesque, H. J. *Knowledge Representation and Reasoning* (Elsevier, 2004).

34. Leaman, R. & Gonzalez, G. BANNER: an executable survey of advances in biomedical named entity recognition. *Pac. Symp. Biocomput.* **2008**, 652–663 (2008).

35. Gerner, M., Nenadic, G. & Bergman, C. Linnaeus: A species name identification system for biomedical literature. *BMC Bioinformatics* **11**, 85 (2010).

36. Jimeno, A. *et al.* Assessment of disease named entity recognition on a corpus of annotated sentences. *BMC Bioinformatics* **9**, S3 (2008).

37. Jessop, D. M., Adams, S. E., Willighagen, E. L., Hawizy, L. & Murray-Rust, P. OSCAR4: a flexible architecture for chemical text-mining. *J. Cheminform.* **3**, 41 (2011).

38. Rebholz-Schuhmann, D., Arregui, M., Gaudan, S., Kirsch, H. & Jimeno, A. Text processing through web services: calling Whatizit. *Bioinformatics* **24**, 296–298 (2008).

39. Shah, N. H. *et al.* Comparison of concept recognizers for building the open biomedical annotator. *BMC Bioinformatics* **10**, S14 (2009).

40. Noy, N. F. *et al.* Bioportal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Res.* **37**, W170–W173 (2009).

41. Pafilis, E. *et al.* Reflect: augmented browsing for the life scientist. *Nature Biotech.* **27**, 508–510 (2009).

42. Frijters, R. *et al.* CoPub: a literature-based keyword enrichment tool for microarray data analysis. *Nucleic Acids Res.* **36**, W406–W410 (2008).

43. 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).

44. Wermter, J., Tomanek, K. & Hahn, U. High-performance gene name normalization with GeNo. *Bioinformatics* **25**, 815–821 (2009).

45. Hakenberg, J., Plake, C., Leaman, R., Schroeder, M. & Gonzalez, G. Inter-species normalization of gene mentions with GNAT. *Bioinformatics* **24**, i126–i132 (2008).

46. Leitner, F. *et al.* The *FEBS Letters*/BioCreative II.5 experiment: making biological information accessible. *Nature Biotech.* **28**, 897–899 (2010).

47. 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).

48. Hoffmann, R. & Valencia, A. Implementing the iHOP concept for navigation of biomedical literature. *Bioinformatics* **21** (Suppl. 2), ii252–ii258 (2005).

49. Goh, K.-I. *et al.* The human disease network. *Proc. Natl Acad. Sci. USA* **104**, 8685–8690 (2007).

50. Feldman, I., Rzhetsky, A. & Vitkup, D. Network properties of genes harboring inherited disease mutations. *Proc. Natl Acad. Sci. USA* **105**, 4323–4328 (2008).

51. Krallinger, M. *et al.* How to link ontologies and protein–protein interactions to literature: text-mining approaches and the BioCreative experience. *Database* **2012**, bas017 (2012).

52. Ananiadou, S., Pyysalo, S., Tsujii, J. & Kell, D. B. Event extraction for systems biology by text mining the literature. *Trends Biotechnol.* **28**, 381–390 (2010).

53. Geifman, N. & Rubin, E. Towards an age-phenome knowledge-base. *BMC Bioinformatics* **12**, 229 (2011).

54. Hearst, M. A. Automatic acquisition of hyponyms from large text corpora. *Proc. 14th Conf. Comput. Ling.* **2**, 539–545 (1992).

55. Brady, S. & Shatkay, H. EpiLoc: a (working) text-based system for predicting protein subcellular location. *Pac. Symp. Biocomput.* **2008**, 604–615 (2008).

56. Jaeger, S., Gaudan, S., Leser, U. & Rebholz-Schuhmann, D. Integrating protein-protein interactions and text mining for protein function prediction. *BMC Bioinformatics* **9**, S2 (2008).

57. Nagel, K., Jimeno-Yepes, A. & Rebholz-Schuhmann, D. Annotation of protein residues based on a literature analysis: cross-validation against UniProtKb. *BMC Bioinformatics* **10** (Suppl. 8), S4 (2009).

58. Blaschke, C., Oliveros, J. C. & Valencia, A. Mining functional information associated with expression arrays. *Funct. Integr. Genom.* **1**, 256–268 (2001).

59. Kuffner, R., Fundel, K. & Zimmer, R. Expert knowledge without the expert: integrated analysis of gene expression and literature to derive active functional contexts. *Bioinformatics* **21**, (Suppl. 2), i259–i267 (2005).

60. Blaschke, C., Andrade, M. A., Ouzounis, C. & Valencia, A. Automatic extraction of biological information from scientific text: protein–protein interactions. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **1999**, 60–67 (1999).

61. Hunter, L. *et al.* OpenDMAP: an open source, ontology-driven concept analysis engine, with applications to capturing knowledge regarding protein transport, protein interactions and cell-type-specific gene expression. *BMC Bioinformatics* **9**, 78 (2008).

The work presented in this paper demonstrates the information technology infrastructure required to process conceptual knowledge and to derive novel findings.

62. Oda, K. *et al.* New challenges for text mining: mapping between text and manually curated pathways. *BMC Bioinformatics* **9** (Suppl. 3), S5 (2008).

63. Narayanaswamy, M., Ravikumar, K. E. & Vijay-Shanker, K. Beyond the clause: extraction of phosphorylation information from MEDLINE abstracts. *Bioinformatics* **21**, i319–i327 (2005).

64. Yuan, X. *et al.* An online literature mining tool for protein phosphorylation. *Bioinformatics* **22**, 1668–1669 (2006).

65. Saric, J., Jensen, L. J. & Rojas, I. Large-scale extraction of gene regulation for model organisms in an ontological context. *In Silico Biol.* **5**, 21–32 (2005).

66. Rodriguez-Penagos, C., Salgado, H., Martinez-Flores, I. & Collado-Vides, J. Automatic reconstruction of a bacterial regulatory network using natural language processing. *BMC Bioinformatics* **8**, 293 (2007).

67. Kim, J. & Rebholz-Schuhmann, D. Improving the extraction of complex regulatory events from scientific text by using ontology-based inference. *J. Biomed. Semantics* **2**, S3 (2011).

68. Rzhetsky, A., Seringhaus, M. & Gerstein, M. Seeking a new biology through text mining. *Cell* **134**, 9–13 (2008).

The authors argue that the exploitation of the scientific literature will serve as an additional resource for the generation of hypotheses and the validation of human-driven hypotheses.

69. Samwald, M. & Stenzhorn, H. Establishing a distributed system for the simple representation and integration of diverse scientific assertions. *J. Biomed. Semantics* **1** (Suppl. 1), S5 (2010).

70. Sansone, S. A. *et al.* Toward interoperable bioscience data. *Nature Genet.* **44**, 121–126 (2012).

71. Neumann, E. & Prusak, L. Knowledge networks in the age of the semantic Web. *Brief. Bioinformat.* **8**, 141–149 (2007).

72. Gao, Y. *et al.* SWAN: A distributed knowledge infrastructure for Alzheimer disease research. *J. Web Semant.* **4**, 222–228 (2006).

73. Dowell, K. G., McAndrews-Hill, M. S., Hill, D. P., Drabkin, H. J. & Blake, J. A. Integrating text mining into the MGI biocuration workflow. *Database* **2009**, bap019 (2009).

74. Jamieson, D. G., Gerner, M., Sarafraz, F., Nenadic, G. & Robertson, D. L. Towards semi-automated curation: using text mining to recreate the HIV-1, human protein interaction database. *Database* **2012**, bas023 (2012).

75. KafkasŞ., Varoğlu, E., Rebholz-Schuhmann, D. & Taneri, B. Diversity in the interactions of isoforms linked to clustered transcripts: a systematic literature analysis. *J. Proteom. Bioinf.* **4**, 250–259 (2011).

76. Attwood, T. K. *et al.* Prints and its automatic supplement, preprints. *Nucleic Acids Res.* **31**, 400–402 (2003).

77. Licata, L. *et al.* MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res.* **40**, D857–D861 (2012).

78. 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).

79. Thorn, C. F., Klein, T. E. & Altman, R. B. Pharmacogenomics and bioinformatics: PharmGKB. *Pharmacogenomics* **11**, 501–505 (2010).

80. Kuhn, M., Campillos, M., Letunic, I., Jensen, L. J. & Bork, P. A side effect resource to capture phenotypic effects of drugs. *Mol. Syst. Biol.* **6**, 343 (2010).

In this study, semantic resources for the description of phenotypes were used to determine effects induced by drugs, (that is, the authors identify effects and side effects of drugs).

81. Collier, N. *et al.* BioCaster: detecting public health rumors with a Web-based text mining system. *Bioinformatics* **24**, 2940–2941 (2008).

BioCaster is an information technology solution that monitors public information streams, such as Twitter, to detect expressions that indicate disease outbreaks. This study demonstrates that social information in combination with scientific information can be very useful for the prediction of disease-related events.

82. Elkin, P. L., Tuttle, M. S., Trusko, B. E. & Brown, S. H. BioProspecting: novel marker discovery obtained by mining the bibleome. *BMC Bioinformatics* **10** (Suppl. 2), S9 (2009).

83. van Haagen, H. H. *et al.* Novel protein-protein interactions inferred from literature context. *PLoS ONE* **4**, e7894 (2009).

84. Ceci, F., Pietrobon, R. & Goncalves, A. L. Turning text into research networks: information retrieval and computational ontologies in the creation of scientific databases. *PLoS ONE* **7**, e27499 (2012).

85. Pesquita, C. *et al.* Metrics for GO based protein semantic similarity: a systematic evaluation. *BMC Bioinformatics* **9** (Suppl. 5), S4 (2008).

86. Coulet, A., Shah, N. H., Garten, Y., Musen, M. & Altman, R. B. Using text to build semantic networks for pharmacogenomics. *J. Biomed. Informat.* **43**, 1009–1019 (2010).

87. Percha, B., Garten, Y. & Altman, R. B. Discovery and explanation of drug-drug interactions via text mining. *Pacific Symp. Biocomput.* **2012**, 410–421 (2012).
88. Campillos, M., Kuhn, M., Gavin, A.-C., Jensen, L. J. & Bork, P. Drug target identification using side-effect similarity. *Science* **321**, 263–266 (2008).
89. Belleau, F., Nolin, M. A., Tourigny, N., Rigault, P. & Morissette, J. Bio2RDF: towards a mashup to build bioinformatics knowledge systems. *J. Biomed. Inform* **41**, 706–716 (2008).
90. Patrinos, G. P. *et al.* Microattribution and nanopublication as means to incentivize the placement of human genome variation data into the public domain. *Hum. Mutat.* 26 June 2012 (doi:10.1002/humu.22144).
91. Grau, B. *et al.* OWL 2: The next step for OWL. *Web Semantics* **6**, 309–322 (2008).
92. Jensen, L. J. & Bork, P. Ontologies in quantitative biology: A basis for comparison, integration, and discovery. *PLoS Biol.* **8**, e1000374 (2010).
93. Chen, H., Yu, T. & Chen, J. Y. Semantic web meets integrative biology: a survey. *Brief. Bioinf.* 6 April 2012 (doi:10.1093/bib/bbs014).
94. Chen, C.-K. *et al.* Mousefinder: candidate disease genes from mouse phenotype data. *Hum. Mutat.* **33**, 858–866 (2012).
95. Washington, N. L. *et al.* Linking human diseases to animal models using ontology-based phenotype annotation. *PLoS Biol.* **7**, e1000247 (2009).
96. King, R. D. *et al.* The automation of science. *Science* **324**, 85–89 (2009).
The authors mimicked genuine scientific work through automatic analysis of experimental results, derivation of novel hypotheses and by controlling a robot to execute novel experiments. Text mining and literature analysis played an important part in the interpretation of the results from the data mining step to generate valid hypotheses.
97. Wilkinson, M. D., Vandervalk, B. & McCarthy, L. The semantic automated discovery and integration (SADI) Web service design-pattern, API and reference implementation. *J. Biomed. Semantics* **2**, 8 (2011).
SADI is a framework that registers Web-based services in such a way that they can be easily detected for the processing of data in the Web. Such work helps to set the stage for future progress towards experimental data residing and data analysis occurring on the Web to improve efficiency and to generate new hypotheses.
98. Krauthammer, M. & Nenadic, G. Term identification in the biomedical literature. *J. Biomed. Inform.* **37**, 512–526 (2004).
99. Liakata, M., Saha, S., Dobnik, S., Batchelor, C. & Rebholz-Schuhmann, D. Automatic recognition of conceptualization zones in scientific articles and two life science applications. *Bioinformatics* **28**, 991–1000 (2012).
100. Krallinger, M. *et al.* Evaluation of text-mining systems for biology: overview of the Second BioCreative community challenge. *Genome Biol.* **9** (Suppl. 2), S1 (2008).
101. Smith, B. *et al.* The obo foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotech.* **25**, 1251–1255 (2007).
102. Richter, J. D., Harris, M. A. A., Haendel, M. & Lewis, S. Obo-edit — an ontology editor for biologists. *Bioinformatics* **23**, 2198–2200 (2007).
103. Noy, N. F. *et al.* Creating semantic web contents with Protege-2000. *IEEE Intelligent Systems* **16**, 60–71 (2001).
104. Jonquet, C., Shah, N. H. & Musen, M. A. The open biomedical annotator. *Summit Translat. Bioinforma* **2009**, 56–60 (2009).
105. Douglas, S. M., Montelione, G. T. & Gerstein, M. PubNet: a flexible system for visualizing literature derived networks. *Genome Biol.* **6**, R80 (2005).

Competing interests statement

The authors declare no competing financial interests.

FURTHER INFORMATION

BioCaster: http://born.nii.ac.jp
CoPub: http://services.nbic.nl/copub/portal
CiteULike: http://www.citeulike.org
Diseasome: http://diseasome.eu
European Bioinformatics Institute: http://www.ebi.ac.uk
European Patent Office: http://www.epo.org
Google Scholar: http://scholar.google.co.uk
GoPubMed: http://www.gopubmed.org/web/gopubmed
iHOP: http://www.ihop-net.org/UniPub/iHOP
Mendeley: http://www.mendeley.com
National Agricultural Laboratory Catalog: http://agricola.nal.usda.gov
National Center for Biotechnology Information: http://www.ncbi.nlm.nih.gov
***Nature Reviews Genetics* Series on Computational tools:** http://www.nature.com/nrg/series/computational/index.html
OBO Flatfile Format Syntax and Semantics: http://www.cs.man.ac.uk/~horrocks/obo
Open Biomedical Annotator: http://bioportal.bioontology.org/annotator
Paperpile: http://paperpile.com
Papers: http://www.mekentosj.com/papers
***PLoS Neglected Tropical Diseases*: Impact of environment and social gradient on *Leptospira* infection in urban slums:** http://imageweb.zoo.ox.ac.uk/pub/2008/plospaper/latest
PharmGKB: http://www.pharmgkb.org
PolySearch: http://wishart.biology.ualberta.ca/polysearch/index.htm
PubMed: http://www.ncbi.nlm.nih.gov/pubmed
Reflect: http://reflect.embl.de
RefMED: http://dm.postech.ac.kr/refmed
ScienceDirect.com: http://www.sciencedirect.com
SIDER — Side Effect Resource: http://sideeffects.embl.de
Textpresso: http://www.textpresso.org
Transcript Based Isoform Interaction Database (TBIID): http://tbiid.emu.edu.tr
UK PubMed Central: http://ukpmc.ac.uk

ALL LINKS ARE ACTIVE IN THE ONLINE PDF