Review TRENDS in Biotechnology Vol.24 No.12

Text mining and its potential applications in systems biology

Sophia Ananiadou[1,2], Douglas B. Kell[3,4] and Jun-ichi Tsujii[1,2,5]

[1] School of Computer Science, National Centre for Text Mining, The Manchester Interdisciplinary Biocentre, The University of Manchester, 131 Princess St, Manchester M1 7ND, UK
[2] National Centre for Text Mining, The Manchester Interdisciplinary Biocentre, The University of Manchester, 131 Princess St, Manchester M1 7ND, UK
[3] School of Chemistry, The University of Manchester, Faraday Building, Sackville St, Manchester M60 1QD, UK
[4] The Manchester Centre for Integrative Systems Biology, The Manchester Interdisciplinary Biocentre, The University of Manchester, 131 Princess St, Manchester M1 7DN, UK
[5] Department of Computer Science, University of Tokyo, Room 615, 7th Building of Science, University of Tokyo, Hongo 7-3-1, Bunkyo-ku, Tokyo, 113-0033, Japan

With biomedical literature increasing at a rate of several thousand papers per week, it is impossible to keep abreast of all developments; therefore, automated means to manage the information overload are required. Text mining techniques, which involve the processes of information retrieval, information extraction and data mining, provide a means of solving this. By adding meaning to text, these techniques produce a more structured analysis of textual knowledge than simple word searches, and can provide powerful tools for the production and analysis of systems biology models.

Introduction

With an overwhelming amount of biomedical knowledge recorded in texts, it is not surprising that there is so much interest in techniques that can identify, extract, manage, integrate and exploit this knowledge; moreover, discover new, hidden or unsuspected information. It is noteworthy to compare the number of MEDLINE® searches in March 2006 (82.027 million) with the number in January 1997 (0.163 million) (Figure 1). MEDLINE® contains ~15 million references to journal articles in the life sciences, and its size is increasing at a rate of more than 10% each year. With the popularity of open access journal publishing, such as BioMed Central (http://www.biomedcentral.com/), full text articles are becoming more available. The availability of huge textual resources provides the scientist with the chance to search for correlations or associations such as protein–protein interactions [1,2] and gene–disease associations [3–6]; however, the traditional information retrieval framework, which relies on keyword-based approaches, cannot address this information overload. For this reason, scientists have focussed their attention on text mining (TM) techniques, which enable them to collect, maintain, interpret, curate and discover the knowledge needed for research or education efficiently and systematically.

Consequently, in the past five years there has been an upsurge of research papers and overviews [7–12] on the topic of TM from biomedical literature – the primary goal of TM [13] is to retrieve knowledge that is hidden in text and to present the distilled knowledge to users in a concise form. TM can discover associations, patterns and clusters of related texts; however, even such abstractions from the raw data provide users with a bewildering number of possible associations. One important potential area for the application of TM is systems biology. In most analyses, systems biology involves the iterative interplay between computational modelling, high-throughput and high-content experimentation, and technology development [14]. Being a highly interdisciplinary subject, it involves collating knowledge from wide areas of biology and the exact sciences, and is a particularly favourable domain for the exploitation of TM technology. The main purpose of this review is to outline the basic techniques of TM and to set down some areas of their application to modern systems biology.

Hypothesis generation using TM

Systems biology is one of the key examples of a field where the mode of scientific knowledge discovery is shifting from a hypothesis-driven mindset to an integrated holistic mode that combines hypotheses with data [14,15]. Data in systems biology can be found in heterogeneous forms, including structured data from databases, experimental data and unstructured data from texts. The amount of unstructured textual data is increasing at such a pace it is difficult to discover knowledge and generate scientific hypotheses without the use of knowledge extraction techniques, which are largely based on TM. In the data-rich but hypothesis-poor sciences [16], including functional genomics and most of biomedicine, the normative hypothesis-driven, deductive scientific method becomes increasingly difficult to sustain because it is unable to deliver advances in knowledge quickly enough [17]. As a complement to hypothesis-driven deductive science, we are now witnessing the emergence of data-driven inductive methods of scientific discovery. These are characterized by the rapid 'mining' of candidate hypotheses from the literature, which are then subsequently tested or validated against available experimental data. The notion of

Corresponding author: Ananiadou, S. (sophia.ananiadou@manchester.ac.uk).
Available online 12 October 2006.

 0167-7799/$ – see front matter doi:10.1016/j.tibtech.2006.10.002

Glossary

Annotation: a layer of representation attached to text. For example, linguistic annotations reveal the linguistic structure in text; thus, the sentence 'secretion of TNF was abolished by BHA' can be encoded as NN (noun: secretion), IN (preposition: of), NN (noun: TNF). Syntactic tree annotations show the syntactic structure of the phrase 'secretion of TNF' is a NP (noun phrase) composed of a NP (noun phrase) realized into secretion (NN) and a PP (prepositional phrase) containing a preposition, of, (IN) and a NP TNF (NN). Semantic annotations reveal terms and named entities in the text, for example, genes and proteins.

Deep parsing: provides relationships not explicitly stated among words in a sentence. For example, in the sentence 'p53 is shown to activate transcription', deep parsing encodes this information as 'p53' is a subject of the predicate 'to activate' and 'transcription' is an object.

Homonyms: words having the same form but different meaning, for example, gene names can be the same as general language words, such as *amid*, *can* or *for*.

F-measure: the harmonic mean of the precision (or sensitivity) and recall (or specificity) values, that is, F = (2.specificity.sensitivity/specificity + sensitivity).

Full parsing: finds deep syntactic relations from the whole of a sentence, for example, a relation between a passive verb and its semantic object. We demonstrated that the human AMID gene promoter was activated by p53, where p53 is the subject of the sentence, and human AMID gene promoter is the object.

Parsing: (syntactic parsing) involves assigning a syntactic structure to a sentence using grammar and a dictionary.

Precision: measures the proportion of the entities and/or relations that the system has returned correctly: it measures the accuracy of the system.

Predicate argument structure: a normalized form representing syntactic relations, as in the example 'ENTITY1/NN ACTIVATE/VB ENTITY2/NN'. In this sentence, activate is the predicate, which contains the main meaning of the predicate argument structure, and ENTITY1 and ENTITY2 are its arguments, carrying information about the participants described by a predicate.

Ontologies: conceptual models that aim to support consistent and unambiguous knowledge sharing and provide a framework for knowledge integration.

Recall: measures the proportion of correct entities and/or relations the system has returned: it measures the 'coverage' of the system.

Sensitivity: the conditional probability that the case is correctly classified {=true positives/[true positives + false negatives]}.

Specificity: the conditional probability that non-cases are correctly classified {=true negatives/[true negatives + false positives]}

Term: the linguistic realisation of a specialized concept in a given domain. Unlike words, the main purpose of terms is the classification of specialized knowledge.

Training data (training corpus): used for statistical and/or machine-learning approaches in text mining. Training data have to be carefully constructed (they should not be too general nor too specific) to avoid skewed results.

Token: the elementary, linguistically plausible unit (e.g. word, number, punctuation symbol,). A text is segmented into tokens as an important and necessary pre-processing step in natural language processing.

conceptual biology [18] complementing empirical evidence [19] is, to some extent, driven by the increasing availability of large textual digital repositories but most crucially by the TM tools that add value to them. Swanson pioneered the research of knowledge discovery from texts [20] by exploring the benefits of inferring associations in a series of experiments using simple semi-automated methods to aid human discovery (Arrowsmith; http://arrowsmith.psych.uic.edu/arrowsmith_uic/index.html). Titles from MEDLINE® were used to make connections between seemingly dissociated arguments: the connection between migraine and magnesium deficiency [21], which has been subsequently validated experimentally; between indomethacin and Alzheimer's disease [22]; and between *Curcuma longa* and retinal diseases, Crohn's disease and disorders related to the spinal cord [23]. Weeber *et al.* [24] used similar techniques, based on bibliographic evidence, to suggest using thalidomide for treating a series of diseases such as acute pancreatitis and chronic hepatitis C.

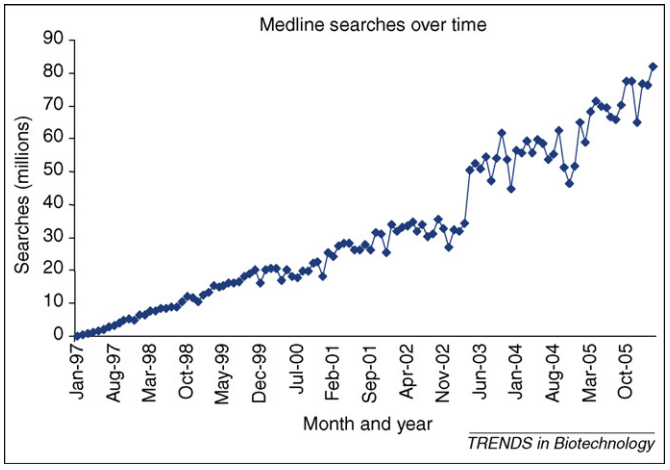

Figure 1. The growth of Medline access. Data are replotted from http://www.nlm.nih.gov/bsd/medline_growth_508.html.

Hypothesis generation in TM relies on the fact that 'chance' connections or associations between disconnected entities or facts can emerge to be meaningful. Here, we shall briefly describe the building blocks of the TM and natural language processing techniques that enable us to make associations for hypothesis generation (Box 1).

TM steps for knowledge discovery

The process of TM encompasses the following major steps: information retrieval, information extraction and data mining.

Information retrieval

Information retrieval (IR) [25] is the activity of finding documents that answer an information need with the aid of indexes. Almost all computer users make habitual use of IR systems (search engines) such as Google™. The user, however, is nevertheless faced with reading many documents to discover the facts reported in them. Apart from general-purpose search engines, many IR tools have been designed specifically to query the databases of biomedical publications such as PubMed© [26–29]. Systems based on IR techniques include:

- Textpresso (http://www.textpresso.org/) [30] – uses a custom ontology to query a collection of documents for information on specific classes of biological concepts (e.g. gene, allele or cell) and their relations (e.g. association and/or regulation).

Box 1. Natural language processing and text mining

Natural language processing (NLP) is the activity of processing natural language texts by computer to access their meaning. NLP systems can analyze (parser) natural language using lexical resources (dictionaries), where words have been organized into groups after a grammar (syntactic level) and a semantic layer has assigned meaning to these words or groups of words. Text mining discovers and extracts knowledge from unstructured data, whereas data mining discovers knowledge from structured data. In this view, text mining comprises three major activities: information retrieval, which gathers relevant texts; information extraction, which identifies and extracts a range of specific types of information from texts of interest; and data mining, which finds associations among the pieces of information extracted from many different texts.

- Query Chem [31] (http://www.QueryChem.com) – combines chemical structure with text based IR using chemical databases and Web API (Google) to retrieve information and relations between compound structures and their properties.
- iHOP [32] (http://www.ihop-net.org/UniPub/iHOP/) – visualizes the interactions between genes.
- EBIMed(http://www.ebi.ac.uk/Rebholz-srv/ebimed/index.jsp) – also retrieves sentences based on detecting co-occurrences between biological entities.
- GoPubMed (http://www.gopubmed.org/) – a thesaurus-driven, GO-based abstract-classification system.
- PubMatrix (http://pubmatrix.grc.nia.nih.gov/) – searches PubMed comparing lists of terms (see Glossary), for example, genes or proteins, given by the user with a set of functionalities, and outputs several papers associated with the list of terms and the functionalities [27].
- PubFinder (http://www.glycosciences.de/tools/PubFinder) – leverages a small set of seed abstracts provided by the user that are relevant to a specific scientific topic. The result is a 'hit-list' of references ranked according to likelihood.

Information Extraction

When the aim is to identify and tabulate the facts reported in large numbers of documents in a literature source, information extraction (IE) becomes a more relevant technology. The goal is to extract information from text without requiring the end user of the information to read the text. Having used a search engine, the user must read each document to know the facts reported in it. IE can be used to support a fact-retrieval service or as a step towards text mining based on conceptually annotated text.

Data mining

Data mining (DM) is used to discover unsuspected associations between known facts extracted by IE – this step encapsulates the integration of text mining with data mining. Most data mining techniques applied to biology assume highly structured biological data, unlike the unstructured textual data used by TM techniques. Unstructured textual data have already been used to improve the results of PSI–BLAST (position specific iterated BLAST; http://helix.nih.gov/docs/gcg/psiblast.html), and sequence homology searches [33–35] have successfully integrated TM with DM for the sequence-based functional classification of proteins using supervised machine-learning methods. Finally, and because these clusters are rarely properly validated [36], TM has been used to go a step further from gene expression clustering and interpret these clusters by associating them with published literature [37,38]. In the following sections, we concentrate on the challenges of terminological processing and novel techniques for information extraction.

Recognizing biological entities in text. Why it is difficult?

Terms are the backbone of specialized knowledge because they denote the biological entities of the documents. Unfortunately, the naming of biological entities is often inconsistent and imprecise [39]. Metabolites, proteins and genes often have a variety of names (terms) for denoting the same concept. For example, the metabolite glucose-6-phosphate is referred to as variants and permutations of α or β, D- or L-glucose (or hexose)-6-(mono)-phosphate. Furthermore, within the same text a term can be given in an extended compounded form then later expressed through various mechanisms, including orthographic variation (usage of hyphens and slashes e.g. amino acid and amino-acid), lower and upper cases (NF-KB and NF-kb), spelling variations (tumour and tumor), various Latin and/or Greek transliterations (oestrogen and estrogen), and abbreviations (RAR and retinoic acid receptor). Further complexity is introduced when authors vary the forms they use in different ways (e.g. different reductions, such as thyroid hormone receptor and thyroid receptor, or the SB2 gene and SB2) or use embedded variant forms within larger forms (CREB-binding protein, where CREB is actually cAMP-response-element-binding protein). Therefore, a term is increasingly viewed as an equivalence class of termforms, the rich variety of which have to be recognized, indexed, linked and mapped to the abundant biological databases and ontologies (see Glossary). Ontologies are crucial for text mining because they provide semantic interpretation to text and also constrain the possible interpretations of biological entities (terms) (Figure 2): when we provide semantic interpretation to text, we link terms to concepts in ontologies [40,41], whereby textual evidence is used to update and to maintain existing ontologies [42].

Named entity recognition (NER) is the first step of IE. NER relies on automatic term recognition (ATR), which extracts terms from a collection of documents (whereas ATR detects terms as opposed to general language words). It assigns to terms such as 'monocyte' an appropriate label, for example, CELL_TYPE, and the recognized terms are than classified into broader domain classes (e.g. genes, proteins or tissues). An example of such a system is TerMine (http://www-tsujii.is.s.u-tokyo.ac.jp/termine/), which is currently developed by the National Centre for Text Mining (http://nactem.ac.uk/).

The majority of approaches in the biology domain integrate term recognition and term classification in a single step. Some approaches have been based on using dictionaries [1,43] to locate terms in the text; however, many terms cannot be recognized if we use straightforward dictionary or database look-up owing to term variation and homonyms (see Glossary). For example, when names from FlyBase were used as a terminological source for recognition of gene names in the literature, the results showed an extremely low precision (see Glossary) – 2% for full articles and 7% for abstracts – with recall (see Glossary) in the range 31% (for abstracts) to 84% (for full articles) [43]. This is, of course, species-dependent, and for the fly the results are not as good as those for yeast are. To overcome the problems inherent to dictionary-based approaches, several groups [44,45] suggested machine learning and probabilistic techniques to deal with term variation. These boosted their performance.

Rule-based systems use rules that describe common naming structures for certain term classes, based on

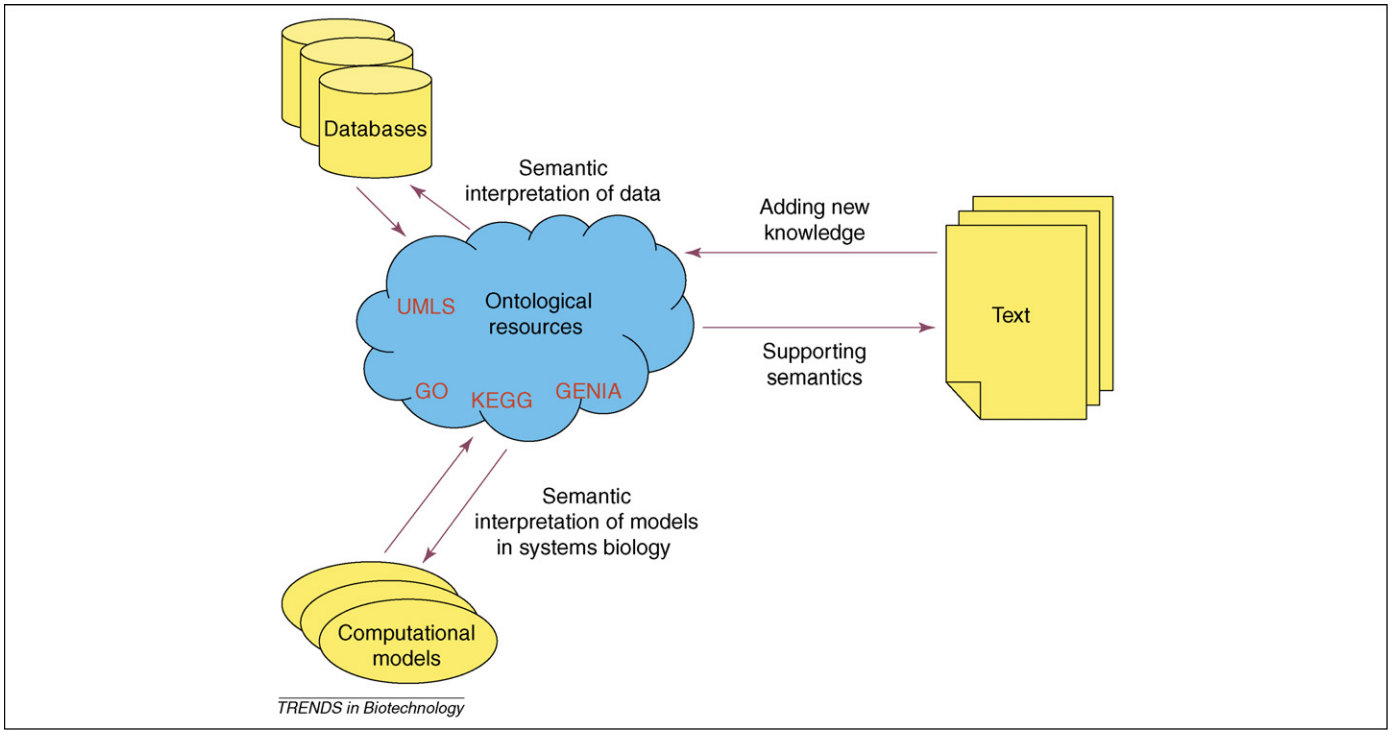

Figure 2. Relationships between text, databases, systems biology models and ontological resources. Ontologies provide descriptions of biological concepts and their relations. Linking domain-specific terms to their descriptions in the ontologies provides a platform for semantic interpretation of textual information. An explicit semantic layer, supported by the use of ontologies, enables text to be mined for interpretable information about biological concepts. The knowledge extracted from text using advanced TM can then be curated and used to update the content of biomedical ontologies, which currently lag behind in their attempts to keep abreast of new knowledge because of its rapid expansion. Scientific databases, systems biology models and textual information are associated with each other through ontologies.

morphological, orthographic and syntactic characteristics [46,47]. These perform better than the other methods described above but have to be customized to new domains. By contrast, machine-learning techniques [48–52] depend on the existence of training data (see Glossary) to learn features that are useful and relevant for NER.

Text mining tasks, particularly NER and IE, use the standard evaluation metrics of precision, recall and F-measure (see Glossary). Most text mining systems for NER and IE (relation extraction) use the F-measure to evaluate their results. The majority of systems compare their results with a 'gold standard': the most popular annotated corpora used by the text mining community as gold standards are GENIA (http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/) [53] and PennBioIE (http://bioie.ldc.upenn.edu/publications/latest_release/data/index.html; Box 2).

Box 2. Evaluating the results of text mining processes

It is imperative when we evaluate text mining tasks that these should be important to the biology community [54]. Two main biological tasks have been used for text-mining-evaluation challenges: document retrieval and biological database curation.

Recent challenge evaluations for text mining in biology include:

- KDD Challenge Task 1 (2002) IE from Biomedical Articles (http://www.biostat.wisc.edu/~craven/kddcup/tasks.html)
- TREC Genomics Track (2003, 2004) (http://trec.nist.gov/pubs/trec12/t12_proceedings.html) [55]
- BioCreAtIvE (2004) [56] (http://biocreative.sourceforge.net/)
- BioNLP, JNLPBA (http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/ERtask/report.html)

Term variation and ambiguity hamper the recognition of biomedical entities

Term variation and term ambiguity make the identification of biological entities difficult. As mentioned above, a concept can be denoted by various realizations, which are known as term variants. A particularly common term variation type in biology is representation by acronyms. In MEDLINE™ abstracts, 64 242 new acronyms were introduced in 2004, with the estimated total being 800 000 [57]. It was reported [58] that 5477 documents could be retrieved by using the acronym JNK, whereas only 3773 documents could be retrieved by using its full term, c-jun N-terminal kinase. Finding all the term variants in text and linking them to the same concept is important for improving the results of IR, to avoid irrelevant information from being retrieved (low precision) and relevant information being overlooked (low recall).

Acronym recognition aims to extract pairs of short forms (acronyms), for example, ADM, and their long (expanded) forms – adrenomedullin abductor digiti minimi adriamycin – occurring in text. Existing methods for acronym recognition can be categorized into three groups: using heuristics and/or scoring rules [59–61]; machine learning [57,62]; and statistical methods [63].

Term ambiguity occurs when the same term refers to many concepts. An example of term ambiguity is the term promoter, which refers to a binding site in a DNA chain, at which RNA polymerase binds to initiate transcription of mRNA by one or more nearby structural genes, whereas in chemistry it refers to a substance that in small amounts can increase the activity of a catalyst. Homologues and

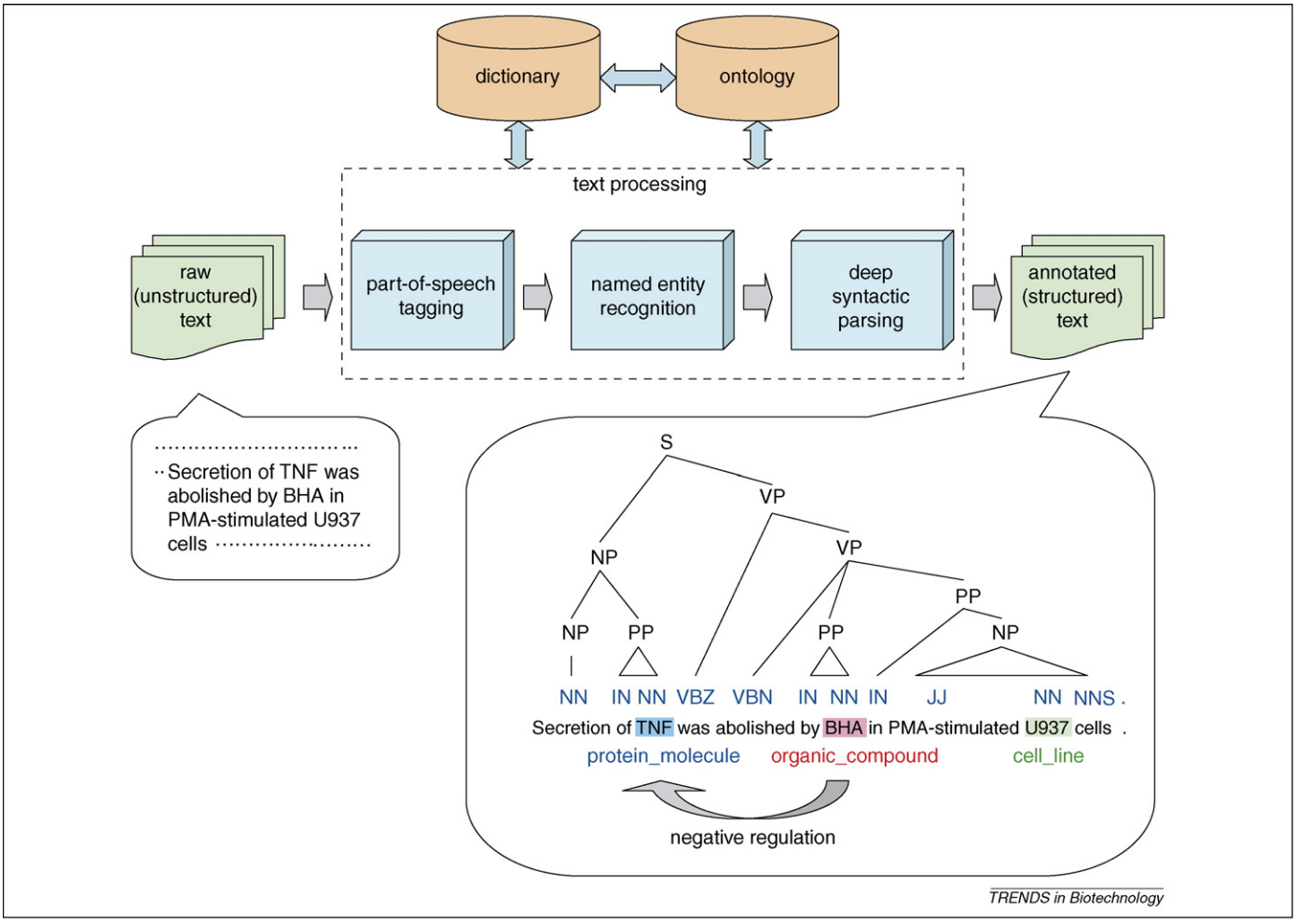

Figure 3. A text mining pipeline from unstructured to annotated data using part-of-speech tagging, named entity recognition and syntactic analysis (parsing), using external resources (i.e. dictionaries and ontologies). Each module enhances text representation with a layer of annotation, which represents explicit linguistic and/or semantic information attached to text in machine-usable form. Such information is inferred by a human reader using (i) linguistic and general knowledge, and (ii) domain-specific expertise. However, for the text to be analysed automatically at a higher semantic level, such knowledge has to be explicitly represented in a machine-readable form. The given figure illustrates the output representation of the sentence 'secretion of TNF was abolished by BHA in PMA-stimulated U937 cells' with multiple layers of annotation, including syntactic, semantic and ontology related.

ad hoc choices of gene or protein names (e.g. yotiao) aggravate the problem of ambiguity. This problem is alleviated by using term disambiguation approaches, such as supervised techniques to assign automatically gene names to their LocusLink ID [64]; applied machine learning techniques to disambiguate genes, proteins and RNA in text [65]; and manual rules combined with a variety of supervised and unsupervised approaches [66]. One way to ensure that mappings to other public databases are consistent is to use data resolution software based on life science identifiers (http://lsid.sourceforge.net/).

Information extraction

Recognizing biological entities is the first step of IE (Figure 3). Text is typically tokenized (see Glossary), to identify the limits of words and sentences, then tagged (part-of-speech tagging) by assigning labels such as NOUN, VERB or ADJECTIVE to each word. Syntactic analysis (Figure 4) identifies the basic textual chunks of a sentence.

To detect and extract the types of evidence needed for hypothesis generation, we need semantic interpretation of the text, upon which we base relation extraction. Relation extraction extracts pairs or triples of biological entities, for example, p53 INDUCES *Peg3* or *Pw1* mRNA expression.

Some IE systems in biology use pattern matching approaches [67], which sometimes have limited generalisation. Moreover, the closer the analysis is to the text, the more patterns are needed to take account of the large amount of surface grammatical variation in texts. Their main limitation is that some measure of semantic processing beyond pattern matching is required that is superior to either text strings or annotations (see Glossary) connected with surface analyses.

Other approaches use a combination of syntactic and semantic parsing (see Glossary). McDonald [68] deploys rules weighted from corpus evidence for extracting pathway relations, whereas Šaric [69] extracts relations concerning the regulation of gene expression. A more promising approach is sublanguage-based IE systems, which exploit the linguistic particularities of the biological language [70,71] to good effect. Few IE systems [72–75] use deep linguistic knowledge (full parsing; see Glossary). The advantage of full parsing is that we can easily make generalizations for more than one type of biological interaction. To achieve this generalisation, we use predicate

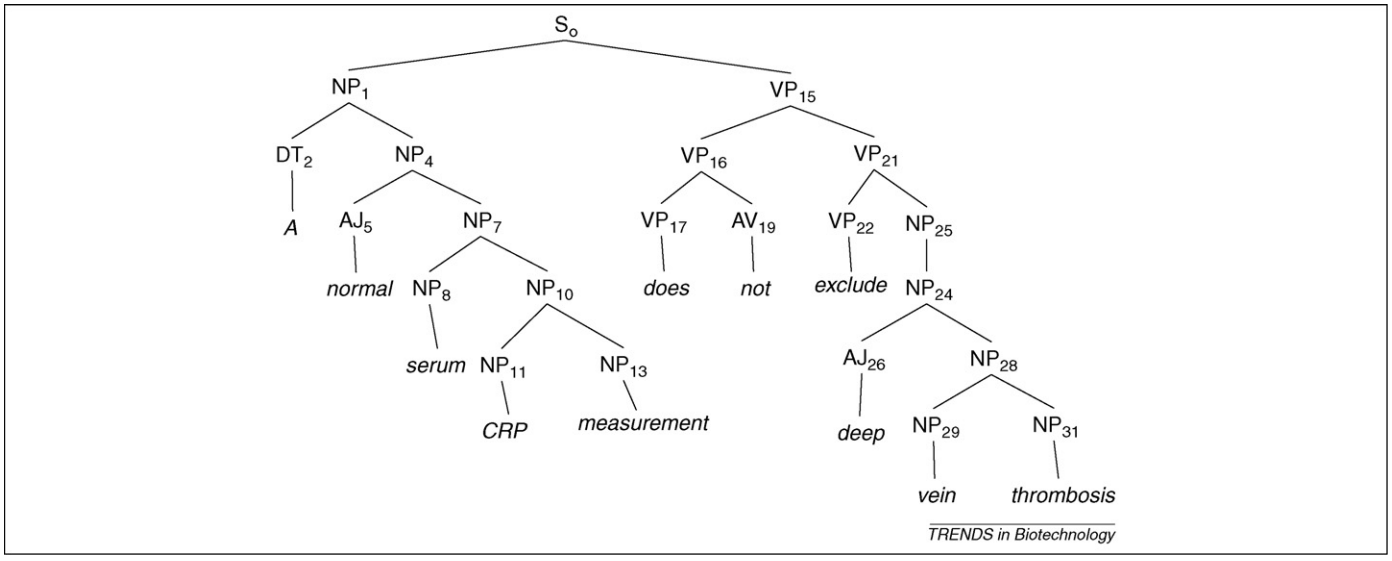

Figure 4. A typical output of syntactic analysis, showing how a sentence S_0 is chunked into noun phrases (NP), verb phrases (VP), adjectives (AJ) and determiners (DT).

argument structures (PAS; see Glossary), which are canonical representations of sentence meanings that represent relations in an abstract manner (Figure 5).

The importance of using PAS is that all the syntactic variations of the sentences in Box 3 can be normalized into one structure {'activate' ARG1 Entity 1 [semantic subject] ARG2 Entity 2 [semantic object]}. We have applied this approach to extract protein–protein interactions [75,76] from the whole of Medline.

Full parsing tuned to biology has been realized in two systems that are currently used at the National Centre for Text Mining: MEDIE (http://www-tsujii.is.s.u-tokyo.ac.jp/medie), which retrieves relations and their concepts from the whole of Medline as a real time application; and InfoPubMed (http://www-tsujii.is.s.u-tokyo.ac.jp/info-pubmed/), which combines full parsing and machine learning techniques to recognize different types of interactions (e.g. inhibit, enhance or promote) between genes and proteins, based on ontological information. MEDIE enables the user to perform semantic querying, thus going beyond keyword searching (Figure 6). These systems rely on a combination of deep linguistic knowledge, the richness of annotations obtained from biological resources (ontologies) and efficient parsing technologies to produce biological interactions and relations extracted from text, thus enabling the biologist to make generalisations and generate hypotheses.

Potential applications in systems biology: from fact discovery to hypothesis generation and model construction

Text mining techniques can be applied in a variety of areas of systems biology, and some applications are already beginning to emerge [75,77–79]. Although these are early days, it is clear that direct linkage of biochemical and signalling models to the literature that underpins them is a goal that is now within our grasp. Systems biology modelling [14,80] starts with a qualitative or structural model: these are commonly derived from genome sequences [81–84] and can be integrated, clearly and usefully, with literature-derived evidence. We note that this task (given the pathway, find the literature) is considerably easier than its converse [71]. A particular area of significance involves hunting for the parameters of the individual reactions of systems biology models [78] because these are mandatory for the purposes of ordinary differential equation (ODE) modelling [85]. The structure, equations and parameters, including starting or

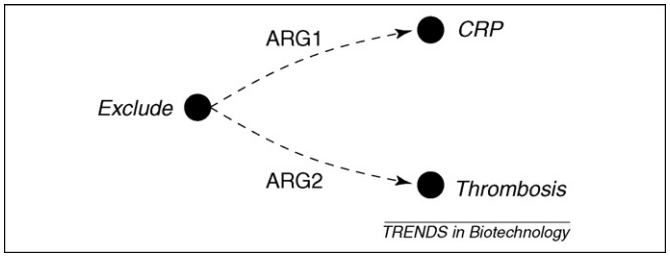

Figure 5. A predicate argument structure (PAS) for the verb exclude, linking it with CRP and thrombosis. PAS is a normalized form representing syntactic relations. In this sentence, exclude is the predicate, which contains the main meaning of the predicate argument structure, and CRP and thrombosis are its arguments, carrying information about the participants described by a predicate. ARG1 denotes a subject relation and ARG2 an object relation. This PAS represents the sentence 'CRP excludes thrombosis'.

Box 3. Syntactic variations

Syntactic variations characterize the different realizations of a predicate (verb); thus, for the verb 'activate', we can have several syntactic variations.

Active Main Verb: Entity1 recognizes and activates Entity2.

After an Auxiliary Verb: Entity1 can activate Entity2 through a region in its carboxyl terminus.

Passive: Entity2 is activated by Entity1a and Entity1b

Past Particle: Entity2 activated by Entity1 are not well characterized.

Verb of a relative clause: the herpes virus encodes a functional Entity1 that activates human Entity2.

Infinitive: Entity1 can functionally cooperate to synergistically activate Entity2.

Gerund in a prepositional phrase: the Entity1 has key roles by activating Entity2.

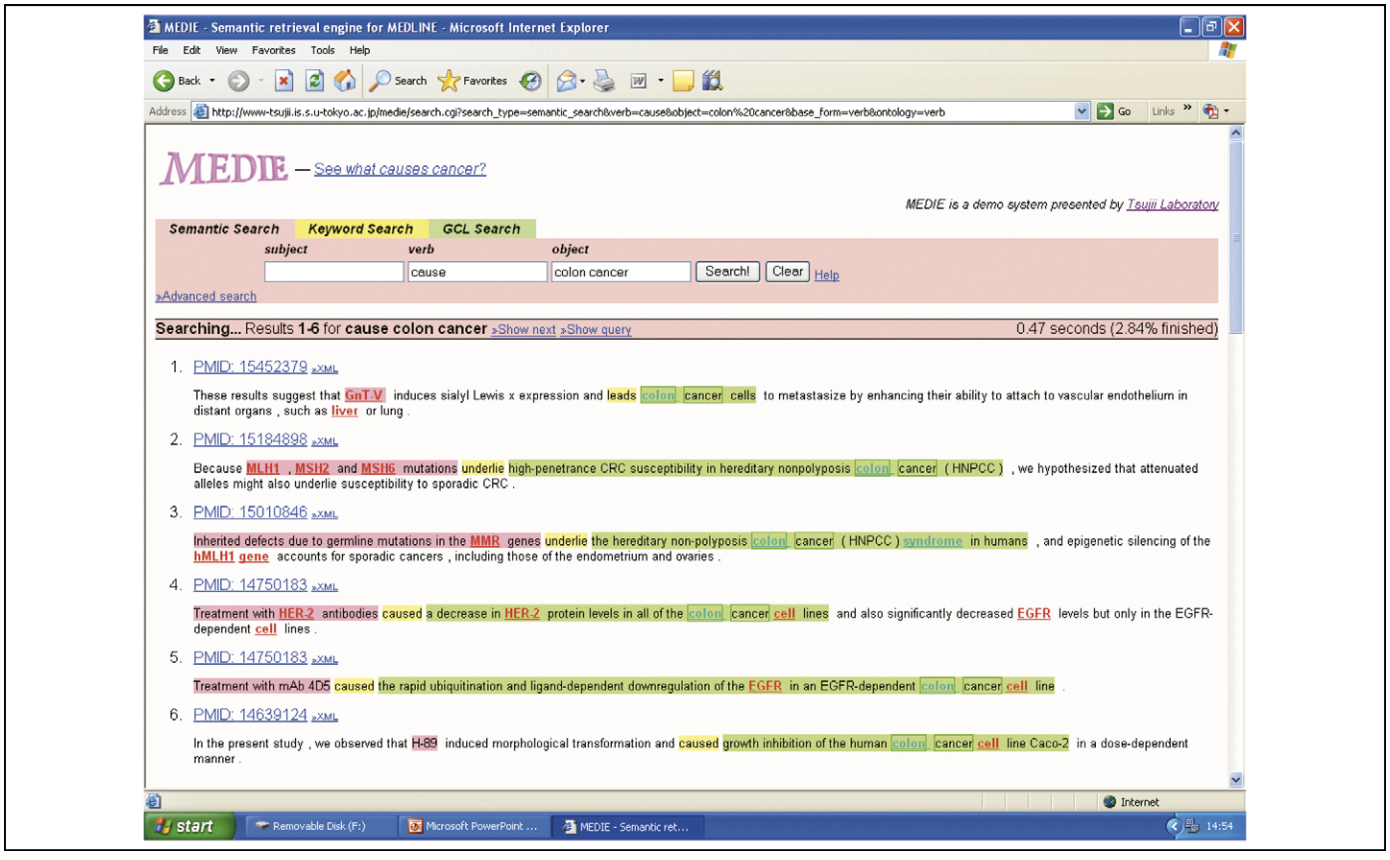

Figure 6. This figure shows part of the output of MEDIE answering the query 'what causes colon cancer?' The answers are retrieved from 14 785 094 MEDLINE articles, indexed by full parsing analysis and using predicate argument structures.

fixed concentrations, define the ODE system and can be stored in a transmissible form as SBML [86] (systems biology markup language; http://www.sbml.org). A desirable goal now is to extend SBML to include the evidence for the models it describes. Another alternative is to have a BioPAX file [87] (http://www.biopax.org/) linked to a complementary SBML file. Overall, we have taken the view [14] that distributed environments using systems such as Taverna [88,89], or others [90–92,] to enact the necessary bioinformatic workflows might well provide the best way forward for linking systems biology modeling activities. Because the difficulties of interoperability seem, in fact, to be much more about data structures (syntax) than about their meaning (semantics) [93], this undertaking might turn out to be considerably easier than anticipated.

Outlook

We consider that important future directions for the exploitation of TM in systems biology include the following.

(i) The availability of full texts is clearly of great significance because abstracts usually lack the relevant information. This is particularly true of the values of kinetic and binding parameters.

(ii) A close integration of TM and DM techniques will benefit more widespread applications, for example, chemical structural similarity searches [31,94], the integration of medical records with genomic data and evidence from the literature for pharmaceutical applications. This will best be done in a distributed manner.

(iii) Visualization from text mining results. Current visualization methods [95] are still rather crude and there is much room for improvement here.

(iv) Better benchmarks for evaluating text mining tools that are relevant to biological needs [54].

In conclusion, although the exploitation of text mining technologies is still in its early phases, they are now becoming sufficiently mature that they can be expected to become tools in the armory of every biologist and biotechnologist.

Acknowledgements

We thank the BBSRC, EPSRC and JISC for financial support.

References

1 Ono, T. *et al.* (2001) Automated extraction of information on protein–protein interactions from the biological literature. *Bioinformatics* 12, 155–161

2 Blaschke, C. *et al.* (2002) Information extraction in molecular biology. *Brief. Bioinform.* 3, 154–165

3 Hao, Y. *et al.* (2005) Discovering patterns to extract protein–protein interactions from the literature. Part II. *Bioinformatics* 21, 3294–3300

4 Korbel, J. *et al.* (2005) Systematic association of genes to phenotypes by genome and literature mining. *PLoS Biol.* 3, e134

5 Marcotte, E.M. *et al.* (2001) Mining literature for protein–protein interactions. *Bioinformatics* 17, 359–363

6 Chun, H.W. *et al.* (2006) Extraction of gene disease relations from MEDLINE using domain dictionaries and machine learning. In *Pacific Symposium on Biocomputing 2006* (Altman, A.B. *et al.*, eds), pp. 4–15, World Scientific Publishing Co

7 Jensen, L. *et al.* (2006) Literature mining for the biologist: from information retrieval to biological discovery. *Nat. Rev. Genet.* 7, 119–129

8 Ananiadou, S. and McNaught, J., (eds) (2006) *Text Mining for Biology and Biomedicine*. Artech House
9 Hirschman, L. *et al.* (2002) Accomplishments and challenges in literature data mining for biology. *Bioinformatics* 18, 1553–1561
10 Cohen, K.B. and Hunter, L. (2004) Natural language processing and systems biology. In *Artificial Intelligence Methods and Tools for Systems Biology* (Dubitzky, W. and Azuaje, F., eds), Kluwer Academic Publishers
11 Hunter, L. and Cohen, K.B. (2006) Biomedical language processing: what's beyond PubMed? *Mol. Cell* 21, 589–594
12 Yandell, M.D. and Majoros, W.H. (2002) Genomics and natural language processing. *Nat. Rev. Genet.* 3, 601–610
13 Hearst, M. (1999) Untangling text data mining. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL 1999)*, pp. 3–10
14 Kell, D.B. (2006) Metabolomics, modelling and machine learning in systems biology: towards an understanding of the languages of cells. The 2005 Theodor Bücher Lecture. *FEBS. J.* 273, 873–894
15 Kell, D.B. (2004) Metabolomics and systems biology: making sense of the soup. *Curr. Opin. Microbiol.* 7, 296–307
16 Kell, D. and Oliver, S. (2004) Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era. *Bioessays* 26, 99–105
17 Brent, R. and Lok, L. (2005) Cell biology. A fishing buddy for hypothesis generators. *Science* 308, 504–506
18 Blagosklonny, M.V. and Pardee, A.B. (2002) Conceptual biology: unearthing the gems. *Nature* 416, 373
19 Bekhuis, T. (2006) Conceptual biology, hypothesis discovery, and text mining: Swanson's legacy. *Biomedical Digital Libraries* (http://www.bio-diglib.com/content/3/1/2)
20 Swanson, D. (1986) Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspect. Biol. Med.* 30, 7–18
21 Swanson, D.R. (1998) Migraine and magnesium: eleven neglected connections. *Perspectives in Biology and Medicine* 31, 526–557
22 Swanson, D. and Smalheiser, N. (1994) Assessing a gap in the biomedical literature: magnesium deficiency and neurologic disease. *Neurosci. Res. Commun.* 15, 1–9
23 Srinivasan, P. and Libbus, B. (2004) Mining MEDLINE for implicit links between dietary substances and diseases. *Bioinformatics* 20 (Suppl. 1), i290–i296
24 Weeber, M. *et al.* (2003) Generating hypotheses by discovering implicit associations in the literature: a case report of a search for new potential therapeutic uses for thalidomide. *J. Am. Med. Inform. Assoc.* 10, 252–259
25 Baeza-Yates, R. and Ribeiro-Neto, B. (1999) *Modern Information Retrieval*. ACM Press
26 Srinivasan, P. (2001) MeSHmap: a text mining tool for MEDLINE. *Proc. AMIA Sym* 642–646
27 Becker, K. *et al.* (2003) PubMatrix: a tool for multiplex literature mining. *BMC Bioinformatics* 4, 61
28 Ding, J. *et al.* (2005) Using the biological taxonomy to access biological literature with PathBinderH. *Bioinformatics* 21, 2560–2562
29 Perez-Iratxeta, C. *et al.* (2003) Update on XplorMed: a web server for exploring scientific literature. *Nucleic Acids Res.* 31, 3866–3868
30 Muller, H.M. *et al.* (2004) Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biol.* 2, 1984–1998
31 Klekota, J. *et al.* (2006) Query Chem: a Google-powered web search combining text and chemical structures. *Bioinformatics* 22, 1670–1673
32 Hoffman, R.V.A. (2005) Implementing the iHOP concept for navigation of biomedical literature. *Bioinformatics* 21 (Suppl. 2), ii252–ii258
33 Chang, J.T. *et al.* (2001) Including biological literature improves homology search. In *Pacific Symp. on Biocomputing*, pp. 374–383
34 MacCallum, R.M. *et al.* (2000) SAWTED: structure assignment with text description-enhanced detection of remote homologues with automated Swiss-Prot annotation comparisons. *Bioinformatics* 16, 125–129
35 Eskin, E. and Agichtein, E. (2004) Combining text mining and sequence analysis to discover protein functional regions. In *Pacific Symp. on Biocomputing*, pp. 288–299
36 Handl, J. *et al.* (2005) Computational cluster validation in post-genomic data analysis. *Bioinformatics* 21, 3201–3212
37 Blaschke, C. *et al.* (2001) Mining functional information associated with expression arrays. *Funct. Integr. Genomics* 1, 256–268
38 Glenisson, P. *et al.* (2004) TXTgate: profiling gene groups with text-based information. *Genome Biol.* 5, R43
39 Ananiadou, S. *et al.* (2004) Introduction: named entity recognition in biomedicine. *J. Biomed. Inform.* 37, 393–395
40 Bodenreider, O. (2004) The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res.* 32, 267–270
41 Tsujii, J. and Ananiadou, S. (2005) Thesaurus or logical ontology, which one do we need for text mining? *Language Resources and Evaluation* 39, 77–90
42 Spasic, I. *et al.* (2005) Text mining and ontologies in biomedicine: making sense of raw text. *Brief Bioinform.* 6, 239–251
43 Hirschman, L. *et al.* (2002) Rutabaga by any other name: extracting biological names. *J. Biomed. Inform.* 35, 247–259
44 Tsuruoka, Y. and Tsujii, J. (2004) Improving the performance of dictionary-based approaches in protein name recognition. *J. Biomed. Inform.* 37, 461–470
45 Yeganova, L. *et al.* (2004) Identification of related gene/protein names based on an HMM of name variations. *Comput. Biol. Chem.* 28, 97–107
46 Fukuda, K. *et al.* (1998) Toward information extraction: identifying protein names from biological papers. In *Proceedings of the 3rd Pacific Symposium on Biocomputing (PSB 1998)*, pp. 707–718, World Scientific
47 Narayanaswamy, M. *et al.* (2003) A biological named entity recognizer. In *Proceedings of the 8th Pacific Symposium on Biocomputing (PSB 2003), Lihue*, pp. 427–428
48 Kazama, J.i. *et al.* (2002) Tuning support vector machines for biomedical named entity recognition. In *Proceedings of the ACL Workshop on Natural Language Processing in the Biomedical Domain, Philadelphia, U. S. A.*, pp. 1–8
49 Yamamoto, K. *et al.* (2003) Protein name tagging for biomedical annotation in text. In *ACL Workshop NLP in Biomedicine* (Ananiadou, S.T., ed.), pp. 65–72
50 Zhou, G. *et al.* (2004) Recognizing names in biomedical texts: a machine learning approach. *Bioinformatics* 20, 1178–1190
51 Morgan, A.A. *et al.* (2004) Gene name identification and normalization using a model organism database. *J. Biomed. Inform.* 37, 396–410
52 Collier, N. and Takeuchi, H. (2004) Comparison of character-level and part of speech features for name recognition in biomedical texts. *J. Biomed. Inform.* 37, 423–435
53 Kim, J. *et al.* (2003) GENIA corpus – a semantically annotated corpus for bio-text mining. *Bioinformatics* 19 (Suppl. 1), i180–i182
54 Hirschman, L. and Blaschke, C. (2006) Evaluation of text mining in biology. In *Text Mining for Biology and Biomedicine* (Ananiadou, S. and McNaught, J., eds), pp. 213–245, Artech House
55 Cohen, A.M. and Hersh, W.R. (2006) The TREC 2004 genomics track categorization task: classifying full text biomedical documents. *J. Biomed. Discov. Collab.* 1, 4 (http://www.j-biomed-discovery.com/content/1/1/4)
56 Hirschman, L. *et al.* (2005) Overview of BioCreAtIvE: critical assessment of information extraction for biology. *BMC Bioinformatics* 6, S1
57 Chang, J.T. and Schutze, H. (2006) Abbreviations in biomedical text. In *Text Mining for Biology and Biomedicine* (Ananiadou, S. and McNaught, J., eds), pp. 138–165, ARTECH House
58 Wren, J.D. *et al.* (2005) Biomedical term mapping databases. *Nucleic Acids Res.* 33, D289–D293
59 Adar, E. (2004) SaRAD: a simple and robust abbreviation dictionary. *Bioinformatics* 20, 527–533
60 Schwartz, A. and Hearst, M. (2003) A simple algorithm for identifying abbreviation definitions in biomedical text. In *Proceedings of the 8th Pacific Symposium on Biocomputing (PSB 2003), Lihue, Hawaii, U. S. A.*, pp. 451–462
61 Wren, J.D. and Garner, H.R. (2002) Heuristics for identification of acronym-definition patterns within text: towards an automated construction of comprehensive acronym definition dictionaries. *Methods Inf. Med.* 41, 426–434
62 Pakhomov, S. (2002) Semi-supervised maximum entropy-based approach to acronym and abbreviation normalization in medical texts. In *Association for Computational Linguistics (ACL)*, pp. 160–167

63 Okazaki, N. and Ananiadou, S. (2006) A term recognition approach to acronym recognition. In *Proceedings of Coling/ACL Conference 2006*, pp. 643–650
64 Podowski, R. *et al.* (2004) AZuRE, a scalable system for automated term disambiguation of gene and protein names. *CSB* 415–424
65 Hatzivassiloglou, V. *et al.* (2001) Disambiguating proteins, genes and RNA in text: a machine learning approach. *Bioinformatics* 1, 97–106
66 Yu, H. and Agichtein, E. (2003) Extracting synonymous gene and protein terms from biological literature. *Bioinformatics* 19 (Suppl. 1), i340–i349
67 Huang, M. *et al.* (2004) Discovering patterns to extract protein–protein interactions from full texts. *Bioinformatics* 20, 3604–3612
68 McDonald, D. *et al.* (2004) Extracting gene pathway relations using a hybrid grammar: the Arizona relation parser. *Bioinformatics* 20, 3370–3378
69 Šaric, J.L. *et al.* (2004) Large-scale extraction of gene regulation for model organisms in an ontological context. *In Silico Biol.* 5, 21–32
70 Friedman, C.E.A. (2001) GENIES: a natural language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics* 17, S74–S82
71 Rzhetsky, A. *et al.* (2004) GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data. *J. Biomed. Inform.* 37, 43–53
72 Kim, J.J. and Park, J.C. (2004) Bioie: retargetable information extraction and ontological annotation of biological interactions from the literature. *JBCB* 3, 551–568
73 Daraselia, N. *et al.* (2004) Extracting human protein interactions from MEDLINE using a full-sentence parser. *Bioinformatics* 20, 604–611
74 Ninomiya, T. *et al.* (2006) Extremely lexiconized models for accurate and fast HPSG parsing. In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 155–163
75 Miyao, Y. *et al.* (2006) Semantic retrieval for the accurate identification of relational concepts in massive textbases. In *Proceedings of Coling/ACL Conference 2006*, pp. 1017–1024
76 Yakushiji, A. *et al.* (2001) Event extraction from biomedical papers using a full parser. In *Pacific Symposium on Biocomputing* 6, 408–419
77 Corney, D.P. *et al.* (2004) BioRAT: extracting biological information from full-length papers. *Bioinformatics* 20, 3206–3213
78 Hakenberg, J. *et al.* (2004) Finding kinetic parameters using text mining. *OMICS* 8, 131–152
79 Hoffmann, R. *et al.* (2005) Text mining for metabolic pathways, signaling cascades, and protein networks. *Sci. STKE* 283, pe21
80 Kell, D.B. and Knowles, J.D. (2006) The role of modeling in systems biology. In *System Modeling in Cellular Biology: from Concepts to Nuts and Bolts* (Szallasi, Z. *et al.*, eds), pp. 3–18, MIT Press
81 Schilling, C.H. *et al.* (2002) Genome-scale metabolic model of *Helicobacter pylori* 26695. *J. Bacteriol.* 184, 4582–4593
82 Borodina, I. *et al.* (2005) Genome-scale analysis of *Streptomyces coelicolor* A3(2) metabolism. *Genome Res.* 15, 820–829
83 Herrgård, M.J. *et al.* (2006) Integrated analysis of regulatory and metabolic networks reveals novel regulatory mechanisms in *Saccharomyces cerevisiae*. *Genome Res.* 16, 627–635
84 Arakawa, K. *et al.* (2006) GEM System: automatic prototyping of cell-wide metabolic pathway models from genomes. *BMC Bioinformatics* 7, 168
85 Mendes, P. and Kell, D. (1998) Non-linear optimization of biochemical pathways: applications to metabolic engineering and parameter estimation. *Bioinformatics* 14, 869–883
86 Hucka, M. *et al.* (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* 19, 524–531
87 Luciano, J.S. (2005) PAX of mind for pathway researchers. *Drug Discov. Today* 10, 937–942
88 Oinn, T. *et al.* (2004) Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics* 20, 3045–3054
89 Oinn, T. *et al.* (2006) Taverna/myGrid: aligning a workflow system with the life sciences community. In *Workflows for eScience*, pp. 299–318, Springer
90 Wilkinson, M.D. and Links, M. (2002) BioMOBY: an open source biological web services proposal. *Brief. Bioinform.* 3, 331–341
91 Lu, Q. *et al.* (2005) KDE Bioscience: platform for bioinformatics analysis workflows. *J. Biomed. Inform.* 39, 440–450
92 Curcin, V. *et al.* (2005) Web services in the life sciences. *Drug Discov. Today* 10, 865–871
93 Wilkinson, M. *et al.* (2005) BioMOBY successfully integrates distributed heterogeneous bioinformatics web services. The PlaNet exemplar case. *Plant Physiol.* 138, 5–17
94 Singh, S.B. *et al.* (2003) Text influenced molecular indexing (TIMI): a literature database mining approach that handles text and chemistry. *J. Chem. Inf. Comput. Sci.* 43, 743–752
95 Tao, Y. *et al.* (2005) Visualizing information across multidimensional post-genomic structured and textual databases. *Bioinformatics* 21, 1659–1667