Review

Cell PRESS

Event extraction for systems biology by text mining the literature

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

[1] School of Computer Science, University of Manchester, Oxford Road, Manchester M13 9PL, UK
[2] National Centre for Text Mining, Manchester Interdisciplinary Biocentre, University of Manchester, 131 Princess Street, Manchester M1 7DN, UK
[3] Manchester Interdisciplinary Biocentre, University of Manchester,131 Princess Street, Manchester M1 7DN, UK
[4] Department of Computer Science, University of Tokyo, Hongo 7-3-1, Bunkyo-ku, Tokyo 113-0033, Japan
[5] School of Chemistry, University of Manchester, 131 Princess Street, Manchester M1 7DN, UK

Systems biology recognizes in particular the importance of interactions between biological components and the consequences of these interactions. Such interactions and their downstream effects are known as events. To computationally mine the literature for such events, text mining methods that can detect, extract and annotate them are required. This review summarizes the methods that are currently available, with a specific focus on protein–protein interactions and pathway or network reconstruction. The approaches described will be of considerable value in associating particular pathways and their components with higher-order physiological properties, including disease states.

Background

Biological systems consist of entities and the relationships between them in terms of how they interact and the downstream effects of such interactions. Classically, molecular biology focused on the entities involved, but now, in the era of systems biology [1–3], we are beginning to appreciate the magnitude of these interactions. Although text mining (TM) has already been useful in systems biology applications through identification [4], normalization [5,6] and disambiguation (see Glossary) of key entities [7,8] such as genes, proteins and enzymes, its true capabilities are only now beginning to be realized through automatic recognition of relevant biological events and relations from the literature. To help elucidate the roles played by biomolecules in important biological processes – and, in turn, in phenotypic outcomes, such as disease and the manifestation of agricultural or biotechnological traits – TM systems have to tackle the complex problem of extracting and identifying the context and type of such relationships. Without explicit recognition of the underlying mechanisms of biological processes in terms of the involvement of specific entities, TM results are either too noisy or too restricted to be useful. TM techniques have to recognize diverse surface forms in text that describe the same biological processes and identify which biological entities are involved. Therefore, more advanced analytical methods are necessary, namely methods that undertake deeper semantic analysis. To achieve this aim, text miners have developed techniques that automatically extract biological events pertaining to processes such as protein–protein interactions and protein–disease associations from the literature.

Automatic event extraction has a broad range of applications in systems biology, ranging from support for the creation and annotation of pathways to automatic population or enrichment of databases. Event extraction systems can be trained to recognize a wide range of activities, including protein–protein interactions, pathway enrichment and construction, gene regulatory events, and metabolic or signaling reactions. The purpose of this review is to describe the resources and techniques involved in developing tools for event extraction from the literature for systems biology applications. To develop event extraction systems, annotated corpora are needed. In this review, we examine the features of a number of event annotation corpora used by TM systems for training and development. We also examine a number of approaches to event extraction, ranging in sophistication from pattern matching to systems that include rich linguistic features based on full parsing.

From entities to relations and events

Text mining in biology [9–11] has focused mainly on recognizing biologically relevant entities, locating synonyms in text (including acronyms and other term variants) [5,7,12], and finally mapping them to unique identifiers in curated databases (normalization) such as UniProt (www.uniprot.org/) and Entrez Gene (www.ncbi.nlm.nih.gov/gene/) [10]. Named entity recognition (NER) [13] and normalization [14] have been helpful for increasing the specificity of document searches in TM systems such as KLEIO (www.nactem.ac.uk/software/kleio/) and for significantly reducing errors compared with simple keyword-based retrieval. Other search systems, such as FACTA [15], use co-occurrence statistics for normalized names in text to enhance the discovery of hidden associations among entities.

Mere textual co-occurrence of entities, however, does not necessarily indicate meaningful relationships. It has been reported that only 30% of protein pairs co-occurring in the same sentences have an actual interaction [16]; furthermore, a considerable amount of experimental noise is present [17] that is best addressed via a methodology that uses multiple assays to increase confidence. In general, ensembles of different techniques are much more effective

Corresponding author: Ananiadou, S. (sophia.ananiadou@manchester.ac.uk).

0167-7799/$ – see front matter doi:10.1016/j.tibtech.2010.04.005 Trends in Biotechnology xx (2010) 381–390

Glossary

Co-occurrence: the occurrence of terms together in text can serve as an indication of a relationship between them. For example, the co-occurrence of two protein names within a single sentence can suggest an interaction between the proteins. Mutual information (MI) can be further used to examine the strength of the suggested relation. MI compares the joint probability of two items occurring $[p(x,y)]$ with the probability of independent occurrence $[p(x)\times p(y)]$. The higher the MI value, the greater is the amount of shared information; in other words, the higher the MI value, the greater is the confidence in hypothesizing that the occurrence of one determines or predicts the occurrence of the other.

Disambiguation: natural language text frequently contains words that have more than one possible interpretation. Disambiguation tasks involve selection of the correct interpretation among ambiguous alternatives, typically drawing on information from the context of the ambiguous expression.

Information extraction: component of text mining that takes natural language text from a document source, extracts essential facts about one or more predefined fact types, and represents each fact as a template with slots filled on the basis of what is found from the text. To this end, various techniques are deployed to recognize entities and relations, which are then used to construct fact templates.

Metadata: 'data about data' (i.e. structured information regarding another piece of information). In a search context, metadata typically refer to keywords that identify concepts that are important for indexing of documents.

Named entity recognition (NER): task of automatically identifying mentioned names that refer to types of entities, such as genes and proteins, in text.

Normalization: in text, a particular concept can be denoted by various surface realizations, which are called term variants. For example, TIF2, TIF-2, transcription intermediary factor-2 and transcriptional intermediate factor 2 all denote the same concept. Usually, one of these term variants is considered as the preferred term. Normalization refers to the automated process by which all term variants are grouped together into an equivalent class.

Ontologies: conceptual models used to support consistent and unambiguous knowledge sharing and to provide a framework for knowledge integration. For example, a biomolecular ontology might define concepts such as organic compounds, proteins and DNA, and organize them to specify that the latter two are subtypes of the first. In addition to organizing concepts in 'is a' hierarchies, ontologies can specify other general relations such as 'part of' and 'located in', as well as domain-specific relations, such as 'translated into' and 'transcribed into'.

Ordered pair: for an ordered pair, (a,b) differs from (b,a). An ordered entity pair representation can be used to model directed relations, such as phosphorylation, in which the roles of the entities are different, whereas unordered pairs are appropriate for simple symmetric relations, such as binding.

Parsing: also referred to as syntactic analysis, parsing is the process of determining the syntactic structure of sentences. There are various approaches to parsing. One major division is between constituency (or phrase structure) and dependency approaches. The former can involve the building of increasing levels of hierarchy from the basic constituents (nouns, verbs, adjectives) to more complex constituents (noun phrases, verb phrases, sentences) in syntactic representation; the latter establishes relations (or dependencies) between the organizing verb and its dependent arguments. Syntactic analysis can also be categorized into full (deep) and partial (shallow) parsing, depending on whether the entire sentence structure or only part is resolved, such as the major top-level phrases. Deep parsing provides relationships not explicitly stated among words in a sentence; this is why it is commonly used for event extraction. For example, in the sentence 'p53 is shown to activate transcription', deep parsing encodes this information as follows: 'p53' is the subject of the predicate 'to activate' and 'transcription' is an object. Deep parsing often uses predicate argument structures.

Predicate argument structure: a normalized form representing syntactic relations, as in the example 'ENTITY1 INHIBITS ENTITY2'. Here, the formal symbol INHIBITS is the predicate, which contains the main meaning of the predicate argument structure, and the formal symbols ENTITY1 and ENTITY2 are its arguments, carrying information about the participants described by the predicate.

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

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

Semantic typing: assignment of a type with specified meaning to identify the category of an item. The definitions of types, such as 'protein' and 'regulation', would typically be defined in an appropriate ontology.

Tagging: in natural language processing, tagging is used to refer to tasks like part-of-speech tagging in which tags or labels representing grammatical parts-of-speech are assigned to a sequence of words or word-like units, such as 'monocyte:NN, noun'. Other tasks add labels as tagging; for example, NER can be performed by marking each word with an additional label (e.g. monocyte: cell-line).

than a single technique. In order to mine real interacting pairs of proteins from the literature, 70% of the noise (i.e. false positives) must be filtered out.

Text retrieval systems, such as PubMed, permit only Boolean combinations of names. However, for semantic and focused queries to retrieve not just articles, but also pertinent biological facts from the literature, it is essential to recognize named entities (e.g. protein, metabolite) as well as biological events (e.g. activate, phosphorylate). This leads to techniques such as relation recognition (RR) and event recognition (ER), which are active areas of research for the biomedical TM community [18,19] and are part of information extraction (IE).

Relations and events

Events are characterized by verbs (e.g. transcribe, regulate) or nominalized verbs (e.g. transcription, regulation). In the sentence 'In *E. coli*, *glnAP2* may be activated by *NifA*', an event is specified by the verb 'activated', and the event participants or arguments are defined as 'In *E. coli*', '*glnAP2*' and '*NifA*'. Semantic roles and concept labels or tags are also assigned to these arguments. For example, the verb 'activated' expects to have: (i) a first argument, such as *NifA*: Activator, which acts as an agent or cause (a semantic role); (ii) a second argument, *glnAP2*: Gene, which acts as a theme; and (iii) the argument In *E. coli*: Wild_Type_Bacteria, expressing a location**.** Concept labels are based on existing ontologies.

There are several definitions of relations and events in biotext mining, such as genotype–phenotype associations [20,21], disease–gene associations [16,22] and regulatory events [23]. In this review, we discuss events and relations that are mainly expressed within the boundaries of a single sentence and not across sentences or papers. The extraction of genotype–phenotype associations typically requires information scattered around several sentences or fragments of sentences in different articles, which is beyond the scope of this review. Entity names appear as continuous spans in text and are mapped to identifiers during the normalization stage. Objects, such as relations and events, generally appear as discontinuous spans in text, and have respective internal structures: that is, a relation or an event generally involves more than one entity, and the entities involved play distinct roles in events. In some cases, events or relations are recursively embedded; for example, a simple event, protein binding, can play a role in a larger, composite event such as regulation.

Once discovered, events or relations need to be represented in a way that is suitable for computational manipulation during subsequent processing. A relation (Figure 1a–c) is typically represented as a pair of entities, linked by an arc that is either directed or undirected. The arc is given a label usually corresponding to a semantic type (e.g. an ontological class). Both participating entities must be specified and their roles are fixed in advance; for example, in a regulation relation, the first entity is always the regulator (agent or cause) and the second is the target (theme). When the roles are the same, the relation is symmetric, or undirected. Event representations (Figure 1d–f) capture the association of multiple partici-

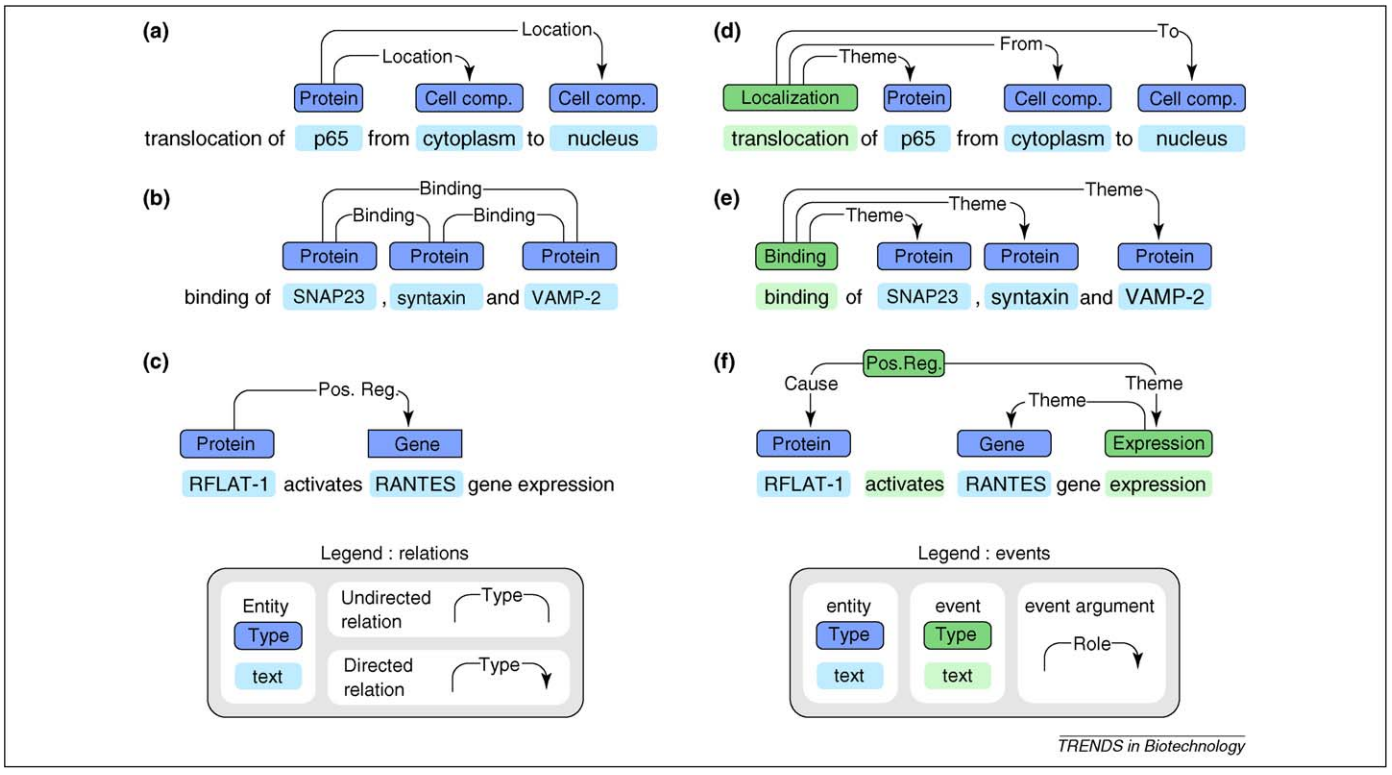

Figure 1. Relation (left) and event (right) representations for (**a**,**d**) localization, (**b**,**e**) binding and (**c**,**f**) regulation. Relations between entities can be directed (arrow) or undirected (no arrow) and are labeled with a semantic type. Event representations can capture the association of multiple participants in different roles, are associated with specific expressions in the text and can participate in other events.

pants of varying numbers and with varying semantic roles [23] that are, in principle, determined by the needs of the domain. For example, in Figure 1d, the localization event captures the identification of the localized entity (p65), the source (cytoplasm) and the destination (nucleus). There is a difference between an instance of an event in a text and the abstract specification of an event as contained in a TM system. An abstract specification could mention many potential participants or arguments; the specification for a localization event might mention a cause. Some participants are typically obligatory (an event would not make sense without these), whereas others are optional. In other words, the semantic needs of an event specification are filled according to what might be found in the text, and not all of its needs might be met in any one instance. However, if the obligatory needs are not met, an event specification fails to apply. Moreover, in Figure 1b, binding is an undirected relation, implying a model in which X binding to Y and Y binding to X are taken to be equivalent. By contrast, regulation (Figure 1c) is a directed relation because the roles of the participants are different.

It is important to note that relations and events can function as participants in other events, thus allowing the construction of complex conceptual networks. For example, consider the following sentence: '*glnAp2* expression is affected by different carbon sources.' In this sentence, there are two events, one anchored to 'expression' (whose object is *glnAp2*), and the other anchored to 'affected'. The expression event itself is affected. During processing, further information can be discovered and attached to events concerning such aspects as negation, contradiction, speculation, probability and possibility, so that various shades of meaning can be distinguished.

To effectively extract events, analysis of sentence structure is necessary. Event extraction can benefit in particular from the use of semantic processing or deep parsing techniques [24,25] that analyze both the syntactic and semantic structure of texts. The output of deep parsing includes predicate–argument relations among words. These relations are especially useful for event extraction where the meaning of a sentence plays a central role [10]. Some advantages of deep parsers are that surface variations expressing the same information are captured (e.g. Entity1 activates Entity2 and Entity2 is activated by Entity1) and that all such examples can be compacted into a single predicate–argument structure: {activate ARG1 Entity1 (semantic subject) ARG2 Entity2 (semantic object)}. Predicate argument structures have been used successfully as a representation to extract protein–protein interactions [24].

Applications in systems biology

Searching

Once biological events have been discovered and extracted, the question arises as to how to use this information. Metadata are used to index digital documents for retrieval purposes. Such metadata are typically bibliographical in nature (e.g. author, volume, page numbers, keywords, etc.) [26,27]. However, in principle, metadata do not need to be confined to these traditional types [28]. Thus, it is also possible to store and annotate extracted entities, events and relations as document metadata. This immediately offers the possibility of a much more sophisticated semantic

search in comparison to conventional search engines. Thus, a user can conduct a type of query-by-example semantic search by partially or completely specifying elements of an event type of interest. Given the enormous size and richness of the literature [26], researchers are interested in obtaining high retrieval accuracy.

The application of text mining techniques to sophisticated metadata from text can be adapted to specific user needs. Systems of interest include iHOP [29], CiteXplore (www.ebi.ac.uk/citexplore/), GoWeb [30] and MedlineRanker [31]. Although these systems include more extracted information (i.e. proteins and protein–protein interactions) than conventional search systems, they do not use events for searching. An example of a search system that is, in fact, based on events is MEDIE [32]. MEDIE is an intelligent search engine designed to retrieve biomedical events from the whole of MEDLINE, relying on a sophisticated indexing system derived from multiple syntactic and semantic analyses. Similar to other semantic metadata-based systems, MEDIE uses NER [33] to automatically identify semantic types, such as genes, proteins and diseases. Entities are then normalized [6] into corresponding IDs and linked to databases, such as UniProt, Entrez Gene and Unified Medical Language System (UMLS: www.nlm.nih.gov/research/umls/), via an ontological database, GENA (http://gena.ontology.ims.u-tokyo.ac.jp:8081/search/servlet/gena). Such databases provide mappings between entities in text and existing biological databases, and facilitate normalization of the synonymy of entities as they occur in text.

MEDIE is also based on a deep parsing technology [24] that produces predicate argument structures (PAS). PAS are useful because they can represent biological events and relations in an abstract manner [32]. When searching with MEDIE, the user fills in a simple form to specify participants in some event of interest. For example, the user might be interested in an activation event and would thus specify the event verb activate, then further specify, for instance, the subject TNFα and the object NF-κB. MEDIE then extracts sentences containing instances of activation events, including events expressed with other verbs that denote the same event type, such as induce and trigger (through reference to the Gene Ontology: www.geneontology.org). The user does not need to specify all fields of the query form; if the object referred to above is not specified, the query then amounts to asking, 'What does TNFα activate?' Two advantages of MEDIE over a conventional search engine are that it returns precise facts rather than entire documents to read and enables the user to perform a semantic search grounded in the user's domain of interest rather than a keyword-based search. In a semantic search, the event specification ties desired concepts together; in a keyword-based search, it is very difficult for a search engine to guess the intended relationships among the query terms.

Protein–protein interaction and curation

Although narrower in focus than event recognition, protein–protein interaction (PPI) extraction has been addressed in several studies during the last decade [34–36]. Compared with the extraction of events, PPI extraction amounts to relation recognition, which only classifies co-occurring pairs of proteins into two subsets, namely a set of interacting pairs and a set of non-interacting pairs. PPI extraction was targeted in the BioCreative II [37] community challenge. The focus on PPI extraction is motivated largely by the needs of database curation. A major obstacle to curation is the high number of false positives found in high-throughput screens [17]. Interaction databases, such as DIP [38] and IntAct [39], are primarily populated from literature analyses by human curators, and automatic methods supporting this extraction can considerably reduce curation costs [40]. Recent comparative evaluation [24] of several state-of-the-art TM methods demonstrated that deep parsing technology outperformed other techniques, such as dependency parsing and phrase structure parsing, in terms of accuracy (precision, recall and F-scores) for PPI extraction. Thus, deep parsing technology is also useful for relation extraction.

In addition to PPI, there is considerable interest in the computational analysis of small molecules and the proteins that they can bind. Although such chemoinformatic approaches [41] are much less frequently considered than PPI owing to the considerably smaller size of the community working on such approaches, it has been shown that they have considerable power [42,43].

Linking pathways to literature

Pathways and networks are at the core of systems biology and are becoming increasingly important for biomedical research because they represent collectively attested interpretations of a large number of facts scattered throughout the literature. In addition, models encoded in the Systems Biology Markup Language (SBML: http://sbml.org/) [44] can be used as input to a variety of tools such as CellDesigner (www.celldesigner.org/) [45] and Copasi [46]. Nevertheless, models are still built manually [47] and biologists read a large number of articles to construct these pathway models; thus, they would benefit considerably from the use of TM tools not only to support their maintenance [48], but also to provide direct links from the models to literature evidence [49]. Furthermore, such tools will become essential to keep existing models up to date by revising them according to newly published articles.

Two of the most important applications of event recognition to systems biology are linking pathways to literature evidence and aiding pathway construction and enrichment [50]. In the past, studies involving TM technology for network construction have focused on extracting binary interactions between proteins or genes [51–53]. Although the resultant networks seem to be pathways, they do not represent any coherent interpretations of the reported facts [50]. Mapping between the results of automatically constructed networks and pathways requires a deeper analysis that emulates the interpretations of biologists, including inferences based on biological background knowledge. Thus, providing evidence from the literature to pathway representations requires the extraction not only of events, but also of the relevant context around them.

A core component of pathway enrichment and construction is the integration of TM technology with pathway

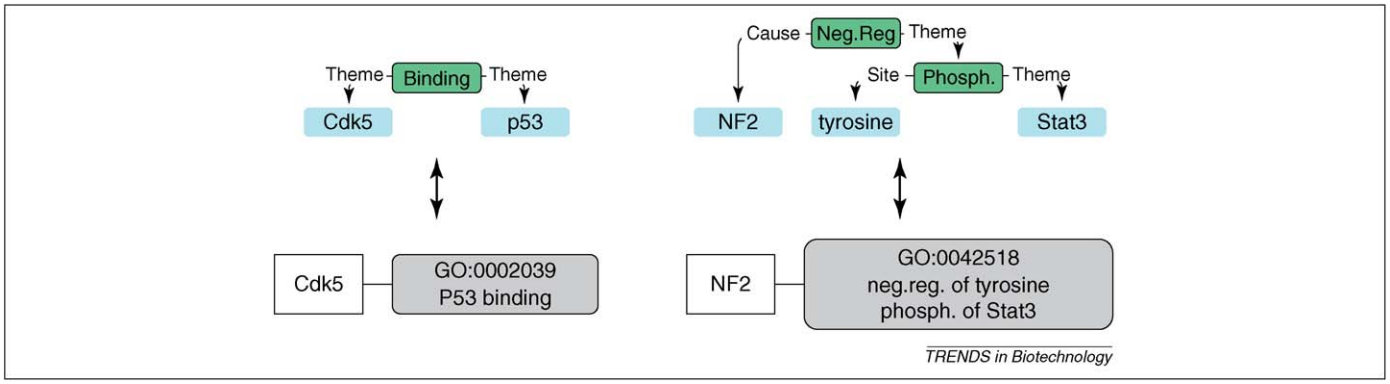

Figure 2. Candidate GO annotations. Recognition of events (top row) can help to derive GO annotations (bottom row) automatically. The diagrammatic representation of event annotation is explained in Figure 1.

visualization software (e.g. CellDesigner) and annotation tools. The maintenance of constructed pathways requires constant monitoring of recent publications. PathText (www.pathtext.org) is a TM-based integrated environment for biological pathway visualization. Unlike existing pathway building platforms, such as WikiPathways [54] and PathCase [55], PathText brings together the strengths of different TM tools, including advanced searches based on event extraction. PathText links several text mining systems (FACTA, KLEIO and MEDIE) pathway visualizations using Payao (CellDesigner) and annotation tools that facilitate curation, thus allowing snippets from publications to be linked to nodes in the pathway or network interactively, as well as on a community-wide basis.

GO term annotation

The Gene Ontology (GO) provides structured, controlled vocabularies of terms describing gene and gene product characteristics. GO is used as a reference in several ongoing efforts to annotate genes and gene products, with consistent and easily searchable terms identifying their characteristics. Although automated methods have been applied to the task, GO annotations of the highest relevance and quality are achieved through manual annotation by curators reading full-text papers. However, the creation of manual GO annotations is an enormous effort undertaken at considerable cost, and the ~500 000 manual annotations that have been created to date cover <3% of UniProt [56]. To reduce GO annotation costs, significant effort has been focused on the development of systems for automatic annotation and annotator support. Figure 2 shows how event annotation could assist in the automatic derivation of GO annotations. Notably, GO annotation was considered as a task in the first BioCreative challenge [57], which found that significant obstacles still remain for systems to achieve practically applicable levels of utility for this task.

Unfortunately, in annotation tasks there are other considerations besides cost and coverage. There is, in particular, a human factor involved. Humans typically arrive at different judgments and decisions when asked to tackle the same task. Thus, even when human annotators agree on the GO codes to be assigned, they tend to indicate different parts of the same articles as evidence of their judgments. Such manifold difficulties encountered in GO assignment indicate that, unlike other tasks, this task cannot be resolved within the IE framework only. Instead, it requires a human understanding of the text, which combines information extracted from text with domain background knowledge to make inferences. An important feature of GO annotations is the use of evidence codes to indicate the kinds of reasoning and evidence that underpin the assignments. Thus, the inferred existence of an enzyme in a particular cell might be based on sequence homologies in the genome, on measurement of the expression of the protein itself, or detection of the actual enzyme reactions of interest; clearly, these forms of evidence differ in their significance.

However, if we now wish to leverage inference and reasoning automatically in an IE system rather than relying on human understanding, we find that the relatively simple relation representation dominant in IE is insufficient to support reasoning. It has been argued [18] that a structured and richly typed event representation is required to facilitate such inferences. These considerations led to the new design of the GENIA corpus event annotation, which draws its event types from GO [58]. Thus, we review the important area of corpus resources to support event extraction.

Resources for event extraction

Corpus annotation

High-quality annotated corpora, or collections of texts, are indispensable for systematic development of IE rules and machine learning methods. Corpora are also used for training and evaluation of text mining systems. Compared with the annotation of named entities, event annotation is much more complicated. Events are usually expressed as discontinuous spans in text (Figure 1). Furthermore, event annotations are application-oriented. For example, researchers studying metabolic pathways are concerned with extraction of enzymes that might be involved in reactions defined in terms of EC numbers, whereas groups engaged in protein functional annotation or regulatory relations among genes are interested in a subset of event classes in GO. With all this activity by different groups with different interests, it is important to investigate ways to reduce the cost of producing annotated corpora using techniques such as accelerated annotation or active learning [59].

PPI corpora

PPI corpora remain highly relevant to event-based IE approaches, in that they share many of the same extraction targets. Furthermore, their representation closely corresponds to that required for PPI database curation and a large body of domain IE work has used these resources for evaluation, so these corpora offer ways to directly assess the relative merits of relation- and event-based IE approaches for a practically relevant task. A number of available PPI corpora are listed at http://mars.cs.utu.fi/PPICorpora/.

BioInfer event corpus

BioInfer was the first publicly available biomedical annotated corpus to incorporate events [60]. The corpus contains 1100 sentences in which the primary annotation types identify entities, events and sentence syntax. BioInfer annotation covers both events (termed causal relationships in the corpus ontology) and static (non-causal) relations using a single predicate formalism. The event annotation is focused on gene and gene product entities, which are annotated as event participants in a cause-, target- or theme-type role. In BioInfer, all explicitly stated events are marked, regardless of their form of expression, (e.g. multi-word expressions not involving verbs can also be annotated as expressing events). An example of the type of annotation performed in the BioInfer corpus is shown in Figure 3a.

GENIA event corpus

The GENIA annotation corpus [58] is one of the most widely used resources in biomedical text mining. Besides events, GENIA annotations include 40 classes of biomedical terms [61], occurrences of gene and gene product names, and sentence syntax. The GENIA event corpus comprises 1000 annotated PubMed abstracts containing over 9000 sentences. Like BioInfer, GENIA marks up all stated events regardless of their form of expression. However, GENIA events allow a wider range of participants than BioInfer does, and these are marked with more role types. This facilitates specification of site arguments for events such as binding, a set of location specifications and time expressions (Figure 3b). The annotation further marks up both negated and speculated statements, identifying three different certainty levels for evidence.

Gene Regulation Event Corpus

The Gene Regulation Event Corpus (GREC: www.nactem.ac.uk/GREC/) [23] consists of 240 MEDLINE abstracts in which sentence-bound events relating to gene regulation and expression have been annotated. Events are centered on both verbs and nominalized verbs. For each event instance, all participants (arguments) in the same sentence are identified and assigned a semantic role from a rich set of 13 roles tailored to biomedical research articles, together with a biological concept type linked to the Gene Regulation Ontology (www.obofoundry.org/). GREC is designed to facilitate the training and development of IE systems and resources in the biomedical domain. GREC is unique in that it annotates not only core relationships between entities, but also a range of other important

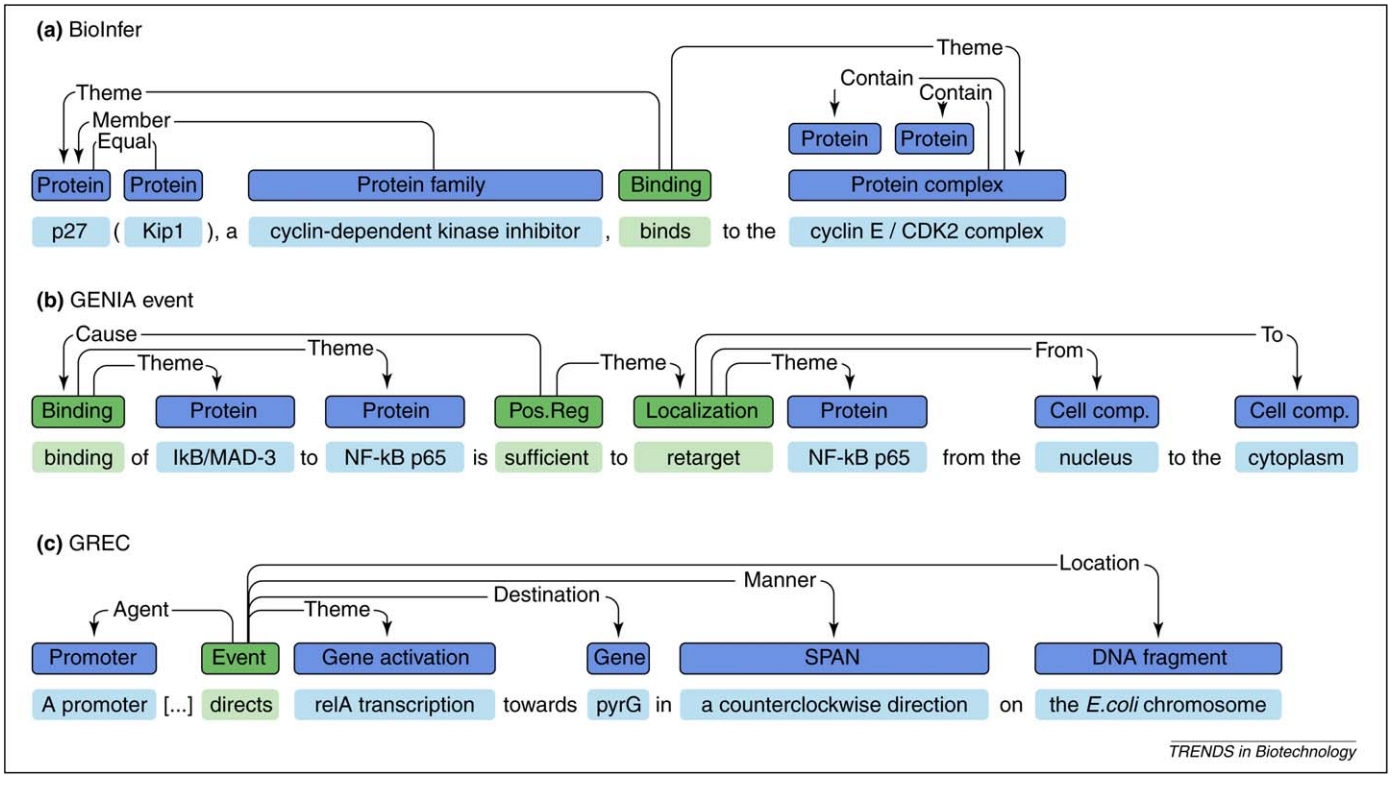

Figure 3. Example event annotations. (**a**) Example from BioInfer showing an event and relations. In addition to events (i.e. binding), the corpus annotation captures static relations (grey text), such as equality of references identifying the same entity, membership of an entity in a family or group, and containment of an entity as part of another. (Type and role names have been edited for consistency.) (**b**) Example from GENIA showing three events: binding, localization and positive regulation. The first event represents binding of the entities IκB–MAD-3 and NF-κB–p65 and the resulting localization event. The causal relation is captured as the third event, positive regulation. (**c**) Example from GREC showing a single event that is centered on the verb 'directs'. All arguments of the event are assigned a semantic role from a set of 13 possible roles. Biological concept types are assigned to arguments where appropriate, whereas the label 'SPAN' is assigned to any arguments that do not correspond to biological concepts.

Table 1. Selected statistics for BioInfer, GENIA and GREC event annotated corpora

	Annotation type			Number of annotations		
	Entity	Event	Role	Sentences	Entities	Events
BioInfer event	110	68	3	1100	6349	1461
GENIA event	40	40	10	9372	45224	36114
GREC	64	19	13	2400	5393	3067

details about these relationships, including location, time, manner and environmental conditions (Figure 3c).

Although GREC is a relatively small annotated corpus, a recent study [62] has shown that combining smaller, richly annotated corpora with larger corpora that are slightly poorer in information content can help to improve the performance of event extraction systems or semantic role labelers. Although the benefits of combining disparate sources in machine learning are well known, this idea is especially attractive, given that the production of large, richly annotated corpora can be very time-consuming.

Although different in size and number and the nature of semantic entity types, semantic roles, and event types (Table 1), these annotated corpora are crucial for the training of event extraction systems. These differences are partly task-dependent, in that certain tasks require annotation of certain types of entities and events. Furthermore, the availability of trained annotators or the need to train annotators, the time available for annotation within a larger project, the text type involved (full papers, abstracts, patents) and the complexity of a particular annotation task can all have an impact on the construction of event-annotated corpora. From the standpoint of general reuse of such corpora, however, any single corpus can be deemed inadequate for some new task; these resources are nevertheless valuable owing to the rich linguistic and biologically pertinent information with which they are annotated. The annotations establish explicit links between the realms of text and biology. Given the value that these corpora represent, every effort is made to fully exploit them for new tasks. Thus, in recent years, the field has moved from the use of a single corpus to a combination of available corpora [24].

Approaches to event extraction

There are several approaches to extraction of events from biomedical texts. These vary in terms of the level of linguistic representations incorporated (pattern matching versus full parsing), the number of lexical and ontological resources used, the analysis adopted (rule-based versus machine learning) and the domain specificity (sublanguage-driven versus general language approaches).

Pattern matching approaches [63–65] can vary from simple approaches, such as sentence extraction (without attempting any syntactic generalization), to slightly more sophisticated techniques, such as using part-of-speech tagging and regular expressions, to construct IE templates. These systems extract sentences containing matched patterns and although some systems include linguistic information, there is limited generalization and several patterns are required, and they are thus non-transferrable to other user cases [66].

Certain systems that have been described [67] opt for a more syntactic approach for extraction of pathway relations. This approach consists primarily of the extraction of relations based on 'triples' (subject, object and verb constructs) obtained from surface-oriented linguistic analysis [10]. However, such simple linguistic analysis cannot capture the information structure of a sublanguage [68], which demands richer relations expressing conditions, manner, destination, etc. Sublanguage-driven IE systems rely on the notion that the informational structure [69] of the domain imposes constraints at all linguistic levels (lexical, syntactic, semantic, discourse), which can be exploited to produce accurate systems. A system that takes a strong sublanguage view of the extraction of biomolecular interactions for signal transduction and biochemical pathways is GENIES [70]. GENIES is a rule-based system that uses a full parsing strategy and filters out ambiguities due to the informational constraints imposed on verbs and their arguments. Such systems require sublanguage grammars and dictionaries, which describe the constraints of the domain. These are typically expensive to build.

To extract events from sublanguages, such as biology, deep linguistic knowledge and ontological information are required. An ontology-driven system that targets events is GenIE [71], which extracts information on biochemical pathways, sequence structures and functions of genomes and proteins. GenIE relies on a deep semantic representation formalism and full linguistic analysis. Importantly, it also requires an ontology of biochemical events. Verbs are clustered into ontological classes, which are then assigned appropriate semantics.

Full parsing methods were applied relatively early in biomedical IE studies [72], but the need for full textual analysis for tasks such as PPI has not been uncontested. However, recent evaluations support the value of full parsing [24], and recently proposed PPI extraction methods that perform competitively with state-of-the-art techniques on the AIMed corpus [73] build on full parsing techniques [74,75]. Importantly, in the recent BioNLP'09 event extraction task, full parsing was applied by the majority of systems and by all systems ranking in the top 50% of the primary task of event extraction, lending strong support to its value in biomedical IE.

BioNLP shared task on event extraction

Shared tasks ('bake-offs'), in which teams from the community compete to analyze the same data within a common evaluation framework, have played a significant role in biomedical IE. They provide standard development and evaluation benchmarks, focusing the attention of the research community on timely issues and acting as a driver for the specification of new tasks and challenges. Task definitions in shared tasks can be seen in part as progressing from basic foundational tasks, such as search and entity detection, to higher IE targets, such as event extraction. Examples of shared tasks include the TREC genomics

Table 2. Event types and their arguments

Event type	Primary arguments[a]	Secondary arguments[c]
Gene expression	Theme (P)	
Transcription	Theme (P)	
Protein catabolism	Theme (P)	
Phosphorylation	Theme (P)	Site
Localization	Theme (P)	AtLoc, ToLoc
Binding	Theme (P)+[b]	Site+
Regulation	Theme (P/Ev), cause (P/Ev)	Site, CSite
Positive regulation	Theme (P/Ev), cause (P/Ev)	Site, CSite
Negative regulation	Theme (P/Ev), cause (P/Ev)	Site, CSite

[a]For each primary (obligatory) event argument, the role of the argument (theme, cause) is shown, with the possible argument filler type shown in parentheses (P, protein; Ev, event).
[b]Binding events can take an arbitrary number (+) of proteins as primary arguments, which form protein complexes.
[c]Secondary arguments are optional, in that they provide extra details about the event that might only be present in certain events of a given type. Site: specific domains or regions that correspond to the theme of an event; AtLoc: the source of an event; ToLoc: the goal or destination of an event; CSite: specific domains or regions that correspond to the theme of an event.

task, the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA 2004: www-tsujii.is.s.u-tokyo.ac.jp/GENIA/ERtask/report.html), the Learning Language in Logic (LLL) challenge (http://genome.jouy.inra.fr/texte/) and BioCreative [76]. Although TREC and JNLPBA focused on named entities and information retrieval, LLL and BioCreative revolved around IE and relations between biomolecules. A step further from these shared tasks, which are crucial for evaluating TM systems for biology applications, is the recent BioNLP'09 task.

The BioNLP'09 shared task focused on event extraction based on protein biology event types. Semantically, these event types (Table 2) share proteins as the theme. The first three types concern protein production and breakdown. Phosphorylation is a representative protein-modification event, and localization and binding are representative fundamental molecular events. Regulation (including its sub-types, positive and negative, as defined for this particular shared task) represents regulatory events and causal relations. The precise definitions of these event types are given in terms of the corresponding GO classes.

Table 2 shows the primary (obligatory) and secondary (optional) arguments for each event type. All events have themes (i.e. entities that are affected by the event) as a primary argument, which is essential for identification of the event. Except for the types corresponding to three regulation events and the binding event, each event type has one primary argument, typed as protein. The regulation events have two primary arguments, which are typed as either protein or event. One of the primary arguments of the regulation events is stated as the cause of the regulation. The binding event type is more complex in that it takes an arbitrary number (one or more) of proteins as primary arguments, which form protein complexes. For some event types, other arguments providing details of the events are also defined as secondary arguments. As an example, the localization event takes the source (AtLoc) and goal (or destination) (ToLoc) locations as its secondary arguments. If we consider the results of the 26 participating groups, we find that simple events corresponding to those shown in Table 2 that take only a theme as their primary argument (e.g. gene expression and phosphorylation) can be extracted by current state-of-the-art methods. However, performance is less than 50% for binding events, which typically involve several participants, and weaker still for the more complex regulation events, which involve other events as arguments.

Concluding remarks

PubMed is increasing at the rate of approximately two papers per minute [26] and it is impossible for any individual to comprehensively read all of the literature related to his or her field. As we move towards more integrative systems biology [77,78], automated reasoning will be required [79]. A goal of TM is to enhance our ability to extract information from the growing corpus of literature to make the process of synthesizing this information more efficient, manageable, comprehensive and precise. Furthermore, as data extracted from the literature become more structured via automated processing, they can be more readily overlaid with other information (e.g. molecular data from experimental platforms) [27,28]. Text mining is increasingly used to support knowledge discovery [80] and hypothesis generation [81], and to make sense out of the mass of biological literature [82–84].

The TM techniques presented in this review recognize diverse surface forms in text describing the same biological processes and identify which biological entities are involved in them. Thus, the use of techniques that undertake a deeper semantic analysis, such as event extraction, is necessary for more advanced systems biology applications. The recent BioNLP shared task showed that the effectiveness and precision of event extraction can be task-dependent: the performance of systems depends on the complexity of the events to be recognized and on user needs. To improve the performance of systems for event extraction, several TM components, such as parsers, named entity recognizers, and taggers, will need to be integrated and subsequently evaluated for annotated gold-standard corpora such as GENIA.

Complex problems, such as industrial design and infrastructure planning, all rely on sophisticated computational techniques for production and analysis of models of real-world systems. Biology is just beginning to move in this direction. However, to provide the types of systems biology models that are required, we must rely on advanced computational techniques. This review shows that event

extraction methods lie at the heart of these techniques, and the various text mining techniques and resources discussed throughout this review should be explored to the fullest. It is through wider implementation and evaluation of these techniques and resources in systems biology that new insights will be gained and improvements made to the benefit of the community as a whole.

Acknowledgments

Work towards this review was partly supported by the Biotechnology and Biological Sciences Research Council in the context of grants BB/E004431/1 and BB/G53025X/1. We would like to thank Paul Thompson and John McNaught (National Centre for Text Mining, University of Manchester) for their helpful comments and support in producing this manuscript.

References

1 Klipp, E. *et al.* (2005) *Systems Biology in Practice: Concepts, Implementation and Clinical Application*, Wiley-VCH
2 Alon, U. (2006) *An Introduction to Systems Biology: Design Principles of Biological Circuits*, Chapman and Hall
3 Palsson, B.Ø. (2006) *Systems Biology: Properties of Reconstructed Networks*, Cambridge University Press
4 Ananiadou, S. and Nenadic, G. (2006) Automatic terminology management in biomedicine. In *Text Mining for Biology and Biomedicine* (Ananiadou, S. and McNaught, J., eds), Artech House Books
5 Okazaki, N. and Ananiadou, S. (2006) Building an abbreviation dictionary using a term recognition approach. *Bioinformatics* 22, 3089–3095
6 Tsuruoka, Y. *et al.* (2008) Normalizing biomedical terms by minimizing ambiguity and variability. *BMC Bioinformatics* 9 (S-3), S2
7 Gaudan, S. *et al.* (2005) Resolving abbreviations to their senses in Medline. *Bioinformatics* 21, 3658–3664
8 Wang, X. *et al.* (2010) Disambiguating the species of biomedical named entities using natural language parsers. *Bioinformatics* 26, 661–667
9 Ananiadou, S. and McNaught, J., eds (2006) *Text Mining for Biology and Biomedicine*, Artech House Books
10 Ananiadou, S. *et al.* (2006) Text mining and its potential applications in systems biology. *Trends Biotechnol.* 24, 571–579
11 Rzhetsky, A. *et al.* (2009) Getting started in text mining: part two. *PLoS Comput. Biol.* 5, e1000411
12 Torii, M. *et al.* (2007) A comparison study on algorithms of detecting long forms for short forms in biomedical text. *BMC Bioinformatics* 8 (Suppl 9), S5
13 Ananiadou, S. *et al.* eds (2004) Special Issue on Named Entity Recognition in Biomedicine. *J. Biomed. Inform.* 37, 393–528.
14 Cohen, K.B. *et al.* (2008) Nominalization and alternations in biomedical language. *PLoS ONE* 3, e3158
15 Tsuruoka, Y. *et al.* (2008) FACTA: a text search engine for finding associated biomedical concepts. *Bioinformatics* 24, 2559–2560
16 Chun, H.W. *et al.* (2006) Extraction of gene–disease relations from Medline using domain dictionaries and machine learning. *Pac. Symp. Biocomput.* 4–15
17 von Mering, C. *et al.* (2002) Comparative assessment of large-scale data sets of protein–protein interactions. *Nature* 417, 399–403
18 Kim, J-D. *et al.* (2009) Overview of BioNLP'09 Shared Task on Event Extraction, In *BioNLP 2009 Companion Volume: Shared Task on Event Extraction*, pp. 1–9, Association for Computational Linguistics
19 Winnenburg, R. *et al.* (2008) Facts from text: can text mining help to scale-up high-quality manual curation of gene products with ontologies? *Brief Bioinform.* 9, 466–478
20 Korbel, J. *et al.* (2005) Systematic association of genes to phenotypes by genome and literature mining. *PLoS Biol.* 3, e134
21 Sam, L.T. *et al.* (2009) PhenoGO: an integrated resource for the multiscale mining of clinical and biological data. *BMC Bioinformatics* 10 (Suppl 2), S8
22 Ozgur, A. *et al.* (2008) Identifying gene–disease associations using centrality on a literature mined gene-interaction network. *Bioinformatics* 24, i277–285
23 Thompson, P. *et al.* (2009) Construction of an annotated corpus to support biomedical information extraction. *BMC Bioinformatics* 10, 349
24 Miyao, Y. *et al.* (2009) Evaluating contributions of natural language parsers to protein–protein interaction extraction. *Bioinformatics* 25, 394–400
25 Miyao, Y. and Tsujii, J. (2008) Feature forest models for probabilistic HPSG parsing. *Comput. Linguist.* 34 (1), 35–80
26 Hull, D. *et al.* (2008) Defrosting the digital library: bibliographic tools for the next generation web. *PLoS Comput. Biol.* 4, e1000204
27 Attwood, T.K. *et al.* (2009) Calling International Rescue: knowledge lost in literature and data landslide! *Biochem. J.* 424, 317–333
28 Pettifer, S. *et al.* (2009) Visualising biological data: a semantic approach to tool and database integration. *BMC Bioinformatics* 10 (Suppl 6), S19
29 Hoffmann, R. and Valencia, A. (2005) Implementing the iHOP concept for navigation of biomedical literature. *Bioinformatics* 21 (Suppl 2), ii252–258
30 Dietze, H. and Schroeder, M. (2009) GoWeb: a semantic search engine for the life science web. *BMC Bioinformatics* 10 (Suppl 10), S7
31 Fontaine, J.F. *et al.* (2009) MedlineRanker: flexible ranking of biomedical literature. *Nucleic Acids Res.* 37, W141–146
32 Miyao, Y. *et al.* (2006) Semantic retrieval for the accurate identification of relational concepts in massive textbases, In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics* (Vol. 2), pp. 1017–1024, Association for Computational Linguistics
33 Tsuruoka, Y. and Tsujii, J. (2004) Improving the performance of dictionary-based approaches in protein name recognition. *J. Biomed. Inform.* 37, 461–470
34 Fundel, K. *et al.* (2007) RelEx – relation extraction using dependency parse trees. *Bioinformatics* 23, 365–371
35 Zweigenbaum, P. *et al.* (2007) Frontiers of biomedical text mining: current progress. *Brief Bioinform.* 8, 358–375
36 Kim, S.H. (2008) Kernel approaches for genic interaction extraction. *Bioinformatics* 24, 118–126
37 Krallinger, M. *et al.* (2008) Overview of the protein–protein interaction annotation extraction task of BioCreative II. *Genome Biol.* 9 (Suppl 2), S4
38 Xenarios, I. *et al.* (2000) DIP: the Database of Interacting Proteins. *Nucleic Acids Res.* 28, 289–291
39 Hermjakob, H. *et al.* (2004) IntAct: an open source molecular interaction database. *Nucleic Acids Res.* 32, D452–455
40 Ramani, A.K. *et al.* (2005) Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome. *Genome Biol.* 6, R40
41 Gasteiger, J. (ed.) (2003) *Handbook of Chemoinformatics: From Data to Knowledge*, Wiley-VCH
42 Dobson, P.D. *et al.* (2009) 'Metabolite-likeness' as a criterion in the design and selection of pharmaceutical drug libraries. *Drug Discov. Today* 14, 31–40
43 Hopkins, A.L. (2008) Network pharmacology: the next paradigm in drug discovery. *Nat. Chem. Biol.* 4, 682–690
44 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
45 Kitano, H. *et al.* (2005) Using process diagrams for the graphical representation of biological networks. *Nat. Biotechnol.* 23, 961–966
46 Mendes, P. *et al.* (2009) Computational modeling of biochemical networks using COPASI. *Methods Mol. Biol.* 500, 17–59
47 Duarte, N.C. *et al.* (2007) Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc. Natl. Acad. Sci. U. S. A.* 104, 1777–1782
48 Spasic, I. *et al.* (2009) KiPar, a tool for systematic information retrieval regarding parameters for kinetic modelling of yeast metabolic pathways. *Bioinformatics* 25, 1404–1411
49 Herrgard, M.J. *et al.* (2008) A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. *Nat. Biotechnol.* 26, 1155–1160
50 Oda, K. *et al.* (2008) New challenges for text mining: mapping between text and manually curated pathways. *BMC Bioinformatics* 9 (Suppl 3), S5
51 Rzhetsky, A. *et al.* (2004) GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data. *J. Biomed. Inform.* 37, 43–53
52 Rajagopalan, D. and Agarwal, P. (2005) Inferring pathways from gene lists using a literature-derived network of biological relationships. *Bioinformatics* 21, 788–793

53 Santos, C. *et al.* (2005) Wnt pathway curation using automated natural language processing: combining statistical methods with partial and full parse for knowledge extraction. *Bioinformatics* 21, 1653–1658
54 Kelder, T. *et al.* (2009) Mining biological pathways using WikiPathways web services. *PLoS ONE* 4, e6447
55 Elliott, B. *et al.* (2008) PathCase: pathways database system. *Bioinformatics* 24, 2526–2533
56 Barrell, D. *et al.* (2009) The GOA database in 2009 – an integrated gene ontology annotation resource. *Nucleic Acids Res.* 37, D396–403
57 Blaschke, C. *et al.* (2005) Evaluation of BioCreAtIvE assessment of task 2. *BMC Bioinformatics* 6, S16
58 Kim, J.D. *et al.* (2008) Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics* 9, 10
59 Tsuruoka, Y. *et al.* (2008) Accelerating the annotation of sparse named entities by dynamic sentence selection. *BMC Bioinformatics* 9 (Suppl 11), S8
60 Pyysalo, S. *et al.* (2007) BioInfer: a corpus for information extraction in the biomedical domain. *BMC Bioinformatics* 8, 50
61 Kim, J. *et al.* (2003) GENIA corpus – a semantically annotated corpus for bio-text mining. *Bioinformatics* 19 (Suppl 1), i180–i182
62 Miwa, M. *et al.* (2010) Event extraction with complex event classification using rich features. *J. Bioinform. Comput. Biol.* 8, 131–146
63 Divoli, A. and Attwood, T.K. (2005) BioIE: extracting informative sentences from the biomedical literature. *Bioinformatics* 21, 2138–2139
64 Huang, M. *et al.* (2004) Discovering patterns to extract protein-protein interactions from full texts. *Bioinformatics* 20, 3604–3612
65 Corney, D.P. *et al.* (2004) BioRAT: extracting biological information from full-length papers. *Bioinformatics* 20, 3206–3213
66 McNaught, J. and Black, W. (2006) Information extraction. In *Text Mining for Biology and Biomedicine* (Ananiadou, S. and McNaught, J., eds), pp. 143–177, Artech House
67 McDonald, D. *et al.* (2004) Extracting gene pathway relations using a hybrid grammar: the Arizona relation parser. *Bioinformatics* 20, 3370–3378
68 Friedman, C. *et al.* (2002) Two biomedical sublanguages: a description based on the theories of Zellig Harris. *J. Biomed. Inform.* 35, 222–235
69 Harris, Z. (2002) The structure of science information. *J. Biomed. Inform.* 35, 215–221
70 Friedman, C. *et al.* (2001) GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics* 17, S74–S82
71 Cimiano, P. *et al.* (2005) Ontology-driven discourse analysis for information extraction. *Data Knowl. Eng.* 55 (1), 59–83
72 Yakushiji, A. *et al.* (2001) Event extraction from biomedical papers using a full parser. *Pac. Symp. Biocomput.* 6, 408–419
73 Bunescu, R. *et al.* (2005) Comparative experiments on learning information extractors for proteins and their interactions. *Artif. Intell. Med.* 33, 139–155
74 Airola, A. *et al.* (2008) All-paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning. *BMC Bioinformatics* 9, S2
75 Miwa, M. *et al.* (2009) Protein–protein interaction extraction by leveraging multiple kernels and parsers. *Int. J. Med. Inf.* 78 (12), e39–e46
76 Krallinger, M. *et al.* (2008) Evaluation of text-mining systems for biology: overview of the Second BioCreative community challenge. *Genome Biol.* 9 (Suppl 2), S1
77 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
78 Kell, D.B. (2007) The virtual human: towards a global systems biology of multiscale, distributed biochemical network models. *IUBMB Life* 59, 689–695
79 King, R.D. *et al.* (2009) The automation of science. *Science* 324, 85–89
80 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
81 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
82 Spasic, I. *et al.* (2005) Text mining and ontologies in biomedicine: making sense of raw text. *Brief Bioinform.* 6, 239–251
83 Kell, D.B. (2009) Iron behaving badly: inappropriate iron chelation as a major contributor to the aetiology of vascular and other progressive inflammatory and degenerative diseases. *BMC Med. Genomics* 2, 2
84 Dobson, P.D. and Kell, D.B. (2008) Carrier-mediated cellular uptake of pharmaceutical drugs: an exception or the rule? Nat. *Rev. Drug Discov.* 7, 205–220