**BRIEFINGS IN BIOINFORMATICS.** VOL 15. NO 5. 856–877
Advance Access published on 22 February 2013
doi:10.1093/bib/bbt006

# Biological network extraction from scientific literature: state of the art and challenges

*Chen Li, Maria Liakata and Dietrich Rebholz-Schuhmann*

Submitted: 3rd November 2012; Received (in revised form): 24th January 2013

## Abstract

Networks of molecular interactions explain complex biological processes, and all known information on molecular events is contained in a number of public repositories including the scientific literature. Metabolic and signalling pathways are often viewed separately, even though both types are composed of interactions involving proteins and other chemical entities.

It is necessary to be able to combine data from all available resources to judge the functionality, complexity and completeness of any given network overall, but especially the full integration of relevant information from the scientific literature is still an ongoing and complex task.

Currently, the text-mining research community is steadily moving towards processing the full body of the scientific literature by making use of rich linguistic features such as full text parsing, to extract biological interactions. The next step will be to combine these with information from scientific databases to support hypothesis generation for the discovery of new knowledge and the extension of biological networks.

The generation of comprehensive networks requires technologies such as entity grounding, coordination resolution and co-reference resolution, which are not fully solved and are required to further improve the quality of results. Here, we analyse the state of the art for the extraction of network information from the scientific literature and the evaluation of extraction methods against reference corpora, discuss challenges involved and identify directions for future research.

***Keywords:*** *text mining; network extraction; event extraction*

## INTRODUCTION

A biological network is represented in a graph structure composed of nodes that denote biomolecules and edges between the nodes representing the interactions or reactions between the biomolecules. Network representations serve many purposes in bioinformatics, and most importantly, networks are used to judge the functional behaviour of interaction networks on the molecular level. Network representations are used to simulate, analyse and visualize the responses of protein interaction networks, metabolic pathways, specific synapses and even whole systems such as an organ, e.g. the liver or the brain. Studying the topological structure and the functional responses of such systems aims to reveal yet undiscovered mechanisms that could explain or improve specific

Corresponding author. Chen Li, EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SD, UK. Tel: +44(0)1223 494 537; Fax: +44(0)1223 494 468; E-mail: chenli@ebi.ac.uk

**Chen Li** works at the European Bioinformatics Institute (EMBL-EBI). He is interested in biomedical text mining by using machine learning and natural language processing methods. In the past, he also contributed to Systems Biology community by participating the development of public services.

**Maria Liakata** has been recently appointed to the position of Assistant Professor at the Department of Computer Science, University of Warwick, UK. She holds a Leverhulme Trust Early Career Research Fellow and is a visiting researcher at EMBL-EBI. Her work has contributed to advances in knowledge discovery from corpora, automation of scientific experimentation and automatic extraction of information from the scientific literature.

**Dietrich Rebholz-Schuhmann** holds a master in medicine and computer science. He is visiting Research Group Leader at the EBI, where he established several world-class biomedical text-mining services. Currently he pursues research in multilingual biomedical semantic resources at the University of Zuerich, and manages the Journal of biomedical semantics.

unfavourable health conditions. In particular, the networks of signalling and metabolic pathways are at the focus of on-going research to explore the causes of disease, and increasingly these different types are investigated jointly for reconciling the outcomes from regulatory or metabolic mechanisms. These developments are part of the research in Systems Biology, which aims to build large-scale networks of complete living systems.

Following the increasing interest in biomedical networks and their availability in electronic form, many public repositories were created for hosting such network data [1–3]. The electronic representation of a network can serve several purposes: not only does it provide a visual representation of a biological system but, at the same time, qualitative models can help interpret experimental data, help predict reactions between entities, even be used in functional genomics to help infer new gene functions [4]. Biological models can also be used to simulate processes according to quantitative information that has been attributed to the interaction between particular entities. Qualitative models are encoded usually in some logical formalism such as Prolog, whereas quantitative models are usually encoded according to standards such as the Systems Biology Markup Language (SBML) [5], CellML [6] and BioPAX [7], and then kept in special repositories [8, 9]. Data population of the repositories, especially open platforms like WikiPathways [10] and BioModels Database [8, 11], are community-driven efforts, as the creation and encoding of the networks require significant contributions from domain experts within the research community. This process is time consuming, especially if the demands on the data quality and reusability are high, and would profit from computer-assisted support enabling better and faster curation of the data and semantic integration with other data resources [12].

The latest developments in natural language processing (NLP) and the integration of biomedical semantics lead the way to support manual curation, semi-automated curation and semantic enrichment of biological networks [13]. Advances in text mining (TM) provide significant benefit to the on-going study of biological networks, ranging from molecular biology [14, 15] to systems biology [16]. In this article, we distinguish analyses concerning complete biological networks from those related to individual biological events or sets of events only, i.e. we focus on advances in TM that pave the way to automated extraction of complete networks. Furthermore, we emphasize the conditions that TM has to fulfil to be able to identify and distinguish between different types of networks, such as signalling pathways versus metabolic pathways. Apart from presenting existing methods for network extraction and resources available, the review will discuss remaining challenges or problems mainly linked to coreference resolution, identification or inference of the outcome of events, hidden relations between entities, and event interconnections.

## BIOLOGICAL NETWORKS AND TEXT MINING

The term biological network may be used to refer to various aggregations of the relations between biomolecules, diseases or phenotypes, etc. In this article, we define a biological network as a directed graph consisting of consecutive interactions between biomolecules, such as proteins, genes and chemical substances. This definition includes signalling pathways, metabolic pathways and regulatory networks.

In the era of 'omics' data, the desired outcome of many interdisciplinary studies in biology is the combination of semantic resources and high throughput data to produce reliable biological networks. Progress in the field of biological TM supports the extraction of such networks automatically from the literature [17–21]. This process of network extraction is complex, and the implementation of a given system may vary (e.g. entity normalization may happen after relation/event extraction). We can distinguish six phases in a typical process as illustrated in the pyramid structure of Figure 1. First, biological entities and concepts are tagged in the text; this corresponds to *Entity mentions*. In the *Entity normalization* step, the entity mentions are disambiguated and linked to entries in biological databases or ontologies. *Relation extraction* is the next processing step, which aims to detect biological relationships between any recognized entities (e.g. 'inhibit', 'bind', etc). The following stage, *event extraction*, is used to identify the directionality (i.e. which are the agents and which are the targets) and the polarity (i.e. the positive or negative effect of a biochemical reaction. e.g. up-regulation and down-regulation) of a biological interaction and further qualifies any given relation as a biological event. *Network extraction* is the step of interconnecting all entities through all possible events, so as to better understand and predict the

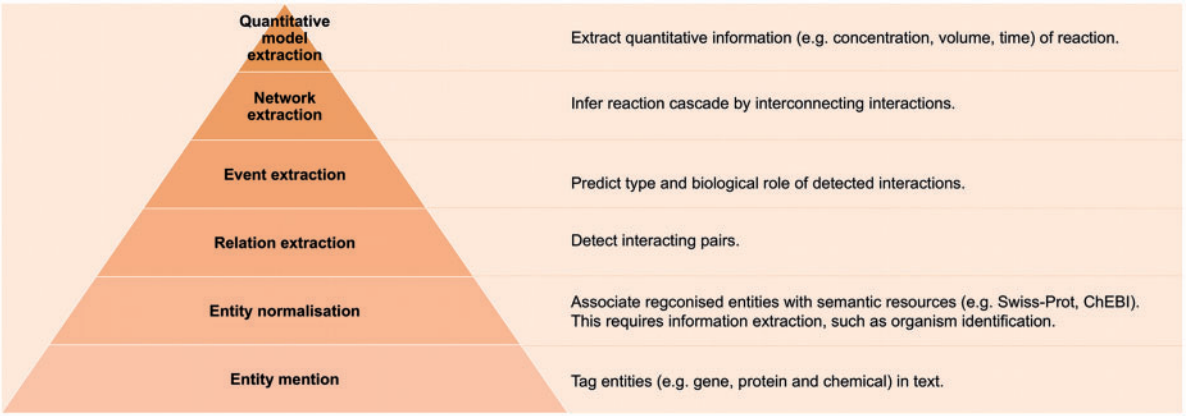


**Figure 1:** Six phases of TM for network extraction. As we advance from identifying entities to creating a quantitative model of a reaction network, the TM technology required becomes increasingly more complicated. But it is closer to its initial mission, assisting the analysis of biological systems. Each step relies on the lower level process, and even the further low level, e.g. quantitative model extraction relies on network extraction and entity recognition etc. Relation and event extraction could happen without entity normalization, but the latter is necessary for event interconnection to allow network extraction. Current TM research has reached event extraction and had attempts in network extraction.

response for complete biological phenomena from the network, such as the initiation of cell growth as the response to an extracellular signal from ligand binding. Such a network can be modelled as a set of molecular interactions that induce binary signals (on/off) and can be encoded in a logical model. *Quantitative modelling* requires the extraction of quantitative parameters for a given biological network, which give specific details on individual reactions. These parameters include mass and concentration of reactants and products, reaction time etc. These parameters are relevant for simulating fine-grained biological phenomena and provide a much better insight into how biological systems function. Although the tasks at the top of the pyramid are challenging, they are required to help deliver information for the simulation of complex biological systems.

The level of maturity of the various TM technologies required to address the above processing steps differs significantly. Current TM research has focused on the first four steps of the pyramid with attempts at network extraction. Named entity recognition (NER) is well established as a mature technology in the TM research domain. Selected entity mention solutions recognize entities at a minimum of 80% balanced F-score and even better [22], although the quality of the gold standard corpus can lead to a bias in the evaluation, and cross-corpus [23] or cross-tagger evaluation can lead to variable results. There is also great differentiation between recognition performance for different types of entities, with genes and proteins having the highest recognition rates. Relation extraction and event extraction are currently the TM tasks that receive the most attention in the biological TM research community. Organized shared tasks and challenges addressing these tasks [15, 24] have gained a lot of prominence in the research community, and their corpora and performances set the standard for research in this area. However, the focus in the challenges resides solely with the identification of relations and events between proteins as opposed to relations involving other molecules, such as chemical entities. Meanwhile, researchers have stepped up their efforts on automated [17] or semi-automated [18] network extraction, and, at the same time, little effort has been spent on the extraction of quantitative models from the literature. KiPar [25], KID [26] and KIND [27] are rule-based systems for extracting quantitative information from text. Nonetheless, adding quantitative information to biological networks is complex and challenging, as this information is often not available in the text itself but hidden in graphs and tables. The recently announced BioNLP'13 [28] shows promise in that it is the first time systems will be asked to tackle pathway curation as a task, thus investigating current TM capability for automatic pathway extraction.

On occasions, the term *relation* is used as a synonym for an *event*. Ananiadou *et al*. [16] distinguish a relation from an event in the following way.

A relation expresses the existence of an interaction between a pair of entities. For example, in 'calcium ions penetrate the site and trigger *VAMP*, *syntaxin* and *SNAP-25* to bind together in a lock and key fashion' [29], each binary binding among *VAMP*, *syntaxin* and *SNAP-25* is the relation between a pair, whereas the formation of the complex composed of three bindings is a biological event. Therefore, an event represents a functionally complete behaviour of biomolecules, which may include several relations. Furthermore, an event may have attributes, which further characterize the relations, such as the directionality of protein binding and the polarity of regulation. An event requires the identification of more complex information than a relation, as it represents a comprehensive biochemical reaction involving different entities. Therefore, it should also include information about the outcome of a reaction (refer section 'Event interconnection'). To characterize an event fully, one needs to be able to capture various aspects of the scientific discourse, such as whether an event is hypothesized or the outcome of a study, or the level of certainty of the event [30, 31]. This review will also discuss a relation or an event that may span across several sentences or exceed the confine of bigger text units.

Even without considering quantitative information, the concept of a network is greater than that of a bag of events. Disregarding the issue of redundancy between events, the sequence of the reactions/events plays an important role in deciding its molecular role in a regulatory mechanism under specific conditions. Under certain circumstances, a network may be topologically and quantitatively dynamic. For example, when other cellular gradients influence a metabolic pathway, the pathway may change over time. These challenges make network extraction conceptually and technologically much more complicated than extracting a bag of events (refer section 'Event interconnection'). However, even obtaining a bag of unique events where coreference between entities and entity mentions have been resolved is far from trivial.

Signalling pathways and metabolic pathways pose different challenges to TM event extraction, as signalling pathways are mainly composed of binary protein–protein interactions (PPI) whereas metabolic pathways consist of chemical–protein interactions (CPI). Both networks are often studied independently, although they might be coupled in order to model similar processes at different levels of a functional hierarchy, e.g. insulin signal transduction pathway and regulation of blood glucose.

## Signalling pathways

A signalling pathway is a series of PPI that transmit signals into a cell. This mechanism is invoked when an extracellular molecule activates a cell surface receptor protein, usually by binding to a receptor's site. A signalling pathway mainly involves the processes of protein binding, phosphorylation and localization. There are three main challenges in the automated extraction of signalling pathways (Figure 2).

The first major issue is that only a subset of all interacting entities is directly reported in the scientific literature and in particular as a relation between entities within the same sentence [34]. Some of the entities and events are explained in detail in consecutive sentences, where references to previously mentioned entities are made through coreference mentions (Point 1 in Figure 2). Coreference resolution in the scientific literature is still a complex problem and will be discussed in section 'The role of coreference resolution and discourse'. There are also cases where the entities that are the products of reactions are not directly mentioned in the text. In Figure 2, *DecorinEGFR* is a multi-protein complex, which is the product of the event, '*Decorin* . . . bind to and activate *EGFR* . . .'. The product of the event is inferred, but it is referred to later in the text, often using referential metonymy. For example, a protein name may be used to refer to an entire protein complex, which includes the protein, where the complex has been implied as the product of previous reactions. Such is the case in Point 2 in Figure 2, where *EGFR* in the sentence, 'it binds activated *EGFR* via an *SH2 domain*', actually represents the multi-protein complex instead of the original receptor. Thirdly, there can be mention of a part of a molecule, e.g. a binding domain or residue, which denotes a particular protein or complex, previously mentioned or inferred (Point 3 in Figure 2). In this case, the entity normalization requires semantic information from the entity context and from biological reference ontologies. For example, considering the information, e.g. gene locus and attribute, from Gene Ontology can support disambiguation and improve the normalization performance [35]. These challenges often cause topological mistakes in the pathway and reduce the automatic and correct interpretation of the generated network. The generation and use of proper nomenclature for the correct

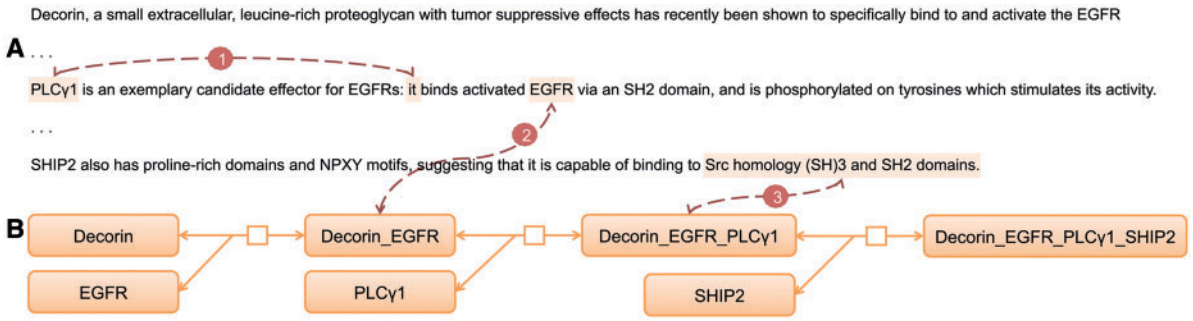


**Figure 2:** Three hurdles for TM in entity recognition for signalling pathways. **A** is text describing signal transduction stimulated by *Decorin* (From PubMed ID:10209155). The first sentence describes an event of *Decorin* binding *EGFR*. The second sentence says that *PLCγ1* binds multi-protein complex *Decorin_EGFR*, which was generated by the binding event of the first sentence. In the third sentence, *SHIP2* binds another new complex, which was generated in the second sentence. The diagram of the pathway snippet is visualized in (**B**). The three challenges are (i) anaphoric coreference: in this example, we have a classic case of pronominal anaphora. (ii) part-for-whole metonymy (part-for-whole metonymy: part of an entity is used to refer to the whole entity) (synecdoche) [32]: the *EGFR* protein is used as a metonym for the entire protein complex resulting from the binding of *Decorin* and *EGFR*. (iii) predicative metonymy (predicative metonymy: an association is made between a source and target word based on concomitance. It consists in conveniently making-way of a predicate for a non-standard, but related, argument) [33]: in this example '*Src homology (SH)3* and *SH2 domains*' are the binding sites of the protein complex which has resulted from the binding of *Decorin_EGFR* and *PLCγ1*. We assume that binding events take protein or protein complexes as their arguments and have a site as a modifier. Thus, in 'SHIP2 is capable of binding to *Src homology (SH)3* and *SH2 domains*', the binding sites are used in place of the protein complex which is the target of the binding event.

representation of macromolecular complexes would be an important step to solving this problem. For example, the protein complex generated by the binding of EGF to EGFR has been called *EGF/EGFR* [36] or *EGF·EGFR* [37], which supports correct extraction results.

## Metabolic pathways

In contrast to signalling pathways, a metabolic pathway is composed of interactions between proteins as well as proteins and small molecules such as chemicals and enzymes, and possibly small molecules only. Chemical substances are denoted according to the standard nomenclature produced by the International Union of Pure and Applied Chemistry (IUPAC). Organic compounds and inorganic compounds have their own nomenclature systems, respectively. The Chemical Abstract Service developed a scheme to index chemical substances and assigned each chemical substance an identifier known as CAS registry number. This improved nomenclature eases the recognition of reactants and the identification of products of a reaction. It is useful for interconnecting extracted events into a network (refer section 'Event interconnection'). Nevertheless, the recognition of chemical entities by TM systems is not as advanced as the recognition of protein and gene names, which has been at the heart of challenges and shared tasks in biological TM [15, 24, 38]. This is an additional complication to the ones mentioned above for the extraction of signalling pathways. Moreover, enzymes in the case of metabolic pathway extraction rarely appear in the same passage, let alone the same sentence [39].

## The coverage of signalling and metabolic pathways in the BioModels database

To measure the distribution of interactions linked to signalling pathways in contrast to metabolic pathways, and to investigate the challenges posed to TM by the available data, we evaluated BioModels database (BMDB) against TM results on the basis of entity recognition and interaction identification. A generic TM system that maximizes recall using co-occurrences at the level of the abstract has been used. BMDB has been chosen for this experiment because (i) it is a standard resource for networks covering signal transduction and metabolic pathways; (ii) all networks are encoded in a standardized and machine-readable format; (iii) its networks represent biological processes in a species-independent way; (iv) each network has been manually curated

according to guidelines [40], on the basis of each individual corresponding publication, to assure the network fully and faithfully represents the reactions described in the publication.

A corpus has been generated by tagging protein and chemical entities in the whole of MEDLINE. The tagging used the Whatizit pipeline [41] and is based on two dictionaries compiled from UniProt [42] and ChEBI [43]. Firstly, models are grouped into five types (the left chart in Figure 3). The first three types are decided based on the entity type, when a model either contains proteins only (*protein network*), or contains chemicals only (*chemical network*), or has both of them (*protein–chemical network*). Then, by using annotations in the models, the models are selected and put into two groups: *signal transduction pathways* (GO:0007165), and *metabolic process pathways* (GO:0008152). The reactions in BMDB were identified in the corpus by checking co-occurrences of all reactants. If all reactants of a reaction occur in the same abstract, then the assumption is that the reaction can be recovered from the abstract using TM. By this method, fewer chemical entities in BMDB are found in MEDLINE compared with proteins. Twenty-nine per cent of the chemical reactions from BMDB are found in MEDLINE in comparison with 42% of the PPI. Twenty-eight per cent of the reactions between proteins and chemicals are assumed to be recoverable and 29% for chemical reactions. For signal transduction models, 58% of entities could be retrieved and 23% of reactions. The results for metabolic process models are 52% and 19%, respectively. To investigate the trends in the scientific literature regarding biological networks, the *signal transduction pathway* and *metabolic process pathway* models are chronologically measured (the right chart in Figure 3). Pathways are classified into four groups: *before 1996*, *1996–2000*, *2001–05* and *2006–12*. For the pathways published before 1996, reactions are rarely found for either type. This is because there were few publications digitally available before 1996. Many of these publications' records in MEDLINE only contain the titles. For both the entities and the reactions, the rate of those found in MEDLINE became higher after 1996. In BMDB, signal transduction models for *1996–2000* (are relatively smaller) contain fewer pathways and reactions compared with the same type of models published later. However, a higher percentage of their entities were annotated, which explains why higher rates of the entities and reactions in these models are found in MEDLINE. Generally, the entities found for signal transduction are more than those for metabolic process pathways. Meanwhile, the

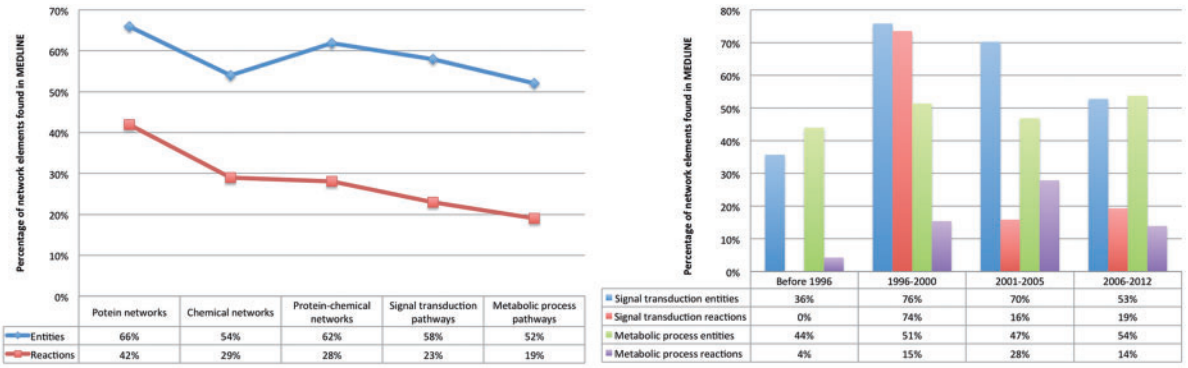


**Figure 3:** The rates of BMDB's entities and reactions found in MEDLINE. In the **left** chart, the models are categorized into five types according to their entity type and network type. The five types are *protein networks*, *chemical networks*, *protein–chemical networks*, *signal transduction* and *metabolic process*. The line on the top denotes the distribution of entities, which have been identified in MEDLINE [44] abstracts by dictionary-based NER. The lower line measures the reactions mentioned, which could be recovered from MEDLINE by abstract-level co-occurrence. In the **right** chart, the signal transduction (GO:0007165) and metabolic process (GO:0008152) models are measured separately based on their publication date. The models are put into four groups: *before 1996*, *1996–2000*, *2001–05* and *2006–12*. The first bar of each cluster reflects the percentage of the signal transduction pathway entities, which have been found in MEDLINE. The second bar of each cluster is the percentage of the signal transduction pathway reactions, which could be found in MEDLINE by abstract-level co-occurrence. The third bar of each cluster is the percentage of the metabolic process entities, which could be found in MEDLINE. The fourth bar of each cluster is the percentage of the metabolic process reactions, which were found in MEDLINE by abstract-level co-occurrence.

number of metabolic pathways in the literatures is gradually increasing. However, the selection of the models in BMDB largely depends on the curators. Although it supports submission from authors, a submitted model still has to go through the curation pipeline to become publicly available. Therefore, the right chart reflects the BMDB's trend regarding biological networks rather than those in the general scientific literature.

These findings suggest that many available relations between entities may be described in full-text articles but missed by the extraction procedure. However, tables and figures are another considerable part of the article where reactions are listed and depicted, and these are currently not usually considered by TM systems. As with other resources, newly published networks are mostly based on pre-existing knowledge. A single article often only presents novel reactions or molecular attributes with respect to known networks. Therefore, TM is unlikely to be able to extract a complete network from a single article even with access to full-text. In this case, a system for automatic network extraction would need to crawl a set of publications, which are bibliographically referenced or conceptually grouped. In addition to the previous, owing to different participating entities and the numbers of their mentions in publications, PPI-constructed signalling pathways and CPI-constructed metabolic pathways need different extraction strategies.

### Relation discovery and curated databases

If a molecular relation found in the scientific literature using TM is not referenced in a bioinformatics database or knowledge base, it is either judged as a false positive (TM mistake) or as an unknown relation (incomplete DB). However, such cases may point to the existence of relations, which hold the potential of new discoveries using biomedical TM. Although TM cannot and could not replace biological experiments for the discovery of biological relations, it can nevertheless provide evidence about possible molecular interactions and help hypothesize relations. In this article, we discuss two types of such relations. The first type covers relations that are explicitly mentioned in the publications, but are missing from curated databases. Explicit mentions of relations in text extracted using TM can be used to populate or enrich databases, although manual curation of the results may be required to make sure they comply with database standards. The second type of relation is not represented as a direct interaction in the text; however, it can be inferred from other relations that are mentioned in the text, in combination with knowledge in existing knowledge resources. Another possibility is to infer such relations by means of combining text with experimental evidence or by finding tentative relations in the text, which could provide evidences for yet uninterpreted experimental findings [45]. For example, after optimizing a logical model about the responses of human cells to seven cytokines by using gene expression data, a direct link between *IGF1* and *Akt* without *PI3K*'s activation has been predicted [46]. The prediction questions the current understanding about the necessity of PI3K for activating Akt in the pathway of regulating glucose uptake. In fact, high-throughput data generated a large number of such predictions, and suggested that, *in vitro*, a protein may play more roles than what has been perceived. Although experimental evaluations covering all predictions are expensive and impractical, evidence automatically collected from the literature can support and prioritize further experiments. In the above example, this second type of relation cannot be extracted or predicted explicitly based on NLP and TM methods only. Nonetheless, such relations can be identified through statistical heuristics [47], with the help of unsupervised machine learning methods [48] or with a combination of TM and logical inference [49].

In addition to the above, the molecular interactions or relations currently already in the knowledge bases would strongly benefit from relation discovery for evaluation and annotation purposes. More specifically, biological database curators would like to be able to link curated data with relations in the publications, as the latter have more detail and more context illustrated in a more explicit way [50], especially when figures are alongside. The available event extraction solutions can be implemented as software and integrated into routine biological workflows [12].

## ENTITY RECOGNITION FOR NETWORK EXTRACTION

### Entity-recognition approaches

Networks are represented as graphs, where entities constitute the nodes. Therefore, all network extraction methods have to perform accurate entity recognition (NER). So far there have been

dictionary-based solutions or machine learning-based approaches for NER. Dictionary-based approaches usually collect names, synonyms and acronyms of entities from bioinformatics databases (e.g. UniProt, ChEBI), compile a dictionary and perform string matching against scientific texts. Following the progress in standardized semantic resources, the dictionary or lexical approach is becoming increasingly more compatible with biomedical ontologies [51, 52]. The advantage of these approaches is that they do not need to be trained and can theoretically be applied to any scientific text for recognizing NEs.

In the case of machine learning approaches, annotated corpora are required to train the entity recognition algorithms. NER is usually regarded as a sequence labelling task, as token order plays an important role in identifying NE components. Hidden Markov Models (HMM) as well as Maximum Entropy Markov Models (MEMMs) have been used to address this [53]. Conditional Random Fields (CRF) is a popular machine-learning alternative to the previous for sequence labelling, often used by biomedical NER, as it combines the advantage of MEMMs in exploiting non-independent contextual features of the entity without a label bias problem [54–58]. Another popular machine learning technique used in NER is Support Vector Machines (SVM) [59, 60]. NER by SVM follows a text classification approach, where each token is given the appropriate NE category based on the morphological characteristics of the named entity and a set of contextual features. These approaches are fast and have a high performance.

One can further distinguish NER solutions into entity mention (EM) and entity normalization (EN). EM solutions detect text components that make reference to a gene or gene product, whereas the EN solutions link the recognized entities to data entries in bioinformatics databases. The use of database information can contribute to the semantic disambiguation and homologous analysis of an entity. By crawling cross-linked resources, such as Gene Ontology or PSI-MI [61], network extraction systems can retrieve affinitive entities and interaction patterns. The host organism is usually determined by contextual lookup. As a result, machine learning solutions, such as SVM [62], play an essential role in EN, as they can cope well with the assembly of a large amount of features from the context of an entity as opposed to rule-based solutions. GeneTUKit [63] uses SVM and was ranked first in the BioCreAtIvE III GN task according to the Threshold Average Precision (TAP-20) measure. GNer [64] is based on a dictionary compiled from Entrez Gene [65] and BioThesaurus, which it combines with SVMs using a set of extraction rules. GNAT [35] is a hybrid system using a dictionary and CRF for EM of genes and then correlates discovered genes with corresponding species.

Chemical entities play an important role in biological networks. They could be, for example, messengers, in signalling pathway, or diverse metabolites. The identification of chemical entities has not been studied as extensively as the identification of genes and proteins. As a result, fewer resources, i.e. freely available TM systems and corpora, have been made available for the identification of chemical entities and small molecules in comparison with gene and proteins. The SCAI (Fraunhofer-Institute for Algorithms and Scientific Computing) corpus for chemical compounds comprises 463 MEDLINE abstracts in the training set and 100 abstracts in the test set [66]. Entities in the corpus are IUPAC terms, trivial names, abbreviations, sum formulas and chemical family names. OSCAR [67, 68] is one of the earliest tools for identifying chemical entities and reaction types in the scientific literature. On the SCAI corpus, ChemSpot [69], combining a dictionary and CRF trained on the IUPAC training corpus, outperformed OSCAR4's F-score by 10.8%.

## Challenges, corpora and resources for NER

Challenges and competitions have been introduced to measure different NER solutions against the same benchmark so as to gain an insight into the parameters of success for different solutions. The resources created and used for the competitions, e.g. corpora, tools and performance measures, have provided a long-term benefit to the community and have helped move the field forward. Since 2004, the BioCreAtIvE initiative [70] has organized three challenges, and covered topics ranging from gene mention [22, 71], gene normalization [70, 72] and functional annotation (FA) [73] to PPI [15]. Each approach is usually good at a particular task or aspect. By harmonizing different systems together, it has been shown that a hybrid system can outperform individual systems [71, 72]. This has lead to the notion that training on harmonized annotations can improve system performance. To this effect, the CALBC (Collaborative Annotation of a

Large Biomedical Corpus) initiative attempted to harmonize contributions from the community to automatically annotate a very large corpus (1 million abstracts) [74]. It explored different harmonization methods based on the semantic groups (protein/gene, chemical, disease and species) and each annotation solution characteristics. This initiative has also introduced the notion of a silver standard corpus, which is created from the harmonized output of different systems. In TM, a gold standard corpus (GSC) is considered essential for evaluating solutions and constitutes training data for many systems [75–77]. However, creating such resources through manual annotation by specially trained experts is expensive. The generated SSC-I silver corpus can be used as training and testing data for NER tasks for protein/gene, chemical, disease and species. There are relatively fewer corpora and community challenges for chemical entities, compared with proteins. Except for SCAI, another corpus for chemical compound has been made available through the collaboration of the European Patent Office and the ChEBI team [44, 78]. The Colorado Richly Annotated Full Text Corpus (CRAFT) [23] contains 67 full text articles annotated with seven different biomedical concepts. Apart from ChEBI entities, it also contains annotations from Cell Ontology, Entrez Gene, Gene Ontology, NCBI Taxonomy, Protein Ontology and Sequence Ontology [79].

NER solutions rely on semantic resources, especially for EN. A number of terminological and semantic resources can potentially benefit NER systems. However, the information in the resources and their cross-linked databases have not been fully exploited for either EN or interaction extraction. Swiss-Prot [43] is a peptide sequence database, which has been widely used as a dictionary for entity recognition, but other information it contains, such as location, organism, interaction and function, has not been considered so far. Similarly, protein structure in Protein Data Bank [80] provides information such as the binding domain, which is useful for *part of whole metonymy* (Point 3 in Figure 2). Protein family databases, e.g. Pfam [81], are useful to link a generic mention with a specific mention. The entities that constitute the essential blocks of metabolism can be linked respectively to entries in databases of chemicals [44, 82–85], enzymes [86], metabolites [87, 88] and drugs [89–91] etc. Although TM can benefit from the information in these resources, it can also help curate and update them [50]. Research in creating links between entities (e.g. genes, phenotypes, diseases) in different knowledge resources to find evidence for disease can also benefit from TM [79,92–95].

## EVENT EXTRACTION FOR NETWORK EXTRACTION

In this review, an *event* represents a biochemical reaction, e.g. PPI or CPI, within a signalling pathway or a metabolic pathway (refer section 'Biological networks and text mining'). Rapid improvement of experimental methods, e.g. two-hybrid screening and mass spectrometry for PPI investigation, generated substantial publications about biochemical reactions. Co-occurrence, pattern-based and machine learning are the three dominant approaches for automated bio-event extraction by TM [96]. In the case of co-occurrence-based methods, the underlying assumption is that a pair of entities is interacting when they appear in a text unit, such as a sentence, a paragraph or an article. This approach is effective when entities concerned are ideally at the level of the same text unit. More sophisticated approaches to event extraction rely on syntactic parsing to extract grammatical relations. These grammatical relations are aligned with expert-defined information extraction templates, or characterized using machine learning algorithms.

### Co-occurrence

A co-occurrence approach to event extraction is less computationally intensive, as it does not involve syntactic parsing. Precisely because of this, there is no way to detect the network directionality or even confirm that the entities are actually interacting. The systems [21, 97–99] built on it can be ran against very large corpora [41, 100]. Co-occurrence-based systems can also be combined with a set of trigger words to detect interaction types [101]. The overall number of connections between specific entities can be quantified to reveal entity clusters to represent sub-networks or detect synonyms etc. to reveal the underlying structure of a network [97, 102, 103]. For example, CoPub [102, 103] uses regular expressions to search for a term or a pair/set of terms to infer relations between co-occurring genes, drugs, pathways and diseases. The discovered relation could be a hidden relation of the first type (refer section 'Relation discovery and curated databases'), if it is novel for the curated databases. Based on an ABC

principle, Frijters *et al.* [104] inferred that entity A interacted with C when A interacts with B and B interacts with C. The relations discovered in this way could be candidate-hidden relation of the second type (refer section 'Relation discovery and curated databases').

Although most co-occurring entities in text are not really interacting [34], co-occurrence captures the maximum number of true positives within a given unit of text. Co-occurrence-based extraction can be used as a baseline, as it reflects the maximum recall that an extraction system can get within a text unit. Meanwhile, owing to the agility, co-occurrence can be adopted for quickly filtering articles related to particular concepts [99]. Subsequent processes, which could be more advanced but computationally intensive, such as syntactic parsing, can be run on the data filtered in this way.

## Patterns, Syntactic Parsing and Machine Learning

Co-occurrence-based event extraction suffers from the problems of a bag-of-words approach, which counts the frequency of words encountered in the same context without considering grammatical relations holding between them or even word order. Entities mentioned in the same discourse unit (sentence, paragraph, abstract) are not necessarily interacting. Furthermore, the presence of a trigger word may not always correspond to an actual interaction between entities, and the problem becomes complicated when several potential trigger words appear in the same text unit. This often results in a large number of false positives, as all combinations between entities and trigger words are produced. For example, for the sentence, '*Phosphorylation* of *p53* disrupts *Mdm2-binding*', a system only identifying entities and trigger words will ignore contextual information and erroneously report that p53 binds Mdm2. In addition, common syntactic patterns such as coordination increase the number of false-positive results. For example, in the sentence, 'Binding of *hnRNP H* and *U2AF65* to respective *G-codes* and a *poly-uridine tract*', the two splicing factors and the two binding targets are forming alternatives, respectively, that would lead to four different pairs (*hnRNP H* → *G-codes*, *hnRNP H* → *poly-uridine tract*, *U2AF65* → *G-codes* and *U2AF65* → *poly-uridine tract*) when using co-occurrence analysis. To accurately acquire the relations, it is necessary to syntactically parse and analyse the sentential structure.

Syntactic parsing is the process of analysing text. It assigns parts of speech to sequential tokens and builds grammatical structure upon tokens. Thus, event extraction systems can determine the existence and attributes of an interaction based on the syntactic characteristics. Shallow parsing, sometimes called chunking, only identifies the sentence constituents, e.g. noun phrase, verb etc. It does not aim to analyse the grammatical relations between the constituents. By contrast, deep parsing is able to tackle grammatical relations and allow the consideration of semantic relations between constituents [105, 106], as it constructs a more detailed structure based on syntactic grammars. There are many parsers that correspond to different grammars and schools of thought. Among them, dependency parsing and head-driven phrase structure parsing are two popular approaches. Dependency parsers generate links between words, where one is the head and the other the dependent, whereas phrase structure parsing organizes syntax into nested structures, usually syntactic trees. It is possible to generate dependencies from both types of parsers [107–112] while the former have a clear appeal for applications which require grammatical relations, such as event extraction.

Pattern-based systems traverse the extracted syntactic structures and align them with a set of patterns to spot syntactic characteristics of interactions. These patterns are either produced by domain experts or can be obtained automatically from the text. They are usually implemented as regular expressions or as templates that include part-of-speech (POS) information [113]. Patterns also can be derived automatically from a corpus, using machine-learning, based on a small set of patterns, being collected from sentences with similar characters of POS and/or grammatical relations [114–116]. In either case, the patterns model the characteristic language structure used in the domain and the corpus in question [117], therefore, the approach can achieve high precision [118–121]. Temkin and Gilder [122] created a set of patterns to extract CPIs by interpreting the output from a context-free grammar (CFG) parser. The interpretation of the type of interaction can vary significantly depending on the context. Thus, contextual information, such as entity types, should be taken into account for determining event types. For example, 'a phosphate group attaches on a protein' is an event of type *phosphorylation* if the argument 'phosphate group' is considered, rather than an event of *Binding*, which could be triggered by the

verb 'attach'. For this reason, machine learning approaches have become popular, as they statistically capture the characteristics of context and arguments relating to syntactic structures.

Machine learning approaches, when analysing grammatical relations obtained from parsing, can adopt different strategies, as they focus on different aspects of the syntactic structure. There is a trend in combining different parsers so as to make the most of the syntactic structure [123, 124]. The challenges [15, 24, 38] that have taken place in the last few years have produced valuable resources and analyses for the bio-TM community, and at the same time have provided standards for event extraction. In BioNLP'11 [24], the extraction of an event can be achieved by checking three aspects: event triggers, argument linking and argument grouping. Some approaches extract triggers, arguments of simple events and arguments of complex events in three separate stages [125, 126]. More recently, models for joint extraction of event triggers and arguments have been proposed [127–130]. McClosky *et al*. [130] transforms candidate events, which consist of preliminarily recognized triggers and arguments, and their arguments into dependency trees. Subsequently, it uses a re-ranking dependency parser, which was modified from the MSTParser [131, 132], to parse the event tree.

Variants of event expression in human language always exceed what a single training annotated source can cover, regardless of its size. To intelligently adapt themselves to arbitrary resources, event extraction systems [129, 133, 134] have used domain-adaptation strategies to get supplementary information from unannotated data when training their models.

## The role of coreference resolution and discourse

Interactions and events may be expressed or denoted in several mostly consecutive sentences. Limiting the unit of event extraction to the sentence level reduces system recall, and this is an issue that current systems for event extraction are beginning to address. Entity coreference resolution aims to link together mentions of the same entity across different discourse units (usually sentences) [135]. Event coreference is also beginning to receive attention, as it caters for the relation between verbs pointing to the same event or entities referring to events. It is expected that entity coreference and event coreference can assist each other [136].

The use of coreference resolution is vital in improving coverage in network extraction, as it can allow interlinking events in the network ('Network construction'). Figure 4 shows four typical cases of coreference in the biomedical literature. The pronoun *It* in sentence A makes reference to *epidermal growth factor receptor* from the previous sentence. This is a classic case of anaphoric coreference. In sentence B, *The complex* is a metonymic reference to the multi-protein complex created during the binding event of the previous sentence. Semantic interpretation is required to infer the existence of this multi-protein entity, so that it can be linked to *the complex* (see also the examples in Figure 2). In sentence C, we have an instance of coreference between an entity and an entire event; i.e. *The association* refers to the *binding* event of the preceding sentence. Finally, in the example sentence D, the two event triggers *combine* and *bind* represent synonyms and constitute an example of event coreference.

The BioNLP Shared Task series focus on biomolecular events in scientific literature. They consist of event extraction tasks along with other tasks, which bear the imprint of the trend from core event extraction [38] to coreference [119] and pathway curation [28] (see BioNLP detail in 'Challenges and corpora for event extraction'). As it was shown that lack of coreference resolution significantly hindered the event extraction performance [38], BioNLP'11 [24] organized a supporting task of identifying coreferential relations between proteins/genes. With all participating systems favouring precision [119], the best performing solution [137], a modified system from Reconcile [138] based on supervised machine learning, obtained 73.7% precision with 22.2% recall (F-score 34.1%).

Available systems for biomedical coreference resolution focus on the correspondence between entities and their various expressions in text (A and B in Figure 4), but little attention has been placed on coreference that involves mentions of events (C and D in Figure 4). There have been both rule-based [133, 139, 140], machine learning [141, 142] and hybrid approaches [136] to coreference. Miwa *et al*. [133] analysed anaphoric coreferences with the help of a set of manually created rules that have been determined through parsing the Protein Coreference (COREF) task training data from BioNLP'11. This solution achieved slightly lower precision (69.8%) than the best system when being evaluated against the BioNLP'11 corpus. Its

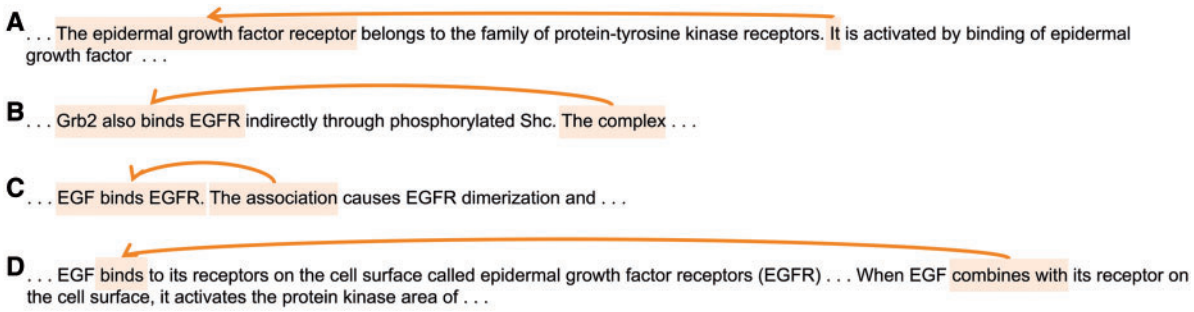


**Figure 4:** Examples of anaphoric coreference. The arch in (**A**) links the pronoun, *it*, with the noun phrase, *The epidermal growth factor receptor*, in the preceding sentence. The example in (**B**) cannot be resolved by syntactic approaches to anaphora resolution. *The complex* is a metonymic reference to the multi-protein complex created during the binding event of the previous sentence. Semantic interpretation is required to infer the existence of this multi-protein entity, so that it can be linked to *the complex* (see also the examples in Figure 2). In (**C**), *The association* does not refer to a previous mention of an entity or a product of an event. It refers to the previous event in its entirety, namely, *EGF binds EGFR*. In (**D**), the second event, *combines with*, refers to the previous event, signalled by the trigger *bind*.

improved recall (53.5%) led to a high F-score performance (60.5%). Yoshikawa *et al.* [141] extracted events between frequently mentioned entities, then inferred the coreferential relations based on transitivity. It tested two coreference-based models: a pipeline based on SVM classifiers, and a joint Markov Logic Network (MLN). In the evaluation against the BioNLP'09 event extraction corpus, both models achieved better performance when applying coreference resolution in comparison with other systems. Although many systems classify pairwise coreference independently from entity mention detection, Song *et al.* [142] proposed a joint learning model with Markov logic, which combined both processes and outperformed other machine learning systems in the CoNLL-2011 shared task [143]. The best overall performance at the CoNLL-2011 shared task was by a rule-based system, the Stanford multi-pass sieve [140], which is a collection of deterministic rules incorporating lexical, syntactic, semantic and discourse information. Lee *et al.* [136] take this work further by addressing coreferential relations between both entities and events (see an example of the latter in Sentence D in Figure 4) simultaneously. They first cluster documents, then extract mentions of both noun and verb phrases within the same document cluster and treat each mention as a singleton cluster. After the application of rule-based filters for entity resolution from [140], they iteratively merge singleton clusters, and a linear regressor, trained on gold coreference labels, indicates the best merge at each stage. Finally, another component from [140] is used to resolve pronominal coreference.

## Challenges and corpora for event extraction

BioCreAtIvE is one of the challenges that have been well recognized and supported by the TM community. Besides the gene mention and gene normalization tasks, BioCreAtIvE has also benchmarked the functional annotation of gene products (BioCreAtIvE I, task 2) including the identification of the biological role of molecules. The subsequent BioCreAtIvE II protein–protein interaction task [15] organized by IntAct [144], a molecular interaction database, and MINT [145], an experimentally verified protein–protein interaction database, contained four sub-tasks. These include *Protein Interaction Article Sub-task 1 (IAS)*, *Protein Interaction Pairs Sub-task 2 (IPS)*, *Protein Interaction Sentences Sub-task 3 (ISS)* and *Protein Interaction Method Sub-task 4 (IMS)*. IAS tests a system's accuracy to spot articles containing PPI descriptions, while IPS assesses the system's ability to extract mentioned PPI in an article. ISS further evaluates whether the system is able to pull all related sentences for a specified PPI pair. To complement the output pairs from systems in ISS with interaction type, IMS was designed to test whether a system can extract description from text for a given interaction type. In 2009, BioCreAtIvE III's PPI task mainly focused on the detection of relevant articles and linking articles to experimental methods.

BioNLP'09 [38] consisted of three challenges: *Core event extraction (GE)*, *Event enrichment* and *Negation and speculation recognition*. BioNLP'11 [24] retained *GE* as the main task, which covered interactions of gene and gene products including *Gene expression*, *Transcription*, *Protein catabolism*,

*Phosphorylation*, *Localization*, *Binding*, *Regulation*, *Positive regulation* and *Negative regulation*. BioNLP'11 added another three main tasks: *Epigenetics and Post-translational Modifications (EPI)*, *Infectious Diseases (ID)*, and *Bacteria Track (BB and BI)*. Except for the main tasks, BioNLP'11 offered three supporting tasks, which are *coreference (CO)*, *entity relations (REL)* and *gene renaming (REN)*. *REL* was for the detection of relations between an entity and its related entity, e.g. a protein and the produced multi-protein complex. The related entities could be metonyms of a protein domain or multi-protein complex. Associating them with entities is helpful for identifying reacting entities and product entities, so as to extract the correct topology and reaction order of a network (refer section 'Event interconnection'). *REN* concerned the detection of bacteria gene homonymies and synonymies. BioNLP'11 corpus included additional five full-text articles to the BioNLP'09 dataset. The recently announced BioNLP'13 [28] consists of six tasks. The GE (Genia event extraction) task aims to construct a knowledge base by combining TM and semantic web technologies, which also demands systems to be able to tackle co-reference. CG (Cancer Genetics) concerns the oncological events in biomedical literature. PC (Pathway Curation) tries to investigate the current TM capability for achieving 'automatic pathway extraction', and the potential to apply current event extraction solutions for pathway curation. GRO (Corpus Annotation with Gene Regulation Ontology) provides a corpus to be annotated with the Gene Regulation Ontology to tackle the problems existing in semantic search. GRN (Gene Regulation Network in Bacteria) attempts to evaluate the applicability of TM in bacterial gene regulation network extraction. BB (Bacteria Biotopes) identifies the natural location of bacteria. In summary, BioNLP tasks are moving from fundamental event extraction to providing support for systematic analysis of biological networks.

Another widely used biomedical corpus is the AIMed corpus, which consists of 225 abstracts, within which 25 are not directly describing specific interactions. With more and more available corpora, work on comparing them [146] and work on performing cross-corpora evaluation has gained appreciation, as each corpus has its own emphasis and systems trained on one corpus may not necessarily port well to a different one.

The databases of molecular interactions [144, 147, 148] cross-reference database entries with corresponding publications. Corpora for specific topics can be generated based on thematic clusters of publication. The semantic resources or ontologies of molecular interactions [61, 145, 149–154] provide substantial meta-data, in which verbs could potentially populate trigger sets, and interaction types are useful for determining or ranking an extracted relation, for example, insulin can never be involved with an interaction within a cell.

Each corpus, as discussed above, has been designed for a particular task, e.g. PPI. Evaluation against these corpora gives an indication of the capability of a TM system to address the particular task in question. However, good performance on one corpus does not automatically guarantee the same level of performance on other corpora. A system, trained on a specific corpus in the context of a different corpus, needs to consider running a domain adaptation task [155]. Another important consideration is that many systems are designed for or trained on scientific abstracts, which does not guarantee equivalent performance on full articles. Indeed, abstracts contain condensed information, but it is the case that most detailed bio-events and information pertaining to networks is found in the body of the articles, which also contains more noise [156]. Several approaches have combined filters to restrict the types of relations and the location of events to consider for extraction [157, 158]. Zoning of scientific articles in the life sciences such as in the case of [31] could also provide a means of filtering and reducing the noise in full articles for the purpose of event extraction.

## NETWORK CONSTRUCTION

Extracted events can reflect a molecule's relationship with the ones it is interacting with. However, such interactions have to be reviewed in the context of a network to understand their role in a biological process. A network needs to be topologically accurate with consecutive interactions.

### Event interconnection

Although TM so far has been focusing on event extraction, the potential use of TM in event interconnection is slowly emerging through the notion of event coreference [136, 142]. Slowly but steadily

TM is moving from event extraction to network extraction, climbing up the pyramid in Figure 1.

Intuitively, events could be connected together based on common entities. However, solely depending on the common entities would require flawless entity coreference and normalization. Apart from the diversity of entity names, for example, it is not uncommon that a generic term is used as a metonymy of another mentioned term, e.g. *glucose* versus *alpha-D-glucose*. In this case, entities can be heuristically merged according to semantic similarity. This method is also effective for merging similar reactions [136, 142, 143, 159].

In BioNLP [24, 38], an event could appear as a *Theme* or a *Cause* of another event. This reflects the succession and recursion of a biological process. Some approaches statistically link co-referring events together. FAUN (Feature Annotation Using Nonnegative matrix factorization) [48] extracts gene term features and constructs them into a matrix. By using different features and weighing each term, the authors treat it as a clustering problem of functionally related genes, addressed by matrix factorization with non-negative normalization (NNMF) in FAUN. Bio-LDA (Latent Dirichlet allocation) extracted 13 338 terms for compound, gene, drug, disease, pathways and side effects. Relationships between concepts were hypothesized based on a LDA model [47].

An important aspect of bio-events that causes the problem for entity identification and coreference, and subsequently network extraction, is the fact that many entities (products) resulting from a particular biochemical reaction are omitted or implied in the text and only later referred to, usually by referential metonymy. For example, an issue resulting from this is the difficulty in recognizing multi-protein complexes (refer section 'Signalling pathways'), which are not specifically named or introduced in the text (e.g. during a binding event) and may be referred to using the name of the proteins they contain. These also point to weaknesses in the usage of the nomenclature, especially for multi-protein complexes, as generated complexes are functionally different entities, but are referred to by the name of the receptor protein. By contrast, products of chemical reactions are unambiguously named. Without properly identifying or inferring the products of biochemical reactions, the generated incomplete network cannot faithfully reflect the mechanism of a pathway, and molecular roles in the pathway.

## Systems for network extraction

To our knowledge, few systems have the capability to extract networks of signalling and metabolic pathways from the literature. Some systems extract related events based on specific entities and connect the events into a network on the basis of common entities [101, 103]. Molecular interactions can be catered for by this kind of system, but the sequence of interactions remain relatively unclear. PathwayFinder [18] extracts pathways with the help of user intervention. During extraction, it provides an interface through which a user can influence the result of NER and modify syntactic patterns. GENIES [17] is the only system that is reported to automatically extract a pathway from a full-text article. It uses a grammar consisting of semantic patterns interleaved with syntactic and semantic constraints to identify relevant relationships and to specify target outputs.

Here, we review the systems using state-of-the-art TM (Table 1), which are designed to collect evidence of molecular interactions or networks to support biologists in routine curation tasks.

## Challenges in network extraction evaluation

Evaluation of network extraction is much more complex than that of event extraction. In the case of network extraction from a single article, one can observe that reactions relevant to the network are scattered across the entire text. The article may also contain redundant mentions of interactions, ambiguous syntactic structures and interactions that make sense from a linguistic point of view but do not bear any significant biological knowledge. For example, in the sentence '*D-ribose*, *creatine* and *calcium* are combined with *vitamin D* as a health', 'combined with' means that the manufacturer mixed them in one intake to provide general healthy support. Without semantic disambiguition, the previous can be interpreted as a binding event between organic compounds and other nutrient. A network would need to be synthesized from individual interlinked events. This is even more important in the case of network extraction from a set of articles. Filtering and paper zoning, mentioned in section 'Challenges and corpora for event extraction', could help in locating the events to be synthesized into a network.

So far, no gold standard corpus exists for the evaluation of network extraction solutions. Systems have been assessed individually and independently,

**Table 1:** Selected systems for discovering relations/networks from literatures

| Systems | Features and limitations | Availability |
|---|---|---|
| Relation discovery | | |
| Arrowsmith[a] | Discovers relations between two sets of MEDLINE articles. It has ~1200 unique users monthly. However, only the titles of two sets of articles are analysed [160]. | Free-access at http://arrowsmith.psych.uic.edu/arrowsmith_uic/ |
| Chilibot[b,c] | Pair-wise search by rules on shallow parsed (CASS) MEDLINE to discover entity interactions. It was used for collecting evidences for studying gene–cancer association [161, 162]. It only provides three types of relations—positive, negative and neutral. | Free-access at http://www.chilibot.net/ |
| EBIMed[a] | Extracts related protein/gene and their interactions from MEDLINE based on co-occurrence [163]. It has been integrated into the Whatizit pipeline, and used for constructing disease knowledge base [164]. It is limited to its pre-processed documents. | Free-access at http://www.ebi.ac.uk/Rebholz-srv/ebimed |
| EventMine[c,d] | Extracts biomedical events based on syntactic analysis by machine learning [125]. The new version includes a rule-based coreference detection component. A modified version was used for extracting exhaustive protein modification event [165]. It was recently used to support scientific discourse annotation [166]. Public service website only has a demonstration with limited usage allowance (~100 requests daily). | Demonstration available at http://www.nactem.ac.uk/EventMine/demo.php |
| FACTA+[c,d] | Browses and visualizes discovered relations and indirect associations. It can recognize trigger word and decide interaction type based on a model trained in BioNLP'09 shared task data [167]. It is used by PathText [20] to link molecular interaction in pathway to literatures. It limits the input file to contain maximum 7 gene lists, and each list contains up to 100 genes. | Free-access at http://refine1-nactem.mc.man.ac.uk/facta/ |
| FAUN | Discover gene functional relationship based on nonnegative matrix factorization. It was used to classify functional genes related with autism [48]. It is limited to documents cross-linked with entries in Entrez Gene. | Demonstration available at https://grits.eecs.utk.edu/faun. Account is created upon request |
| iHOP[a] | Phenotype, pathology and function study tool based on network of co-occurring gene/protein [39]. It reportedly had 500 000 visits monthly, and now provides programmatical access [167]. It is limited to its pre-processed documents. | Free-access at http://www.ihop-net.org/ |
| LAITOR[a] | Co-occurrence-based interaction extraction tool. It is used by PESCADOR [21] as relation extraction component. Using it requires high computational expertise, and it is limited to its pre-processed documents. | Available at http://laitor.sourceforge.net under General Public License (GPL) |
| LitInspector[a] | Searches gene and signal transduction pathway within NCBI's PubMed database based on co-occurrence [19]. It was used for supporting the annotation evaluation in BiblioSphere [169]. | Free-access at http://www.litinspector.org/ |
| LitMiner[a] | Searches genes and their relationships based on co-occurrences [170]. Its java application needs to be downloaded, and it has no web interface. The homepage (http://www.litminer.com) was out of service when being tested. | Available at http://www.litminer.com |
| PathText[c,d] | Integrates MEDIE, FACTA and KLEIO for searching related articles for interactions in systems biology model. It is not available as an independent software. | Available from http://www.pathtext.org as plug-in for CellDesigner [171] and Payao Web 2.0 [172] |
| Protein corral[a,b] | Co-occurrence and rule based PPI extraction system. It is integrated into the Whatizit pipeline. It emphasizes precision over recall. | Free-access at http://www.ebi.ac.uk/Rebholz-srv/pcorral/ |
| Network discovery | | |
| CoPub[a] | Searches a pair/set of terms upon their co-occurrence in annotated MEDLINE. CoPub is used to discover indirect relations between biological concepts based on transitivity [104]. Its search is limited to given keywords, and human, mouse or rat genes only. | Free-access http://services.nbic.nl/copub/portal/ |
| GENIES[b,c] | Extracts pathway from literature by using a set of grammars to analyse syntactic parsing results. It is integrated into GeneWay [173]. It is not freely publicly available. | Not freely available and its patent application was filed (http://www.cat.columbia.edu/genies.licensing.htm) |
| PESCADOR[a] | Extracts interaction network of given entities. It uses LAITOR as the text-mining engine. It was used for molecular interaction analysis [174] and interaction collection for database population [175]. | Free-access at http://cbdm.mdc-berlin.de/tools/pescador/ |
| PubGene[a] | Co-occurrence-based microarray analysis tool. It has been commercialized to a product. | http://www.pubgene.com/ |
| RelEx[b,c] | Relation extraction tool based on syntactic parsing and rules. It was used for interaction analysis, such as those between host–pathogen [176]. It has relatively low recall. | Test set and relation term lists available from http://www.bio.ifi.lmu.de/publications/RelEx/ |

[a]Systems that use co-occurrence of entity pair to extract relations. [b]Systems that use pattern-based approach. [c]Systems that use syntactic parsing to extract relations. [d]Systems that use machine learning approach.

and according to different benchmarks. The closest one can get to a gold standard at the moment is to extract networks from a set of articles, which have been manually linked to biological entities in a curated database. The extracted entities and reactions can then be cross-evaluated against entries in the curated databases. Some interaction databases contain manually curated networks and have the potential to support related work in TM. For example, each reaction network in BMDB is linked to a single publication. As a molecular interaction database, IntAct [144] provides a corpus [177] with sentential evidence of interactions along with publication identifiers and IntAct accession numbers.

It should be noted that the pathway curation task (PC) recently proposed by BioNLP'13 [28] is still an event extraction task, involving entities and relations encountered in biological pathways. Evaluation in this case follows the same guidelines as previous event extraction tasks. Owing to information sparsity and technology limitations, events extracted by current systems are far from being complete bio-events, which would include not only reactants, but also products, even catalyst. These obligatory parts of bio-event may need inference, but are necessary for interconnecting events, which reflects the sequence of reactions in the biological pathway and is more informative for pathway curation.

## CONCLUSION

The study of molecular role and function is key to revealing causative mechanisms of disease. Such investigations frequently depend on distinguishing the interactions involved into pathways. These pathways are often represented as qualitative logical models. In living organisms, the interaction between molecules is not simply an on/off relation, which calls for the determination of experimental conditions and parameters in these network models. Such quantitative modelling provides a more realistic depiction of the interaction networks in complicated vital processes. In this article, we have reviewed the process of network extraction using TM techniques, from entity recognition to event and network extraction. There has been less emphasis on the quantitative aspect, which can be beyond NLP-based TM, as it requires the extraction of information from tables, figures and equations.

We have considered two types of networks, signalling pathways and metabolic pathways. Although in principle the mechanism of network extraction is the same (entity recognition followed by event extraction and subsequently event interconnection), in practice there are different challenges, as entity recognition has focused on proteins and genes and has not reached the same level of maturity for entities such as chemicals and enzymes, which feature in metabolic pathways. A study of the distribution of reactions from BMDB in the literature has also shown a disproportionate relative coverage of these networks within the literature itself, with more information available for signalling pathways.

Even for signalling pathways, NER faces a number of challenges. These issues also have significant impact on entity coreference resolution. Better handling of multi-word expressions and co-ordination structures would improve NER. Another important issue is that of incomplete nomenclature; many protein complexes are metonymically referenced by the name of the protein they contain. Also, proteins may be referred to by means of their nucleotide residues. This leads to incorrect entity normalization. Strict naming for such protein and protein complexes would alleviate this problem and reduce instances of referential metonymy. This generally benefits protein-related molecular biology research rather than TM only, because protein and interaction database can unambiguously annotate entities, and support automated analysis upon the databases [178]. Another problem is the incompleteness of existing data resources, which makes it difficult to normalize to the appropriate concepts that entities refer to. Furthermore, there is the recognition of the products of a biochemical reaction, which are often not explicitly mentioned in the context of the reaction, but are rather implied and referenced later in the text. This phenomenon is also especially problematic for entity coreference and subsequent interlinking of extracted events.

Event extraction involves the recognition of particular event types/relations and their participants. There is only a limited number of relations considered at the moment and this reflects a gap in the available knowledge resources and corpus annotations. Moreover, a major challenge in event extraction is the resolution of complex and embedded events, spanning more than a single sentence, which requires access to long-range dependencies within the discourse. State of the art systems for event extraction are beginning to make the resolution of complex events possible via joint models for recognition of triggers of event types and event

arguments, which also take into account the dependency paths. Both entity and event coreference resolution are also becoming aspects of current event extraction systems and more effort needs to be focused in this direction, especially in the case of verb coreference, which can help solve complex cases of entity coreference and determine the directionality and polarity of an extracted relation.

It is important to make use of the large amount of experimental data to help the analysis of molecular interactions and pathways. Combining such information with evidence from the literature and using the literature to update and even create useful models of reactions in knowledge resources is key for this process. We argue that a number of factors need to be combined to facilitate the extraction of such information from the literature. Some of these rely on improving TM technologies such as coreference. However, it is equally important to widen the use of existing ontological and database resources so that entities found in literature can be normalized. To this end, it is important to extend such knowledge resources to obtain a better coverage of the scientific literature.

**Key Points**

- To collect informative evidence and support the discovery of hidden relations for disease and pharmaceutical research, biomedical TM is required to address network extraction and quantitative model extraction, which can be distinguished into six separate phases: entity mention, entity normalization, relation extraction, event extraction, qualitative model extraction and finally quantitative model extraction.
- The characteristics of different biological networks, e.g. signalling pathways and metabolic pathways, necessitate adjustments in various aspects of existing TM solutions.
- Currently, many TM researchers focus on bio-event extraction, and many systems use syntactic parsing to obtain linguistic patterns as well as machine learning in their approaches. Systems using joint models of learning, which address more than one task at the same time show promise and are gaining popularity.
- To achieve network extraction from the literature, TM researchers face challenges in entity normalization, referential metonymy, coreference and event interconnection. Although some of these can be addressed with advances in TM technologies, stricter naming conventions on the part of experts and more complete knowledge resources would be of significant benefit.

## FUNDING

Chen Li is funded by the Cambridge Overseas Trust and the European Molecular Biology Laboratory (EMBL-EBI). Maria Liakata is funded by the Leverhulme Trust and EMBL-EBI.

## *References*

1. Kanehisa M, Goto S, Hattori M, *et al*. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res* 2006;**34**:D354–7.
2. D'Eustachio P. Reactome knowledgebase of human biological pathways and processes. *Methods Mol Biol* 2011;**694**: 49–61.
3. Pathway Interaction Database. http://pid.nci.nih.gov/ (13 February 2013, date last accessed).
4. Hodgkin AL, Huxley AF. A quantitative description of membrane current and its application to conduction and excitation in nerve. *J Physiol* 1952;**117**:500–44.
5. Hucka M, Finney A, Sauro HM, *et al*. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* 2003;**19**:524–31.
6. Lloyd CM, Halstead MDB, Nielsen PF. CellML: its future, present and past. *Prog Biophys Mol Biol* 2004;**85**: 433–50.
7. Demir E, Cary MP, Paley S, *et al*. The BioPAX community standard for pathway data sharing. *Nat Biotechnol* 2010;**28**: 935–42.
8. Li C, Donizelli M, Rodriguez N, *et al*. BioModels database: an enhanced, curated and annotated resource for published quantitative kinetic models. *BMC Syst Biol* 2010; **4**:92.
9. Lloyd CM, Lawson JR, Hunter PJ, *et al*. The CellML model repository. *Bioinformatics* 2008;**24**:2122–3.
10. WikiPathways. http://www.wikipathways.org (13 February 2013, date last accessed).
11. Li C, Courtot M, Le Novère N, *et al*. BioModels.net Web Services, a free and integrated toolkit for computational modelling software. *Brief Bioinformatics* 2010;**11**:270–7.
12. Ohta T, Pyysalo S. From pathways to biomolecular events: opportunities and challenges. In: *Proceedings of the 2011 Workshop on Biomedical Natural Language Processing, ACL-HLT 2011, Portland, OR, June 2011*. Portland, OR, USA: Association for Computational Linguistics; 105–113.
13. Jensen LJ, Saric J, Bork P. Literature mining for the biologist: from information retrieval to biological discovery. *Nat Rev Genet* 2006;**7**:119–29.
14. Krallinger M, Erhardt RAA, Valencia A. Text-mining approaches in molecular biology and biomedicine. *Drug Discov Today* 2005;**10**:439–45.
15. Krallinger M, Leitner F, Rodriguez-Penagos C, *et al*. Overview of the protein-protein interaction annotation extraction task of BioCreative II. *Genome Biol* 2008; **9(Suppl 2)**:S4.
16. Ananiadou S, Pyysalo S, Tsujii J, *et al*. Event extraction for systems biology by text mining the literature. *Trends Biotechnol* 2010;**28**:381–90.
17. Friedman C, Kra P, Yu H, *et al*. GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics* 2001;**17(Suppl 1)**:S74–82.
18. Yao D, Wang J, Lu Y, *et al*. PathwayFinder: paving the way towards automatic pathway extraction. In: *Proceedings of the 2nd Asia-Pacific Bioinformatics Conference (APBC2004), Dunedin, New Zealand,* 2004. Australian Computer Society, Inc. Darlinghurst, Australia.

19. Frisch M, Klocke B, Haltmeier M, *et al*. LitInspector: literature and signal transduction pathway mining in PubMed abstracts. *Nucleic Acids Res* 2009;**37**(Web Server issue): W135–40.
20. Kemper B, Matsuzaki T, Matsuoka Y, *et al*. PathText: a text mining integrator for biological pathway visualizations. *Bioinformatics* 2010;**26**:i374–81.
21. Barbosa-Silva A, Fontaine JF, Donnard ER, *et al*. PESCADOR, a web-based tool to assist text-mining of biointeractions extracted from PubMed queries. *BMC Bioinformatics* 2011;**12**:435.
22. Yeh A, Morgan A, Colosimo M, *et al*. BioCreAtIvE task 1A: gene mention finding evaluation. *BMC Bioinformatics* 2005; **6(Suppl 1)**:S2.
23. Verspoor K, Cohen KB, Lanfranchi A, *et al*. A corpus of full-text journal articles is a robust evaluation tool for revealing differences in performance of biomedical natural language processing tools. *BMC Bioinformatics* 2012;**13**:207.
24. Kim JD, Pyysalo S, Ohta T, *et al*. Overview of BioNLP shared task 2011. In: *Proceedings of BioNLP Shared Task 2011 Workshop, Portland, OR, June 24, 2011, pp. 1–6*. Portland, OR, USA: Association for Computational Linguistics.
25. Spasić I, Simeonidis E, Messiha H. KiPar, a tool for systematic information retrieval regarding parameters for kinetic modelling of yeast metabolic pathways. *Bioinformatics (Oxford, England)* 2009;**25**:1404–11.
26. Heinen S, Thielen B, Schomburg D. KID–an algorithm for fast and efficient text mining used to automatically generate a database containing kinetic information of enzymes. *BMC Bioinformatics* 2010;**11**:375.
27. Tsay JJ, Wu BL, Hsieh CC. Automatic extraction of kinetic information from biochemical literatures. *2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery, Tianjin, China, 2009.* IEEE; 28–32. IEEE Press Piscataway, NJ, USA.
28. BioNLP Shared Task 2013. http://2013.bionlp-st.org/ (13 February 2013, date last accessed).
29. Benagiano V, Lorusso L, Flace P, *et al*. VAMP-2, SNAP-25A/B and syntaxin-1 in glutamatergic and GABAergic synapses of the rat cerebellar cortex. *BMC Neurosci* 2011;**12**:118.
30. Miwa M, Thompson P, McNaught J, *et al*. Extracting semantically enriched events from biomedical literature. *BMC Bioinformatics* 2012;**13**:108.
31. Liakata M, Saha S, Dobnik S, *et al*. Automatic recognition of conceptualisation zones in scientific articles and two life science applications. *Bioinformatics* 2012;**28**:991–1000.
32. Markert K, Hahn U. On the interaction of metonymies and anaphora. In: *Proceedings of International Joint Conferences on Artificial Intelligence, Nagoya, Japan, 1997*;1010–15, Morgan Kaufmann Publishers, San Francisco, USA.
33. Stallard D. Two kinds of metonymy. In: *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics, Ohio State University, Columbus, Ohio, USA, 1993.* Association for Computational Linguistics; 87–94, Association for Computational Linguistics Stroudsburg, PA, USA.
34. Chun HW, Tsuruoka Y, Kim JD, *et al*. Extraction of gene-disease relations from MEDLINE using domain dictionaries and machine learning. In: *Proceedings of the Pacific Symposium on Biocomputing (PSB) 11, January 2006*. Maui, HI, USA; 4–15, Pac Symp Biocomput.
35. Hakenberg J, Plake C, Leaman R, *et al*. Inter-species normalization of gene mentions with GNAT. *Bioinformatics* 2008;**24**:i126–32.
36. Gorgoulis V, Aninos D, Mikou P, *et al*. Expression of EGF, TGF-alpha and EGFR in squamous cell lung carcinomas. *Anticancer Res* 1992;**12**:1183–7.
37. Kholodenko BN, Demin OV, Moehren G, *et al*. Quantification of short term signaling by the epidermal growth factor receptor. *J Biological Chem* 1999;**274**: 30169–81.
38. Kim JD, Ohta T, Pyysalo S, *et al*. Overview of BioNLP'09 shared task on event extraction. In: *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task, Boulder, CO, June 5, 2009*. Boulder, CO, USA: Association for Computational Linguistics; 1–9.
39. Hoffmann R, Krallinger M, Andres E, *et al*. Text mining for metabolic pathways, signaling cascades, and protein networks. *Sci STKE* 2005;**2005**:pe21.
40. BioModels Database Curation Guideline. http://www.ebi.ac.uk/biomodels-main/curationtips (13 February 2013, date last accessed).
41. Rebholz-Schuhmann D, Arregui M, Gaudan S, *et al*. Text processing through Web services: calling Whatizit. *Bioinformatics* 2008;**24**:296–8.
42. UniProt Consortium. Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res* 2012;**40**(Database issue):D71–5.
43. de Matos P, Alcántara R, Dekker A, *et al*. Chemical entities of biological interest: an update. *Nucleic Acids Res* 2010; **38**(Database issue):D249–54.
44. MEDLINE/PubMed. http://www.ncbi.nlm.nih.gov/pubmed/.
45. Bekhuis T. Conceptual biology, hypothesis discovery, and text mining: Swanson's legacy. *Biomed Digit Libr* 2006; **3**:2.
46. Saez-Rodriguez J, Alexopoulos LG, Epperlein J, *et al*. Discrete logic modelling as a means to link protein signalling networks with functional analysis of mammalian signal transduction. *Mol Syst Biol* 2009;**5**:331.
47. Wang HJ, Ding Y, Tang J, *et al*. Finding complex biological relationships in recent PubMed articles using Bio-LDA. *PLoS One* 2011;**6**:e17243.
48. Tjioe E, Berry MW, Homayouni R. Discovering gene functional relationships using FAUN (Feature Annotation Using Nonnegative matrix factorization). *BMC Bioinformatics* 2010;**11(Suppl 6)**:S14.
49. Tari L, Anwar S, Liang S, *et al*. Discovering drug-drug interactions: a text-mining and reasoning approach based on properties of drug metabolism. *Bioinformatics* 2010;**26**: i547–53.
50. Hirschman L, Burns GAPC, Krallinger M, *et al*. Text mining for the biocuration workflow. *Database* 2012;;**2012**:bas020.
51. Thompson P, McNaught J, Montemagni S, *et al*. The BioLexicon: a large-scale terminological resource for biomedical text mining. *BMC Bioinformatics* 2011;**12**:397.
52. Kim JJ, Rebholz-Schuhmann D. Improving the extraction of complex regulatory events from scientific text by using ontology-based inference. *J Biomed Semant* 2011; **2(Suppl 5)**:S3.
53. Tomanek K, Hahn U. Semi-supervised active learning for sequence labeling. In: *Proceedings of the 47th Annual Meeting of*

*the ACL and the 4th IJCNLP of the AFNLP, Suntec, Singapore, August 2–7, 2009*. Suntec, Singapore: Association for Computational Linguistics.

54. McCallum A, Freitag D. Maximum entropy Markov models for information extraction and segmentation. In: *Proceedings of the 17th International Conference on Machine Learning, Stanford, CA, USA,* 2000. Morgan Kaufmann Publishers, San Francisco, USA.
55. McDonald R, Pereira F. Identifying gene and protein mentions in text using conditional random fields. *BMC Bioinformatics* 2005;**6(Suppl 1)**:S6.
56. Settles B. ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. *Bioinformatics* 2005;**21**:3191–2.
57. Hsu CN, Chang YM, Kuo CJ, *et al*. Integrating high dimensional bi-directional parsing models for gene mention tagging. *Bioinformatics* 2008;**24**:i286–94.
58. Sasaki Y, Tsuruoka Y, McNaught J, *et al*. How to make the most of NE dictionaries in statistical NER. *BMC Bioinformatics* 2008;**9(Suppl 11)**:S5.
59. Saha S, Ekbal A, Saha S. A supervised approach for gene mention detection. *Swarm Evol Memetic Comput* 2011;**7076**: 425–32.
60. Shi L, Campagne F. Building a protein name dictionary from full text: a machine learning term extraction approach. *BMC Bioinformatics* 2005;**6**:88.
61. Kaiser J. Proteomics. *Public-private group maps out initiatives.* Science 2002;**296**:827.
62. Neves ML, Carazo JM, Pascual-Montano A. Moara: a Java library for extracting and normalizing gene and protein mentions. *BMC Bioinformatics* 2010;**11**:157.
63. Huang M, Liu J, Zhu X. GeneTUKit: a software for document-level gene normalization. *Bioinformatics* 2011;**27**: 1032–3.
64. Chen Y, Liu F, Manderick B. A machine learning-based system to normalise gene mentions to unique database identifiers. *Int J Data Mining Bioinformatics* 2011;**5**: 640–60.
65. NCBI Entrez Gene. http://www.ncbi.nlm.nih.gov/gene.
66. Kolárik C, Klinger R, Friedrich CM, *et al*. Chemical names: terminological resources and corpora annotation. *Workshop on Building and evaluating resources for biomedical text mining (6th edition of the Language Resources and Evaluation Conference), Marrakech, Morocco, 2008*; 51–58, Springer Publishing, USA.
67. Corbett P, Murray Rust P. Highthroughput identification of chemistry in life science texts. In: *Proceedings of the 2nd International Symposium on Computational Life Science (CompLife 06), Cambridge, UK, 2006*; 107–18, Springer Publishing, USA.
68. Jessop DM, Adams SE, Willighagen EL, *et al*. OSCAR4: a flexible architecture for chemical text-mining. *J Cheminform* 2011;**3**:41.
69. Rocktäschel T, Weidlich M, Leser U. ChemSpot: a hybrid system for chemical named entity recognition. *Bioinformatics* 2012;**28**:1633–40.
70. Hirschman L, Yeh A, Blaschke C, *et al*. Overview of BioCreAtIvE: critical assessment of information extraction for biology. *BMC Bioinformatics* 2005;**6(Suppl 1)**:S1.
71. Smith L, Tanabe LK, Ando RJ, *et al*. Overview of BioCreative II gene mention recognition. *Genome Biol* 2008;**9(Suppl 2)**:S2.
72. Morgan AA, Lu Z, Wang X, *et al*. Overview of BioCreative II gene normalization. *Genome Biol* 2008; **9(Suppl 2)**:S3.
73. Blaschke C, Leon EA, Krallinger M, *et al*. Evaluation of BioCreAtIvE assessment of task 2. *BMC Bioinformatics* 2005;**6(Suppl 1)**:S16.
74. Rebholz-Schuhmann D, Yepes AJ, Li C, *et al*. Assessment of NER solutions against the first and second CALBC Silver Standard Corpus. *J Biomed Semantics* 2011;**2(Suppl 5)**:S11.
75. Kim JD, Ohta T, Tsuruoka Y, *et al*. Introduction to the bio-entity recognition task at JNLPBA. Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications, *JNLPBA '04*, Stroudsburg, PA, USA, 2004. Stroudsburg, PA: Association for Computational Linguistics; 70–75.
76. Kulick S, Bies A, Liberman M, *et al*. Integrated annotation for biomedical information extraction. *Proceeding of the Human Language Technology Conference and the Annual Meeting of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL), 2004*Association for Computational Linguistics; 61–68.
77. Buyko E, Beisswanger E, Hahn U. The GeneReg corpus for gene expression regulation events an overview of the corpus and its in-domain and out-of-domain interoperability. In: *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), May 2010, Valletta, Malta*. Valletta, Malta: European Language Resources Association (ELRA).
78. European Patent Office and ChEBI annotated chemical corpus. http://chebi.cvs.sourceforge.net/viewvc/chebi/chapati/patentsGoldStandard/ (13 February 2013, date last accessed).
79. Bada M, Eckert M, Evans D, *et al*. Concept annotation in the CRAFT corpus. *BMC Bioinformatics* 2012;**13**:161.
80. World Wide Protein Data Bank. http://www.wwpdb.org/ (13 February 2013, date last accessed).
81. Punta M, Coggill PC, Eberhardt RY, *et al*. The Pfam protein families database. *Nucleic Acids Res* 2012;**40**(Database issue):D290–301.
82. Wang Y, Xiao J, Suzek TO, *et al*. PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res* 2009;**37**(Web Server):W623–33.
83. CAS Registry. http://www.cas.org/content/chemical-substances (13 February 2013, date last accessed).
84. Chemical category of MeSH. http://www.nlm.nih.gov/mesh/ (13 February 2013, date last accessed).
85. Reaxys. http://www.info.reaxys.com (13 February 2013, date last accessed).
86. Tipton K, Boyce S. History of the enzyme nomenclature system. *Bioinformatics* 2000;**16**:34–40.
87. Wishart DS, Knox C, Guo AC, *et al*. HMDB: a knowledgebase for the human metabolome. *Nucleic Acids Res* 2009;**37**(Database issue):D603–10.
88. KEGG Compound. http://www.genome.jp/kegg/compound/ (13 February 2013, date last accessed).
89. Gaulton A, Bellis LJ, Bento AP, *et al*. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res* 2011;**40**(Database issue):D1100–07.
90. Knox C, Law V, Jewison T, *et al*. DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res* 2011;**39**(Database issue):D1035–41.

91. KEGG Drug. http://www.genome.jp/kegg/drug/ (13 February 2013, date last accessed).
92. Washington NL, Haendel MA, Mungall CJ, *et al*. Linking human diseases to animal models using ontology-based phenotype annotation. *PLoS Biol* 2009;**7**: e1000247.
93. Oellrich A, Hoehndorf R, Gkoutos GV, *et al*. Improving disease gene prioritization by comparing the semantic similarity of phenotypes in mice with those of human diseases. *PloS One* 2012;**7**:e38937.
94. Ohta T, Pyysalo S, Tsujii J, *et al*. Open-domain anatomical entity mention detection. In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, Jeju, Republic of Korea, July 12, 2012*;27–36, Association for Computational Linguistics.
95. Pyysalo S, Ohta T, Miwa M, *et al*. Event extraction across multiple levels of biological organization. *Bioinformatics* 2012;**28**:i575–81.
96. Skusa A, Rüegg A, Köhler J. Extraction of biological interaction networks from scientific literature. *Brief Bioinformatics* 2005;**6**:263–76.
97. Jenssen TK, Laegreid A, Komorowski J, *et al*. A literature network of human genes for high-throughput analysis of gene expression. *Nat Genet* 2001;**28**:21–8.
98. Tsuruoka Y, Tsujii J, Ananiadou S. FACTA: a text search engine for finding associated biomedical concepts. *Bioinformatics* 2008;**24**:2559–60.
99. Barbosa-Silva A, Soldatos TG, Magalhães ILF, *et al*. LAITOR–literature assistant for identification of terms co-occurrences and relationships. *BMC Bioinformatics* 2010;**11**:70.
100. PubMed Central. http://www.ncbi.nlm.nih.gov/pmc/ (13 February 2013, date last accessed).
101. Chen H, Sharp BM. Content-rich biological network constructed by mining PubMed abstracts. *BMC Bioinformatics* 2004;**5**:147.
102. Frijters R, Heupers B, van Beek P, *et al*. CoPub: a literature-based keyword enrichment tool for microarray data analysis. *Nucleic Acids Res* 2008;**36**(Web Server issue): W406–10.
103. Fleuren WWM, Verhoeven S, Frijters R, *et al*. CoPub update: CoPub 5.0 a text mining system to answer biological questions. *Nucleic Acids Res* 2011;**39**(Web Server): W450–54.
104. Frijters R, van Vugt M, Smeets R, *et al*. Literature mining for the discovery of hidden connections between drugs, genes and diseases. *PLoS Comput Biol* 2010;**6**:e1000943.
105. Miyao Y, Tsujii J. Feature forest models for probabilistic HPSG Parsing. *Comput Linguist* 2008;**34**:35–80.
106. Hara T, Miyao Y, Tsujii J. Text, speech and language technology. *Trends in Parsing Technology* 2011, Vol. 43. Dordrecht; 257–75, Nancy Ide, Springer Publishing, USA.
107. Charniak E, Johnson M. Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, *ACL'05*, Stroudsburg, PA, 2005. Stroudsburg, PA, USA: Association for Computational Linguistics; 173–180.
108. McDonald R, Lerman K, Pereira F. Multilingual dependency analysis with a two-stage discriminative parser. In: *Proceedings of the 10th Conference on Computational Natural Language Learning (CoNLL-X), New York City, June 2006*. New York City: Association for Computational Linguistics; 216–220.
109. De Marneffe MC, MacCartney B. Generating typed dependency parses from phrase structure parses. *LREC, Genoa, Italy,* 2006. European Language Resources Association.
110. Sagae K, Tsujii J. Dependency parsing and domain adaptation with LR models and parser ensembles. In: *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007, Prague, June 2007*. Prague: Association for Computational Linguistics; 1044–1050.
111. Curran JR, Clark S. Linguistically motivated large-scale NLP with C&C and boxer. In: *Proceedings of the ACL 2007 Demo and Poster Sessions, Prague, June 2007*. Prague: Association for Computational Linguistics; 33–36.
112. McClosky D, Charniak E. Self-training for biomedical parsing. In: *Proceedings of ACL-08: HLT, Short Papers, volume Companion Volume, Columbus, OH*. Columbus, OH, USA: Association for Computational Linguistics, June 2008; 101–104.
113. Chiang JH, Yu HC. Literature extraction of protein functions using sentence pattern mining. *IEEE Trans Knowl Data Eng* 2005;**17**:1088–98.
114. Yangarber R, Grishman R, Tapanainen P, *et al*. Unsupervised discovery of scenario-level patterns for information extraction. In: *Proceedings of the sixth conference on Applied natural language processing, Morristown, NJ, 2000*. Morristown, NJ, USA: Association for Computational Linguistics; 282–289.
115. Yangarber R. Counter-training in discovery of semantic patterns. In: *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, Morristown, NJ, 2003*. NJ, USA: Association for Computational Linguistics Morristown; 343–350.
116. Uszkoreit H. Learning relation extraction grammars with minimal human intervention: strategy, results, insights and plans. *Computational Linguistics and Intelligent Text. . ., Berlin, Heidelberg, 2011*. Berlin, Heidelberg: Springer Berlin Heidelberg; 106–26.
117. Protein Corral. http://www.ebi.ac.uk/rebholz-srv/pcorral/ (13 February 2013, date last accessed).
118. Cohen KB, Verspoor K, Johnson HL, *et al*. High-precision biological event extraction: effects of system and of data. *Comput Intell* 2011;**27**:681–701.
119. Kim JD, Nguyen N, Wang Y, *et al*. The genia event and protein coreference tasks of the BioNLP shared task 2011. *BMC Bioinformatics* 2012;**13(Suppl 11)**:S1.
120. Kilicoglu H, Bergler S. Adapting a general semantic interpretation approach to biological event extraction. In: *Proceedings of BioNLP Shared Task 2011 Workshop, Portland, OR, June 2011*;173–82, Association for Computational Linguistics.
121. Liu H, Keselj V, Blouin C, *et al*. Subgraph Matching-Based Literature Mining for Biomedical Relations and Events. In: *Proceedings of the AAAI 2012 Fall Symposium on Information Retrieval and Knowledge Discovery in Biomedical Text, Arlington, VA, USA, 2012*, p. 0, Association for the Advancement of Artificial Intelligence.
122. Temkin JM, Gilder MR. Extraction of protein interaction information from unstructured text using a context-free grammar. *Bioinformatics* 2003;**19**:2046–53.

123. Miwa M, Saetre R, Miyao Y, *et al*. Protein-protein interaction extraction by leveraging multiple kernels and parsers. *Int J Med Inform* 2009;**78**:e39–46.

124. Kang N, van Mulligen EM, Kors JA. Comparing and combining chunkers of biomedical text. *J Biomed Inform* 2011; **44**:354–60.

125. Miwa M, Saetre R, Kim JD, *et al*. Event extraction with complex event classification using rich features. *J Bioinformatics Comput Biol* 2010;**8**:131–46.

126. Björne J, Heimonen J, Ginter F, *et al*. Extracting complex biological events with rich graph-based feature sets. In: *Proceedings of the Workshop on BioNLP: Shared Task, Boulder, CO, June 2009*. Boulder, CO, USA: Association for Computational Linguistics; 10–18.

127. Hoifung Poon Lucy Vanderwende. Joint inference for knowledge extraction from biomedical literature. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, Los Angeles, CA, June 2010*. Los Angeles, CA, USA: Association for Computational Linguistics; 813–821.

128. Riedel S, McClosky D, Surdeanu M, *et al*. Model combination for event extraction in BioNLP 2011. In: *Proceedings of BioNLP Shared Task 2011 Workshop, Portland, OR, June 2011*. Portland, OR, USA: Association for Computational Linguistics; 51–55.

129. Vlachos A, Craven M. Search-based structured prediction applied to biomedical event extraction. In: *Proceedings of the Fifteenth Conference on Computational Natural Language Learning,* CoNLL '11, *Stroudsburg, PA, 2011*. Stroudsburg, PA, USA: Association for Computational Linguistics; 49–57.

130. McClosky D, Surdeanu M, Manning CD. Event extraction as dependency parsing for BioNLP 2011. In: *Proceedings of BioNLP Shared Task 2011 Workshop, Portland, OR, June 2011*. Portland, OR, USA: Association for Computational Linguistics; 41–45.

131. McDonald R, Crammer K. Online large-margin training of dependency parsers. In: *Proceedings of the 43rd Annual Meeting of the ACL, Ann Arbor, June 2005*. Ann Arbor: Association for Computational Linguistics; 91–98.

132. McDonald R, Pereira F, Ribarov K, *et al*. Non-projective dependency parsing using spanning tree algorithms. In: *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), Vancouver, October 2005*. Vancouver: Association for Computational Linguistics; 523–530.

133. Miwa M, Thompson P, Ananiadou S. Boosting automatic event extraction from the literature using domain adaptation and conference resolution. *Bioinformatics* 2012;**28**: 1759–65.

134. Ravikumar K, Liu H, Cohn JD, *et al*. Literature mining of protein-residue associations with graph rules learned through distant supervision. J Biomed Semantics, 2012; **3(Suppl 3)**:S2.

135. Hobbs JR. Coherence and coreference. *Cognit Sci* 1979;**3**: 67–90.

136. Lee H, Recasens M, Chang A, *et al*. Joint entity and event coreference resolution across documents. In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Jeju Island, Korea, July 2012*;489–500, Association for Computational Linguistics.

137. Kim Y, Riloff E, Gilbert N. The taming of reconcile as a biomedical coreference resolver. Proceedings of the BioNLP Shared Task 2011 Workshop*, BioNLP Shared Task '11*, Stroudsburg, PA, 2011. Stroudsburg, PA, USA: Association for Computational Linguistics; 89–93.

138. Stoyanov V, Cardie C, Gilbert N, *et al*. Reconcile: a coreference resolution research platform. *Computing and Information Science Technical Reports* 2010. Cornell Univeristy, USA.

139. Raghunathan K, Lee H, Rangarajan S, *et al*. A multi-pass sieve for coreference resolution. In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, MIT, MA, October 2010*. MIT, MA, USA: Association for Computational Linguistics; 492–501.

140. Lee H, Peirsman Y, Chang A, *et al*. Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task. In: *Proceedings of the 15th Conference on Computational Natural Language Learning: Shared Task, Portland, OR, June 2011*. Portland, OR, USA: Association for Computational Linguistics; 28–34.

141. Yoshikawa K, Riedel S, Hirao T, *et al*. Coreference based event-argument relation extraction on biomedical text. *J Biomed Semantics* 2011;**2(Suppl 5)**:S6.

142. Song Y, Jiang J, Zhao WX, *et al*. Joint learning for conference resolution with Markov logic. In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Jeju Island, Korea, July 2012*;1245–54. Association for Computational Linguistics.

143. Pradhan S, Ramshaw L, Marcus M, *et al*. CoNLL-2011 shared task: modeling unrestricted coreference in OntoNotes. In: *Proceedings of the 15th Conference on Computational Natural Language Learning: Shared Task, Portland, OR, June 2011*. Portland, OR, USA: Association for Computational Linguistics; 1–27.

144. Kerrien S, Aranda B, Breuza L, *et al*. The IntAct molecular interaction database in 2012. *Nucleic Acids Res* 2012; **40**(Database issue):D841–6.

145. Licata L, Briganti L, Peluso D, *et al*. MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res* 2012; **40**(Database issue):D857–61.

146. Pyysalo S, Airola A, Heimonen J, *et al*. Comparative analysis of five protein-protein interaction corpora. *BMC Bioinformatics* 2008;**9(Suppl 3)**:S6.

147. Szklarczyk D, Franceschini A, Kuhn M, *et al*. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res* 2011;**39**(Database issue):D561–8.

148. Kuhn M, Szklarczyk D, Franceschini A, *et al*. STITCH 3: zooming in on protein-chemical interactions. *Nucleic Acids Res* 2012;**40**(Database issue):D876–80.

149. Salwinski L, Miller CS, Smith AJ, *et al*. The database of interacting proteins: 2004 update. *Nucleic Acids Res* 2004; **32**(Database issue):D449–51.

150. Stark C, Breitkreutz BJ, Reguly T, *et al*. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res* 2006;**34**(Database issue):D535–9.

151. Liu T, Lin Y, Wen X, *et al*. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res* 2007;**35**(Database issue):D198–201.

152. Barrell D, Dimmer E, Huntley RP, *et al*. The GOA database in 2009–an integrated gene ontology annotation resource. *Nucleic Acids Res* 2009;**37**(Database issue): D396–403.

153. Turner B, Razick S, Turinsky AL, *et al*. iRefWeb: interactive analysis of consolidated protein interaction data and their supporting evidence. *Database* 2010;**2010**:baq023.

154. Orchard S, Kerrien S, Abbani S, *et al*. Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. *Nat Methods* 2012;**9**:345–50.

155. Mcclosky D. Any domain parsing: automatic domain adaptation for natural language parsing. PhD thesis, Providence, RI, 2010. AAI3430199.

156. Cohen KB, Johnson HL, Verspoor K, *et al*. The structural and content aspects of abstracts versus bodies of full text journal articles are different. *BMC Bioinformatics* 2010;**11**: 492.

157. Fundel K, Küffner R, Zimmer R. RelEx–relation extraction using dependency parse trees. *Bioinformatics* 2007;**23**: 365–371.

158. Chiang JH, Liu HH, Huang YT. Condensing biomedical journal texts through paragraph ranking. *Bioinformatics* 2011;**27**:1143–49.

159. Yang JB, Mao Q, Xiang QL, *et al*. Domain adaptation for coreference resolution: an adaptive ensemble approach. In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Jeju Island, Korea, July 2012*. Jeju Island, Korea: Association for Computational Linguistics; 744–753.

160. Smalheiser NR, Torvik VI, Zhou W. Arrowsmith two-node search interface: a tutorial on finding meaningful links between two disparate sets of articles in MEDLINE. *Comput Methods Programs Biomed* 2009;**94**:190–97.

161. Tiffin N, Okpechi I, Perez-Iratxeta C, *et al*. Prioritization of candidate disease genes for metabolic syndrome by computational analysis of its defining phenotypes. *Physiol Genomics* 2008;**35**:55–64.

162. Chang W, Ma L, Lin L, *et al*. Identification of novel hub genes associated with liver metastasis of gastric cancer. *Int J Cancer* 2009;**125**:2844–53.

163. Rebholz-Schuhmann D, Kirsch H, Arregui M, *et al*. EBIMed–text crunching to gather facts for proteins from Medline. *Bioinformatics* 2007;**23**:e237–44.

164. Gorr SU, Wennblom TJ, Horvath S, *et al*. Text-mining applied to autoimmune disease research: the Sjögren's syndrome knowledge base. *BMC Musculoskelet Disord* 2012;**13**:119.

165. Pyysalo S, Ohta T, Miwa M, *et al*. Towards exhaustive protein modification event extraction. In: *Proceedings of the 2011 Workshop on Biomedical Natural Language Processing, ACL-HLT 2011, Portland, OR, June 2011*, 114–123. Association for Computational Linguistics, Portland, OR, USA.

166. Liakata M, Thompson P, de Waard A, *et al*. A three-way perspective on scientific discourse annotation for knowledge extraction. In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, Jeju, Republic of Korea, July 2012*;37–46. Association for Computational Linguistics.

167. Tsuruoka Y, Miwa M, Hamamoto K, *et al*. Discovering and visualizing indirect associations between biomedical concepts. *Bioinformatics* 2011;**27**:i111–19.

168. Fernández JM, Hoffmann R, Valencia A. iHOP web services. *Nucleic Acids Res* 2007;**35**(Web Server issue):W21–6.

169. Epple A, Scherf M. BiblioSphere—Hypothesis generation in regulatory network analysis. *Bioinformatics for Systems Biology*. Totowa, NJ: Humana Press, 2009;401–12.

170. Demaine J, Martin J, Wei L, *et al*. LitMiner: integration of library services within a bio-informatics application. *Biomed Digit Libr* 2006;**3**:11.

171. CellDesigner. http://celldesigner.org (13 February 2013, date last accessed).

172. Matsuoka Y, Ghosh S, Kikuchi N, *et al*. Payao: a community platform for SBML pathway model curation. *Bioinformatics* 2010;**26**:1381–3.

173. Rzhetsky A, Iossifov I, Koike T, *et al*. GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data. *J Biomed Inform* 2004;**37**:43–53.

174. Koh GCKW, Porras P, Aranda B, *et al*. Analyzing protein-protein interaction networks. *J Proteome Res* 2012; **11**:2014–31.

175. Zhang Y, Zhong L, Xu B, *et al*. SpermatogenesisOnline 1.0: a resource for spermatogenesis based on manual literature curation and genome-wide data mining. *Nucleic Acids Res* 2013;**41**:D1055–62.

176. Thieu T, Joshi S, Warren S, *et al*. Literature mining of host-pathogen interactions: comparing feature-based supervised learning and language-based approaches. *Bioinformatics* 2012;**28**:867–75.

177. Sentences for text-mining from IntAct. ftp://ftp.ebi.ac.uk/pub/databases/intact/current/various/data-mining (13 February 2013, date last accessed).

178. Kikugawa S, Nishikata K, Murakami K, *et al*. PCDq: human protein complex database with quality index which summarizes different levels of evidences of protein complexes predicted from H-Invitational protein-protein interactions integrative dataset. *BMC Syst Biol* 2012; **6(Suppl 2)**:S7.