DATABASE
The Journal of Biological Databases and Curation

Database, Vol. 2012, Article ID bas017, doi:10.1093/database/bas017

Original article

How to link ontologies and protein–protein interactions to literature: text-mining approaches and the BioCreative experience

Martin Krallinger[1], Florian Leitner[1], Miguel Vazquez[1], David Salgado[2], Christophe Marcelle[2], Mike Tyers[3,4], Alfonso Valencia[1] and Andrew Chatr-aryamontri[3,4,*]

[1]Structural and Computational Biology Group, Spanish National Cancer Research Centre (CNIO), Spain, [2]Australian Regenerative Medicine Institute, Monash University, Australia, [3]School of Biological Sciences, University of Edinburgh, Edinburgh, UK and [4]Institute for Research in Immunology and Cancer, Université de Montréal, Montréal, QC, Canada H3C 3J7

***Corresponding author:** Tel: +1 514 343 111 ext. 44668; Fax: +1 514 343 5839; Email: andrew.chatr-aryamontri@umontreal.ca

Submitted 15 October 2011; Revised 14 February 2012; Accepted 28 February 2012

There is an increasing interest in developing ontologies and controlled vocabularies to improve the efficiency and consistency of manual literature curation, to enable more formal biocuration workflow results and ultimately to improve analysis of biological data. Two ontologies that have been successfully used for this purpose are the Gene Ontology (GO) for annotating aspects of gene products and the Molecular Interaction ontology (PSI-MI) used by databases that archive protein–protein interactions. The examination of protein interactions has proven to be extremely promising for the understanding of cellular processes. Manual mapping of information from the biomedical literature to bio-ontology terms is one of the most challenging components in the curation pipeline. It requires that expert curators interpret the natural language descriptions contained in articles and infer their semantic equivalents in the ontology (controlled vocabulary). Since manual curation is a time-consuming process, there is strong motivation to implement text-mining techniques to automatically extract annotations from free text. A range of text mining strategies has been devised to assist in the automated extraction of biological data. These strategies either recognize technical terms used recurrently in the literature and propose them as candidates for inclusion in ontologies, or retrieve passages that serve as evidential support for annotating an ontology term, e.g. from the PSI-MI or GO controlled vocabularies. Here, we provide a general overview of current text-mining methods to automatically extract annotations of GO and PSI-MI ontology terms in the context of the BioCreative (Critical Assessment of Information Extraction Systems in Biology) challenge. Special emphasis is given to protein–protein interaction data and PSI-MI terms referring to interaction detection methods.

Introduction

Advances in laboratory technologies and data analysis methodologies are permitting the exploitation of complex experimental data sets in ways that were unthinkable just a few years ago (1–3). However, although the number of scientific articles containing relevant data is steadily increasing, the majority of published data is still not easily accessible for automated text processing systems. In fact, the information is still buried within the articles rather than being summarized in computer readable formats (4). Therefore, it is necessary to perform the additional step of annotating the experimental data in formats suitable for systematic consultation or computation. This task is performed manually by curators of databases specialized in diverse biological domains, ranging from cellular phenotypes and tissue anatomy to gene function. The importance and the critical role played by such themed biocuration efforts are evident by the multitude of databases reported over the years in the NAR Database special issue (5) and by the birth of dedicated journals such as Database.

Database, Vol. 2012, Article ID bas017, doi:10.1093/database/bas017

Different models have been followed to generate annotations from the literature (6,7). In the museum model, a relatively small group of specialized curators perform a particular literature curation effort, while in the jamboree model a group of experts meet for a short intensive annotation workshop. When various research groups scattered at different locations share common research interests and they jointly organize into a collaborative decentralized annotation effort (working from their own laboratories), the so-called cottage industry model is followed. Devoted expert curators produce quality annotations, but because manual curation is time-consuming and there is a limited number of curators, it is difficult to keep current with the literature. Potential alternatives inspired by successful efforts, such as Wikipedia, are the open community model (8) and the author-based annotations model (9,10). The first does not have major restrictions on the actual annotators, as the whole community can contribute to generate annotations. In some cases, qualified roles for the contributors have been proposed to guarantee a certain level of confidence in the annotations. The idea behind author-based annotations is that the authors themselves provide minimal annotations of their own article during the writing or submission process, going beyond author-provided keywords for indexing purposes.

Each of the manual literature curation models previously introduced here still faces the problem of the increasing volume of literature (11). Therefore, some attempts have been made to generate annotations automatically using automated text mining. Databases constructed according to the automated text-mining model are limited by performance issues but can generate valuable results in case of lack of manual annotations (12,13). A hybrid approach, namely text-mining-assisted manual curation, wherein semi-automated literature mining tools are integrated into the biocuration workflow, represents a more promising solution (14,15).

Controlled vocabularies have been fundamental for all of these diverse annotation types, from the purely manual ones to totally automatic annotations. Key tools in the annotation of experimental data are bio-ontologies, a well-defined set of logic relations and controlled vocabularies that permit an accurate description of the experimental findings (16).

The BioCreative initiative (Critical Assessment of Information Extraction systems in Biology) (17,18) is a community-wide effort for the evaluation of text mining and information extraction systems applied to the biological domain. Its major purpose is to stimulate the development of software that can assist the biological databases in coping with the deluge of data generated by the 'omics' era. We provide here a general overview of the BioCreative experience with biomedical ontologies. For the BioCreative initiatives, it was of particular importance that annotations chosen as part of a challenge task had been generated through a model followed by research groups employing expert curators using well-established biocuration workflows refined over years of manual literature curation.

In particular, we will focus on the attempts that have been made to automatically extract protein–protein interaction (PPI) data taking advantage of ontologies, and to associate ontology terms to the interactions.

Protein interaction biocuration

The opportunity to decipher the mechanisms underlying cellular physiology from the analysis of molecular interaction networks has prompted the establishment of databases devoted to the collection of such data, with great attention to protein and genetic interactions (19–22). Some of the major protein interaction databases (19–25) are now federated in the International Molecular Exchange (IMEx) consortium, whose primary goals are to minimize curation redundancy and to share the data in a common format. All active IMEx members share the same data representation standard, the Human Proteome Organisation Proteomics Standards Initiative Molecular Interactions (HUPO PSI-MI) (26). The PSI-MI provides the logic model and the controlled vocabulary for representation of molecular interactions. Not surprisingly, the members of the IMEx consortium themselves are the main contributors to the development and maintenance of the PSI-MI ontology.

The PSI-MI was introduced with the intent to facilitate data integration among databases specifically for the representation of binary or *n*-nary interactions. It also allows in-depth annotation of the experimental set-up such as the experimental or biological role of the interactors, the experimental method employed for the detection of the interaction, the binding domain of the interactors, and the kinetics of the binding reaction, among other attributes (the PSI-MI ontology can be explored at the EBI ontology look-up service) (27). The PSI-MI is not restricted to the representation of physical interactions but permits the thorough annotation of genetic interactions and even experimental evidence of co-localization among molecules. Each attribute of the interaction is described by a rich controlled vocabulary which is organized in a well-defined hierarchy and continuously updated and maintained by the PSI-MI workgroup. Regrettably, despite the cooperative efforts of the IMEx databases, the complete annotation of interaction data from the biomedical literature, and in particular, the subset of interactions involving human genes and their products, remains far from complete. The time-consuming nature of manual curation severely hampers the achievement of an exhaustive collection of molecular interactions. The thorough annotation of the experimental data contained in a single scientific article can take anywhere from minutes to hours. Hence, any automated support

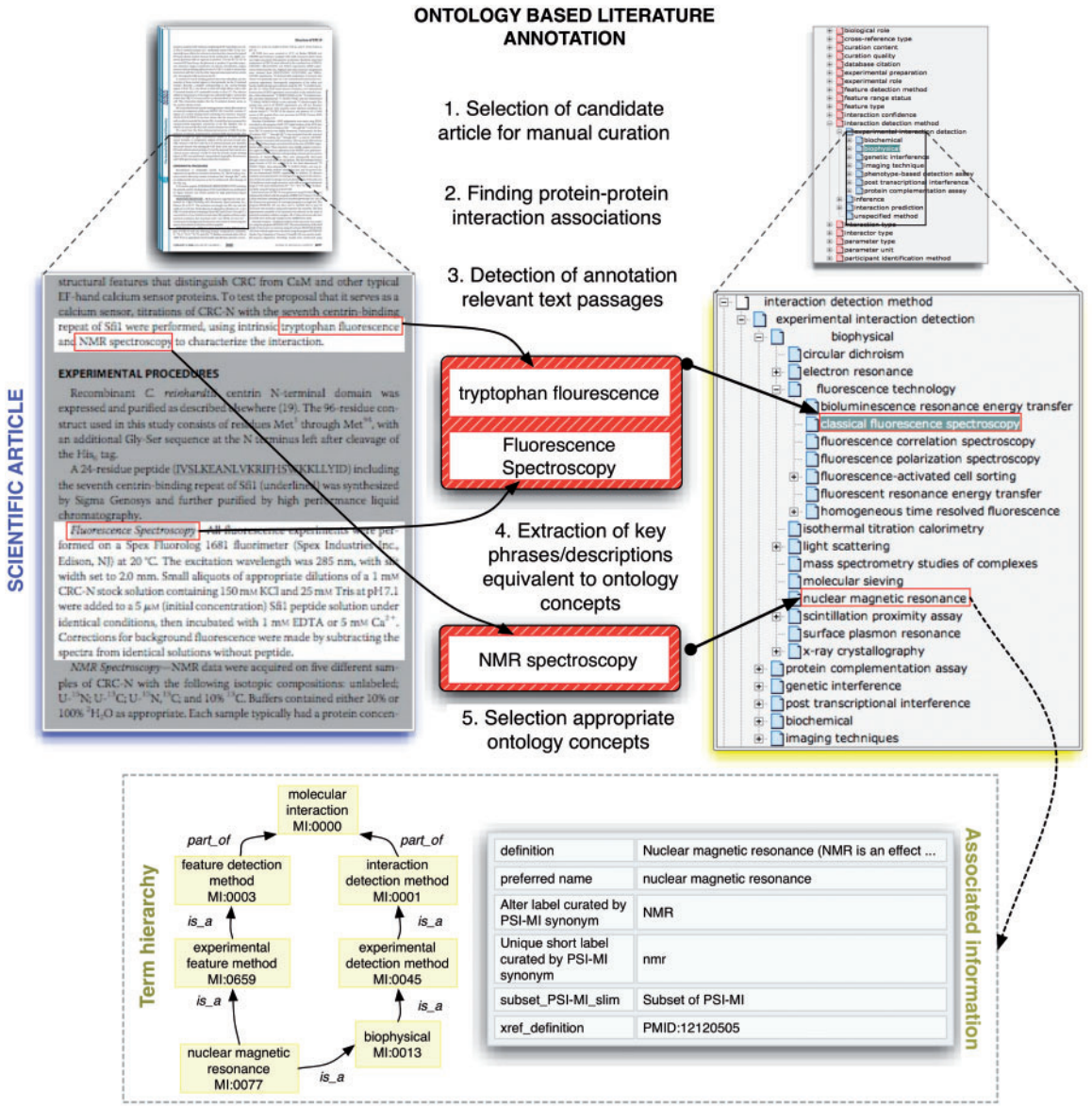

Figure 1. This figure shows schematically how protein interaction data is annotated and/or marked up using ontologies. Systems such as MyMiner (myminer.armi.monash.edu.au/links.php), have been used for text labeling and highlighting purposes in the context of the BioCreative competition. The main steps illustrated in this figure have been addressed in the BioCreative challenges. Finding associations between textual expressions referring to experimental techniques used to characterize protein interactions and their equivalent concepts in the MI ontology is cumbersome in some cases when deep domain inference is required. Experienced curators are able to quickly navigate the term hierarchy to find the appropriate terms while novice annotators often need to search the ontology using method keywords as queries and consult associated descriptive information for potential candidate terms.

that assists the database curators—be it the selection of the relevant literature or identification and annotation of the interactions—is more than welcome by the database community. Figure 1 provides a schematic representation of the manual literature curation of PSI-MI concepts for protein interaction annotation.

A number of initiatives have been started in order to facilitate the automated extraction of information from the biomedical literature and of PPI data in particular. The Structured Digital Abstracts developed by *FEBS Letters* in collaboration with the MINT database (20), for instance, is a structured text appended to the classical abstract that can be easily parsed by text-mining tools. Each biological entity (proteins) and relationship between these entities is tagged with appropriate database identifiers, thus permitting an unambiguous interpretation of the data.

Database, Vol. 2012, Article ID bas017, doi:10.1093/database/bas017

Natural language processing and ontologies

In recent years, we have witnessed a flourishing of ontologies that attempt to accurately represent the complexity of the biological sciences (28). Hence, we now have ontologies describing a wide variety of biological concepts, spanning from clinical symptoms to molecular interactions. They not only attempt to capture in a more formal way the meaning (semantics) of a particular domain based on community consensus (29) but are also a key element for database interoperability and querying, as well as knowledge management and data integration (30).

Some of these ontologies can now be integrated with other ontologies, broadening their descriptive potential (31). Furthermore, the Gene Ontology (GO) (32) has grown considerably over 10 years, counting now almost 35 000 terms, compared to the initial 5000. [for a general introduction to the GO annotation process refer to Hill *et al*. (33)].

The increasing number of biological terms and concepts covered by these ontologies has prompted a growing interest in their potential for use in the development of methods for automatic data extraction from the biomedical literature.

However, while biomedical ontologies are indispensable in the daily practice of database curators, it remains to be established if text mining can really benefit from well-established ontologies. In fact, while an analysis of the lexical properties of the GO indicates that a large percentage of GO terms are potentially useful for text mining tools (34), other evidence suggests that many of the Open Bbiomedical Ontologies (28) are not suitable for effective natural language processing applications (35).

This discrepancy is due to the fact that often the information is not only present as natural language data, but often also requires interpretation of information contained in images or obtained by interpreting the data reported in the articles. As a consequence, not every piece of information is unambiguously linked to a continuous passage of text hence detectable by parsing machines.

The results of the first BioCreative challenge suggest that a combination of several factors can influence the performance of text mining systems in the extraction of GO terms associated with defined genes, including the specificity of the terms and their GO branch membership (36).

Ontologies benefitting from an iterative process of expansion and restructuring based on direct observations (analysis of scientific literature) made by communities of active users more likely will successfully result in a resource for text-mining purpose. Inclusion of such observations in the ontologies will dramatically increase their potential in the context of text mining.

Nevertheless, some popular text-mining-based applications, such as Textpresso (37), NCBO Annotator (38), Geneways (39), Domeo (40) or PubOnto (41), rely on the usage of ontologies. These kinds of systems are currently exploring ontologies mainly as lexical resources of controlled vocabulary terms for text indexing or markup purposes. They assist the end users in improving the detection of annotation-relevant information at a very general level. Efficiently handling complex terms and annotation types is thus still a challenge for such approaches, making the results of the BioCreative tasks particularly interesting to better understand the comparison between manual and automated extractions. Adapting some of the methodologies that participated in BioCreative into such technical frameworks could potentially capture previously missing annotation types or concepts.

BioCreative

The BioCreative challenge was established in 2004 with the purpose of assessing the state-of-the-art of text-mining technologies applied to biological problems. Although it is called a challenge, the primary aim of BioCreative is not to identify a contest winner. Instead the ambition of BioCreative is manifold: (i) to benchmark the performance of text mining applications, (ii) to promote communication between bioinformaticians, text miners, and database curators, (iii) to define shared training and 'gold standard' test data and (iv) to spur the development of high-performance suites. To date, four editions of BioCreative have been organized, each consisting of two or more specific tasks (Table 1). Each task was designed to test the ability of the systems to detect biological entities (gene or proteins) and/or to link them to stable database identifiers, and evaluate how efficiently facts or functional relations can be associated with the biological entities (e.g. protein function and PPI). Figure 2 shows how these BioCreative challenges have evolved over time in the context of related community efforts, resources and applications.

The first edition of the BioCreative challenge (17) was geared to the needs of model organism database curators. It consisted of two main tasks. The first task was further divided into two subtasks: the recognition of gene mentions in the text (42) and the linking of identified proteins from yeast, fly and mouse in abstracts to model organism database identifiers (43). The second task challenged the participants to annotate human gene products, defined by their UniProtKB/Swiss-Prot accession codes (44), with the corresponding GO codes by mining full-text articles (36). In particular, teams were asked to return the textual evidence for the GO term assigned to a defined set of proteins. Figure 3 illustrates schematically the idea behind the associated annotation process where for proteins described in a given paper, GO annotation evidence had to be extracted.

Table 1. Summary of the BioCreative editions related to the identification of ontology terms in articles

Information	BioCreative I, task 1	BioCreative I, task 2	BioCreative II—IMS	BioCreative III—IMS
Description	Return evidence text fragments for protein–GO–document triplets	Predict GO annotations derivable from a given protein–article pair	Prediction of MI annotations from PPI-relevant articles	Prediction of MI annotations from PPI-relevant articles (ranked with evidence passages)
Ontologies	GO	GO	MI ontology	MI ontology
Curators/databases	GOA-EBI	GOA-EBI	MINT and IntAct	BioGRID and MINT
Participants	9	6	2	8
Data/format	Full-text articles, SGML format	Full-text articles, SGML format	Full-text articles, PDF and HTML format	Full-text articles, PDF format
Training	803 articles	803 articles	740 articles	2003 training articles and 587 development set articles
Test	113 articles	99 articles	358 articles	223 articles
Evaluation	Three labels (correct, general, wrong), % correct cases	Three labels (correct, general, wrong), % correct cases	Precision, recall and F-score; mapping to the parent terms	Precision, recall, F-score, ranked predictions (AUC iP/R)
Methods	Term lookup, pattern matching/template extraction, term tokens (information content of GO words, *n*-gram models), part-of-speech of GO words and machine learning	Term lookup, pattern matching/template extraction, term tokens (information content of GO words, *n*-gram models), part-of-speech of GO words and machine learning	Pattern matching, automatically generating variants of MI terms, handcrafted patterns	Cross-ontology mapping, manual and automatic extension of method names, statistic of work tokens building terms (mutual information, chi square), machine learning of training set articles
Result highlights	Precisions from 46% to 80%, accuracy of ~30%	Precisions from 9% to 35%	Precision from 32% to 67%, best *F*-score of 48	Most between 30% and 80%, best *F*-score of 55
Observation	Limited recall, effect of GO term length	Limited recall, difference in performance depending on GO categories, cellular component terms are easier	Difficulties with very general method terms	Difficulties in case of methods not specific to PPIs, problems with recall

Precision and recall were the basic metrics employed to evaluate the performance of the systems during this BioCreative challenge. Precision is the fraction of true positive (TP) cases, i.e. correct results, divided by the sum of TP and false positive (FP) cases. Recall can be considered as the fraction of TP results divided by the sum of TP and false negative (FN) results, i.e. relevant cases missed by the system. To account for both of these measures, the *F*-measure, i.e. harmonic mean of precision and recall was used. For the GO task, database curators had to manually evaluate the automatically extracted evidence passages to determine if they correctly supported the annotations, as exemplified in Figure 4 (36).

The first BioCreative competition saw the participation of 27 teams and some of the text mining algorithms yielded encouraging results in the identification of the gene names and in linking them to database identifiers (80% precision/recall) (43).

The identification of gene mentions in sentences was addressed using machine-learning and natural language processing techniques and benefited from training and test data in the form of labeled text prepared by biologists.

For linking (normalizing) genes mentioned in abstracts, there was a considerable variability in performance depending on the used model organism. In the case of yeast, an *F*-score of 0.92 could be reached, while in the case of fly (*F*-score of 0.82) and mouse (*F*-score of 0.79) the performance was considerable lower due to less consistent naming nomenclature use and high degree of ambiguity of gene names.

Conversely, the results of the functional annotation task proved that the interpretation of complex biological data, and thus linking text to the GO ontology, is extremely challenging for text mining tools. The obtained results indicated that some categories of GO, in particular, the terms expressing sub-cellular location provided by the cellular

Database, Vol. 2012, Article ID bas017, doi:10.1093/database/bas017

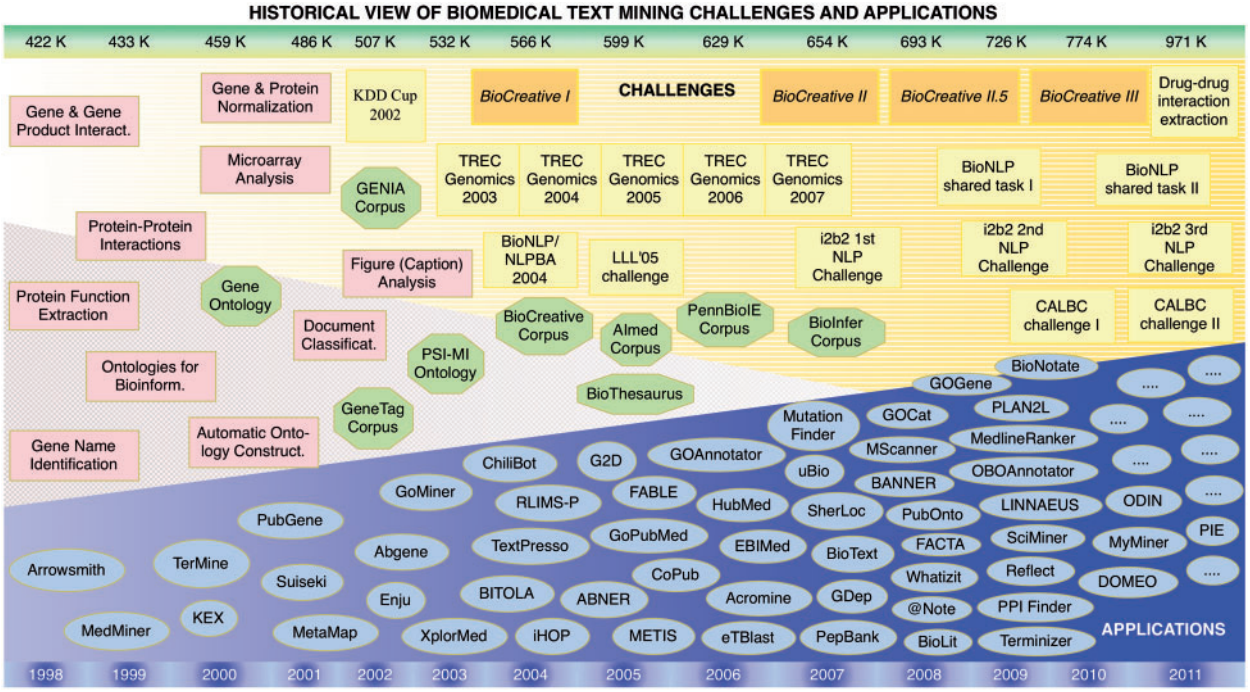

Figure 2. Historical view and timeline of the BioCreative challenges in the context of other community efforts, textual resources (corpora) and applications developed in the area of biomedical text mining. The upper bar shows the number of new records added to PubMed each year, expressed in thousands (K). The lower bar refers to the corresponding year timeline. Pink squares, appearance of biomedical text mining methods; green octagons, relevant ontologies, lexical resources and corpora; yellow boxes, community challenges; blue ovals, biomedical text mining applications.

component (CC) branch seemed to be more amenable for text-mining strategies.

Outcomes of the BioCreative challenge for PPIs

The task of extracting PPI data was introduced in the second edition of BioCreative (45). Several subtasks were defined: detecting the literature containing protein interaction data (Interaction Article Subtask, IAS), identifying the interaction pairs and linking the interacting partners to UniProtKB/Swiss-Prot identifiers (Interaction Pair Subtask, IPS), identifying the experimental methods employed to detect the interaction (Interaction Method Subtask, IMS) and retrieving the textual evidence of the interaction (Interaction Sentences Subtask, ISS). The PPI task was a collaborative effort with IntAct and MINT, databases whose curators annotated the training and test sets used in the various tasks (46).

The experimental methods are important to infer how likely it is that a given protein interaction actually occurs *in vivo*, and it is usually the cumulative evidence rather than a single experiment that defines the reliability of the interaction. At a practical level, for curators, it is fundamental to identify in the article if there are experimental techniques usually associated with the detection of protein interactions (e.g. two hybrid, affinity purification technologies). These facts motivated the introduction of the IMS (45).

For the IMS subtask, the two participating teams were asked to identify from the text the list of the experimental techniques employed for the detection of PPIs, and their results were compared with a reference list generated by manual annotation. The experimental interaction detection techniques allowed for this task consisted of a sub-graph specified in the PSI-MI ontology. The highest score for exact match precision was 48%, but if matching to parent terms in the ontology was allowed, the score raised to an encouraging 65% (45). This improved performance was obtained by considering as correct those predicted terms that, when compared to the manually annotated terms, were either an exact match or a direct parent concept based on the PSI-MI ontology graph structure.

This result is due to the fact that some ontology terms are far too specific to match the vocabulary routinely used in the biomedical literature. For instance, while' coimmunoprecipitation' (MI:0019) is widely used in the scientific literature, its child terms 'anti bait coimmunoprecipitation' (MI:0006) and 'anti tag coimmunoprecipitation' (MI:0007) are not. The two child terms are used for annotation by database curators to further indicate if the experiment has been conducted with an

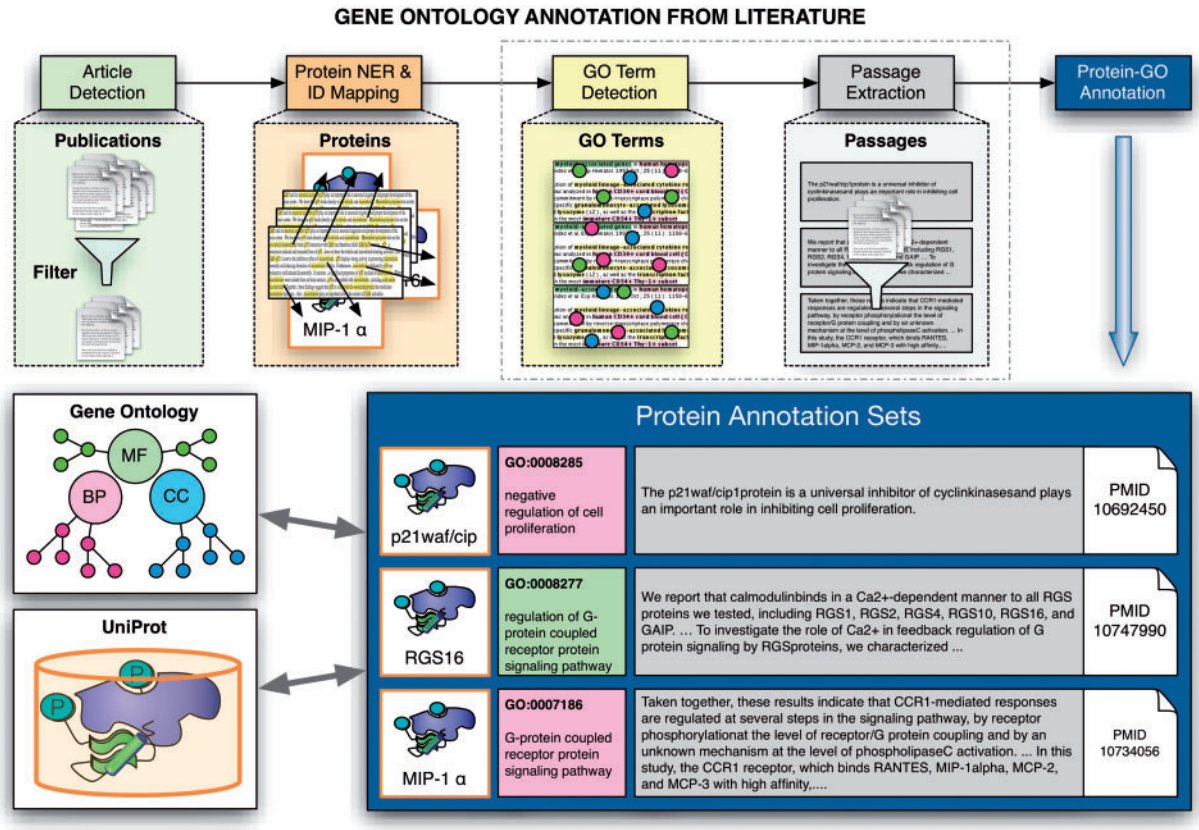

Figure 3. Schematic overview of the extraction of GO annotations from the literature. The process illustrates the individual steps of the annotation process, covering the initial selection of relevant documents for GO annotation of proteins, identification of proteins and their corresponding database identifiers followed by the extraction of associations to GO terms and the retrieval of evidence sentences/passages. The participating teams had to provide the evidence passages for a given document–protein–GO term triplet for one subtask, and to actually detect GO–protein associations (together with evidence passages) for the other subtask.

antibody recognizing the protein or a tag fused to the target protein, respectively. The use of these terms is therefore largely limited to human curator interpretation of the literature rather than explicit text mentions of these terms.

Attempts that might be promising particularly for terms that are lengthy and representative of complex concepts could also consider the use of term definitions. With this respect, GO term definitions had been exploited by Piao *et al*. (47) for identifying and analyzing relations between terms. The definitions of PSI-MI terms have also been used for linking PSI-MI terms to full-text articles by analyzing unigrams and character *n*-grams from the PSI-MI definition and synonyms (48).

Several studies have been published in the biomedical domain with the purpose to quantify through metrics how closely related two terms are in their meanings, i.e. their semantic similarity (49). This is an important issue not only for comparing text-mining results to manual annotations, but also for measuring consistency of manual annotations themselves in inter-annotator agreement studies or to determine the functional similarity between genes annotated with those terms. A simple approach for measuring semantic similarity can be the calculation of the distance between two terms in the graph path underlying the ontology. Semantic similarity calculations have been promising for resources like WordNet (50,51), which is essentially a lexical database of English words together with their semantic relation types with practical usage for text analysis. This resource differs therefore in scope from GO or the PSI-MI ontology, whose primary use is for annotation of gene products. Semantic similarity calculations have shown useful results to quantity functional similarity between gene products based on their GO annotations (49), but using them for directly quantifying the similarity between predicted and manually annotated terms in the context of BioCreative remained problematic.

The IMS task was replicated in the BioCreative III edition (52–54) and saw increased participation, with eight teams. The difference from the previous edition was that participants were asked to provide a list of interaction detection

Database, Vol. 2012, Article ID bas017, doi:10.1093/database/bas017

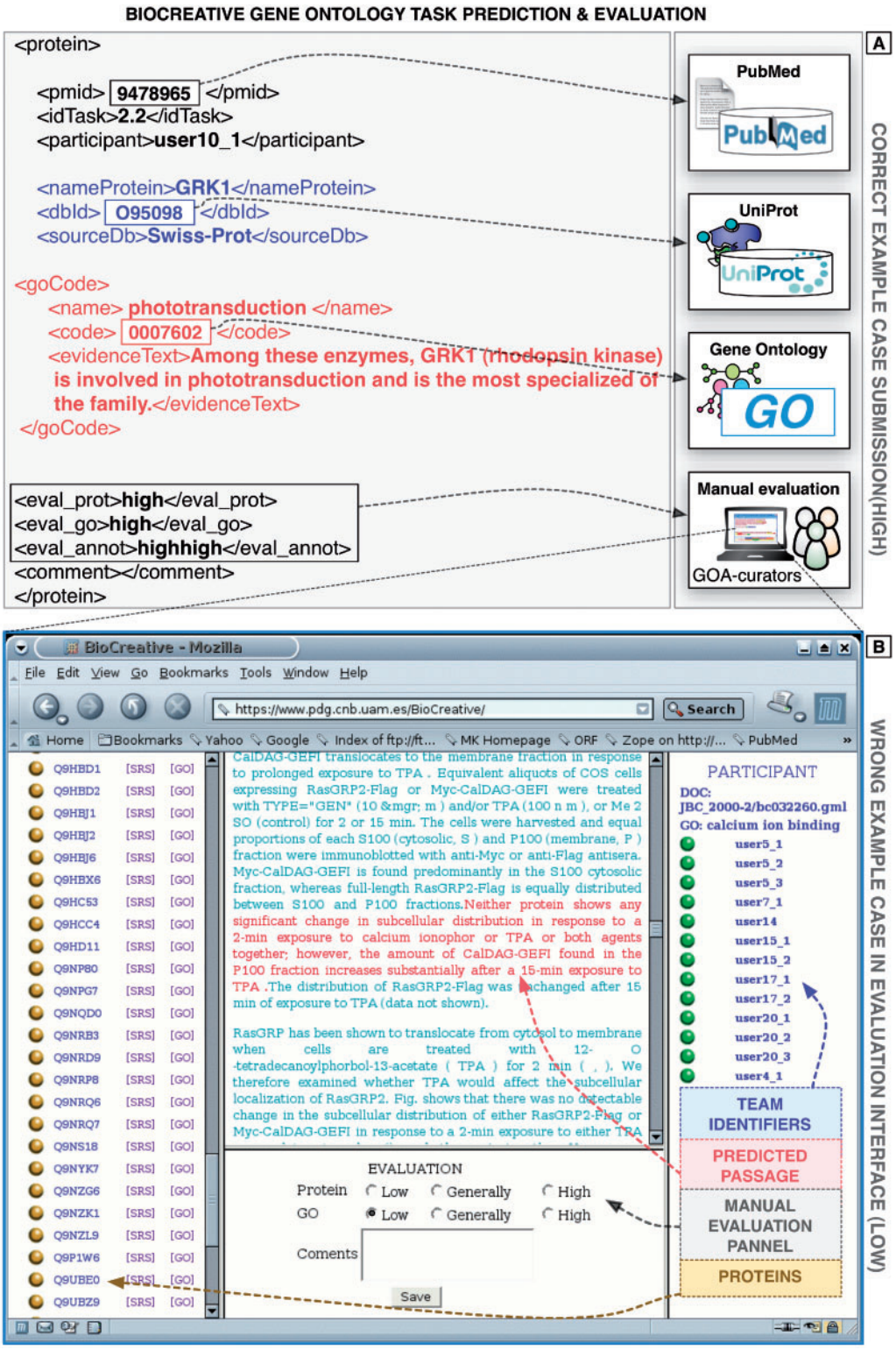

Figure 4. Example predictions of the GO task of BioCreative I. (**A**) Here a correct prediction is shown, containing the information on the corresponding document, protein and GO term as well as the supporting evidence text passages extracted automatically from the full-text article. (**B**) Example prediction (wrong) showing a screen shot of the original evaluation interface developed at the time for this task (based on Apache/PHP). The original evaluation application is not functional anymore and was implemented specifically for this task. Proteins and GO terms were defined unambiguously through corresponding standard identifiers. The database curators manually evaluated both the correctness of the protein as well as the GO terms.

INTERACTION METHOD TASK PREDICTION EXAMPLES: BIOCREATIVE III

ARTICLE ID (PubMed)	MI CONCEPT ID (Ψ PSI-MI)	RANK	CONFIDENCE SCORE	EVIDENCE PASSAGE	PSI-MI TERM & SYNONYMS (Ψ)	
19056683	MI:0018	2	0.998559	TP53INP2 Interacts with GABARAP and GABARAP-like2 Proteins Proteins interacting with TP53INP2 were identi\xef\xac\x81ed by yeast two-hybrid screening of a HeLa cDNA library.\n	two hybrid', '2 hybrid', '2-hybrid', '2H', '2h', 'classical two hybrid', 'Gal4 transcription regeneration', 'two-hybrid', 'yeast two hybrid	EASY
19218236	MI:0096	1	0.668406	due to the bridging effect of SirT1 (see below) . In an in vitro binding assay , GST - DBC1 efficiently pulled down in vitro - translated SUV39H1 (Fig. 2b) , suggesting that the binding is a direct interaction . These	pull down	NORMAL
18625238	MI:0114	3	0.557822	helix in the EF loop (Leu63 - Ala65) . CHIR - AB1 forms homodimers Although the CHIR - AB1 protein used for crystallization was purified from the monomeric peak , crystal packing created a symmetric CHIR - AB1 dimer in which residues \n	x-ray crystallography', 'X-ray', 'x-ray diffraction	DIFFICULT
19481529	MI:0424	1	0.630389	phosphorylated Ser437Ala mutant , suggesting phosphorylation of PACS-2 Ser437 was required for binding 14-3-3 proteins . We then conducted a fluorescence polarization assay to determine quantitatively whether phosphorylated	protein kinase assay	DIFFICULT
18922473	MI:0006	2	0.472072	Interaction between the endogenous TRAF6 and TAK1 in AML12 cells as determined by immunoprecipitation with anti - TAK1 antibody , followed by anti - TRAF6 Western blot . The TGF - \xce\xb2 treatment was for 30 minutes and the total rabbit IgG \n	anti bait coimmunoprecipitation', 'anti bait coip	DIFFICULT

Figure 5. Representative predictions submitted for the MI task of BioCreative III of diverse degrees of difficulty for automated systems. The examples correspond to submissions from various teams. Participating teams had to return the article identifier, the concept identifier for the interaction detection method according to the MI ontology, a rank, a confidence score as well as a supporting text evidence passages extracted from the full-text article. Submissions were plain text files where each field was separated using a tabulator. This figure provides colored highlights of original predictions to better grasp the output. In red, the original term from the MI ontology and its synonyms have been added to facilitate the interpretation of the results. As can be seen some cases are rather straightforward, and could be detected by direct term lookup, while others require generating lexical variants or even more sophisticated machine learning and statistical word analysis.

method identifiers for a set of full-text articles, ordered by their likelihood of having been used to detect the PPIs described in each article and providing also a text evidence passage for the interaction method. Figure 5 shows a set of example predictions of various degrees of difficulty corresponding to BioCreative III submissions. The training and development set were derived from annotations provided by databases compliant with the PSI-MI annotation standards, while the BioGRID and MINT database curators carefully prepared the test set. Participating teams went beyond simple term look-up and many of them considered this task as a multi-class classification problem. The best precision obtained by a submission for this task was of 80.00% at a recall of 41.50% (*F*-score of 51.508) (53). The highest *F*-score was of 55.06 (62.46% precision with 55.17% recall) (53).

A common approach followed by participating teams was, in addition to pattern matching techniques, the use of various kinds of supervised machine learning techniques that explored a range of different features. Machine-learning methods tested included Naïve Bayes multiclass classifiers [team 65, (55)], support vector machines [SVMs; teams 81 (56) and 90 (48)], logistic regression [LR; team 69, (53)] and nearest neighbors [team 100, (53)].

Another common practice was based on dictionary extension approaches using manually added terms based on the training data inspection, the use of cross-ontology mapping based on Medical Subject Headings (MeSH) and Unified Medical Language System (UMLS) terms as well as rule-based expansion of the original dictionary of method terms. Most participating teams explored statistical analysis of words, bigrams and collocations present in the training and development set articles. Exact and partial word tokens building the original method term lists were also exploited too. Finally, pattern-matching techniques together with rule-based approaches combined with machine-learning classifier could be successfully adapted for this task.

Team 88 of BioCreative III (53) used a dictionary-based strategy to recover mentions of interaction method terms. As finding exact mentions of method terms results

Database, Vol. 2012, Article ID bas017, doi:10.1093/database/bas017

generally in limited recall, team 70 (53) used approximate string searches for finding method mentions. Another option to boost recall was followed by team 65 (55), which considered sub-matches at the level of words and applied pattern-matching techniques. Such methods are suitable to handle multi-term words, which comprise an important fraction of the PSI-MI terms. This team used a corpus-driven approach to derive conditional probabilities of terms and the detect (56) complemented pattern matching with a sentence classification method relying on SVMs. This type of machine learning method together with logistic regression was also tested by team 90 (48), trying out many features, like type and text of named entities, words proximity to the entities and information on where in a document these entities where mentioned. Team 69 (53) also applied logistic regression for their participating system. They included features that covered term and lexicon membership properties and carried out a global analysis at the level of the documents as well as at the level if individual sentences. A software that directly resulted from participation at the IMT is the OntoNorm framework (57) from team 89 (58) which integrated dictionary-based pattern-matching together with a binary machine-learning classification system and the calculation of mutual information and chi-squared scores of unigrams and bigrams relevant for method terms.

According to an observation of team 100 (53), how competitive a given strategy was depended heavily on the actual PSI-MI term. They therefore used a PSI-MI term specific knowledge-based approach, applying for instance pattern matching approached for some terms, while others were detected through a nearest neighbors method.

Conclusions

The availability of text-mining tools can assist scientific curation in many ways, from the selection of the relevant literature to greatly facilitate the completion of a database entry (saving a conspicuous amount of time). Furthermore, there is a lot of ferment in the area of ontology driven annotation of biomedical literature as witnessed by the 'Beyond the PDF' initiative (59).

The whole BioCreative experience highlighted that in order to obtain substantial advances in the development of text-mining methodologies, it is necessary to develop close collaboration among different communities: text miners, database curators and ontology developers. In particular, such vicinity instilled into the text-mining community a more mature comprehension of crucial biological questions (e.g. gene species annotation) and the necessity to make methods and results more easily accessible to biologist and database annotators (e.g. user-friendly visualization tools).

What is crucial for text miners in the development of more efficient predictive algorithms is the availability of a large corpus of manually annotated training data. Ideally, such text-bound annotations should cover a variety of representative text phrases mapped to the same concept. How feasible it is to generate large enough annotated text data sets for complex annotation types at various levels of granularity is still unclear.

This necessity prompted various initiatives to compile *ad hoc* curated data sets [e.g. the GENIA corpus (60)]. Unfortunately, such collections are usually created as a specific resource for natural processing language sciences but are not suitable for all applications. Furthermore, their creation is extremely laborious resulting in relatively small collections. Another effort to provide syntactic and semantic text annotations of biomedical articles using various ontologies is the CRAFT corpus initiative, which aims to provide concept annotations from six different ontologies including GO and the Cell Type Ontology (CL) (61). One of the merits of BioCreative has been to permit the public deposition of annotated corpora. BioCreative has also been very effective in identifying the main areas of application, limitations and goals of text mining in the area of protein/gene function and interactions.

Data sets routinely annotated by databases are ideal candidates for the compilation of large reference data sets. Unfortunately, databases do not capture the textual passages linked to the experimental evidence and this represents a significant hurdle to the development of text-mining suites. In addition, it is still very hard to convince databases and publishers to provide access to text-bound annotations (manual text labelling), but this has also difficulties related to technical and organizational aspects.

In this respect, the biological ontologies may represent a powerful tool to overcome these limitations. The identification of the experimental methods (as described by PSI-MI) linked to protein interactions can be an important resource facilitating the retrieval of protein interactions, but this requires an extra effort to increase the aliases of the dictionary and/or to identify the critical textual passages.

Ideally, an effective strategy to effectively employ bio-ontologies in text-mining technologies would consist of an in-depth annotation of text passages associated with the ontology terms, thus creating an effective dictionary. This could serve as valuable data for machine learning approaches as well as be useful for automatic term extraction techniques to enrich iteratively the lexical resources behind the original ontologies. On the other hand, there is a need to consider more closely the use of text-mining methods for the actual development and expansion of controlled vocabularies and ontologies, relying for instance on corpus-based term acquisition. Such an approach has shown promising results for the metabolomics (29) and animal behavior (62) domains where term recognition

and filtering methods using generic software tools has been explored. At the current stage, it is possible to say that the BioCreative effort has successfully promoted the exploration of a set of sophisticated methods for the automatic detection of ontology concepts in the literature, some of which can generate promising results. What is still missing is to determine more systematically which methods are more robust or competitive for particular types of concepts or terms as well as to have more granular annotations at the level of labeling textual term evidences. Ultimately, the incorporation of concept recognition systems into text-mining tools will greatly depend on their availability and flexibility to handle more customized term lists and ontology relation types.

Acknowledgements

We would like to thank Lynette Hirschman and Christian Blaschke for their active feedback in the BioCreative tasks described in this article.

Funding

This work was supported by the National Center for Research Resources (NCRR) and the Office of Research Infrastructure Programs (ORIP) of the National Institutes of Health (NIH) (1R01RR024031 to M.T.) (R24RR032659 to M.T.); the Biotechnology and Biological Sciences Research Council (BB/F010486/1 to M.T.); the Canadian Institutes of Health Research (FRN 82940 to M.T.); the European Commission FP7 Program (2007-223411 to M.T.); a Royal Society Wolfson Research Merit Award (to M.T.); the Scottish Universities Life Sciences Alliance (to M.T.); Projects BIO2007 (BIO2007-666855) (to M. K. and A.V.), CONSOLIDER (CSD2007-00050) (to M. K. and A.V.), MICROME (Grant Agreement Number 222886-2) (to M. K. and A.V.). Funding for open access charges: National Institutes of Health (1R01RR024031).

Conflict of interest. None declared.

References

1. Neumann,B., Walter,T., Heriche,J.K. *et al*. (2010) Phenotypic profiling of the human genome by time-lapse microscopy reveals cell division genes. *Nature*, **464**, 721–727.
2. Smogorzewska,A., Desetty,R., Saito,T.T. *et al*. (2010) A genetic screen identifies FAN1, a Fanconi anemia-associated nuclease necessary for DNA interstrand crosslink repair. *Mol. Cell*, **39**, 36–47.
3. Birney,E., Stamatoyannopoulos,J.A., Dutta,A. *et al*. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.
4. Seringhaus,M. and Gerstein,M. (2008) Manually structured digital abstracts: a scaffold for automatic text mining. *FEBS Lett.*, **582**, 1170.
5. Galperin,M.Y. and Cochrane,G.R. (2011) The 2011 Nucleic acids research database issue and the online molecular biology database collection. *Nucleic Acids Res.*, **39**, D1–D6.
6. Stein,L. (2001) Genome annotation: from sequence to biology. *Nat. Rev. Genet.*, **2**, 493–503.
7. Elsik,C.G., Worley,K.C., Zhang,L. *et al*. (2006) Community annotation: procedures, protocols, and supporting tools. *Genome Res.*, **16**, 1329–1333.
8. Huss,J.W. III, Lindenbaum,P., Martone,M. *et al*. (2010) The Gene Wiki: community intelligence applied to human gene annotation. Nucleic Acids Res., **38**, D633–D639.
9. Leitner,F., Chatr-aryamontri,A., Mardis,S.A. *et al*. (2010) The FEBS Letters/BioCreative II.5 experiment: making biological information accessible. *Nat. Biotechnol.*, **28**, 897–899.
10. Superti-Furga,G., Wieland,F. and Cesareni,G. (2008) Finally: the digital, democratic age of scientific abstracts. *FEBS Lett.*, **582**, 1169.
11. Baumgartner,W.A. Jr, Cohen,K.B., Fox,L.M. *et al*. (2007) Manual curation is not sufficient for annotation of genomic databases. *Bioinformatics*, **23**, i41–i48.
12. Rebholz-Schuhmann,D., Kirsch,H., Arregui,M. *et al*. (2006) Protein annotation by EBIMed. *Nat. Biotechnol.*, **24**, 902–903.
13. Couto,F.M., Silva,M.J., Lee,V. *et al*. (2006) GOAnnotator: linking protein GO annotations to evidence text. *J. Biomed. Discov. Collab.*, **1**, 19.
14. Dowell,K.G., McAndrews-Hill,M.S., Hill,D.P. *et al*. (2009) Integrating text mining into the MGI biocuration workflow. *Database*, Vol. 2009, Article ID bap019, doi:10.1093/database/bap019.
15. Wiegers,T.C., Davis,A.P., Cohen,K.B. *et al*. (2009) Text mining and manual curation of chemical-gene-disease networks for the comparative toxicogenomics database (CTD). *BMC Bioinformatics*, **10**, 326.
16. Alterovitz,G., Xiang,M., Hill,D.P. *et al*. (2010) Ontology engineering. *Nat. Biotechnol.*, **28**, 128–130.
17. Hirschman,L., Yeh,A., Blaschke,C. *et al*. (2005) Overview of BioCreAtIvE: critical assessment of information extraction for biology. *BMC Bioinformatics*, **6 (**Suppl 1), S1.
18. Leitner,F., Mardis,S.A., Krallinger,M. *et al*. (2010) An Overview of BioCreative II.5. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **7**, 385–399.
19. Aranda,B., Achuthan,P., Alam-Faruque,Y. *et al*. (2010) The IntAct molecular interaction database in 2010. *Nucleic Acids Res.*, **38**, D525–D531.
20. Ceol,A., Chatr-Aryamontri,A., Licata,L. *et al*. (2010) MINT, the molecular interaction database: 2009 update. *Nucleic Acids Res.*, **38**, D532–D539.
21. Salwinski,L., Miller,C.S., Smith,A.J. *et al*. (2004) The database of interacting proteins: 2004 update. *Nucleic Acids Res.*, **32**, D449–D451.
22. Stark,C., Breitkreutz,B.J., Chatr-Aryamontri,A. *et al*. (2011) The BioGRID interaction database: 2011 update. *Nucleic Acids Res.*, **39**, D698–D704.
23. Mewes,H.W., Ruepp,A., Theis,F. *et al*. (2011) MIPS: curated databases and comprehensive secondary data resources in 2010. *Nucleic Acids Res.*, **39**, D220–D224.
24. Chautard,E., Fatoux-Ardore,M., Ballut,L. *et al*. (2011) MatrixDB, the extracellular matrix interaction database. Nucleic Acids Res., **39**, D235–D240.

25. Goll,J., Rajagopala,S.V., Shiau,S.C. *et al*. (2008) MPIDB: the microbial protein interaction database. *Bioinformatics*, **24**, 1743–1744.
26. Kerrien,S., Orchard,S., Montecchi-Palazzi,L. *et al*. (2007) Broadening the horizon–level 2.5 of the HUPO-PSI format for molecular interactions. *BMC Biol.*, **5**, 44.
27. Cote,R.G., Jones,P., Apweiler,R. *et al*. (2006) The ontology lookup service, a lightweight cross-platform tool for controlled vocabulary queries. *BMC Bioinformatics*, **7**, 97.
28. Smith,B., Ashburner,M., Rosse,C. *et al*. (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.*, **25**, 1251–1255.
29. Spasic,I., Schober,D., Sansone,S.A. *et al*. (2008) Facilitating the development of controlled vocabularies for metabolomics technologies with text mining. *BMC Bioinformatics*, **9 (**Suppl. 5), S5.
30. Bodenreider,O. (2008) Biomedical ontologies in action: role in knowledge management, data integration and decision support. *Yearb. Med. Inform.*, 67–79.
31. Tirmizi,S.H., Aitken,S., Moreira,D.A. *et al*. (2011) Mapping between the OBO and OWL ontology languages. *J. Biomed. Semantics*, **2 (**Suppl. 1), S3.
32. Ashburner,M., Ball,C.A., Blake,J.A. *et al*. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
33. Hill,D.P., Smith,B., McAndrews-Hill,M.S. *et al*. (2008) Gene ontology annotations: what they mean and where they come from. *BMC Bioinformatics*, **9 (**Suppl. 5), S2.
34. McCray,A.T., Browne,A.C. and Bodenreider,O. (2002) The lexical properties of the gene ontology. *Proc. AMIA Symp.*, 504–508.
35. Beisswanger,E., Poprat,M. and Hahn,U. (2008) Lexical properties of OBO ontology class names and synonyms. In: *Proceedings of the Third International Symposium on Semantic Mining in Biomedicine*. Turku, Finland, pp. 13–20.
36. Blaschke,C., Leon,E.A., Krallinger,M. *et al*. (2005) Evaluation of BioCreAtIvE assessment of task 2. *BMC Bioinformatics*, **6 (**Suppl. 1), S16.
37. Muller,H.M., Kenny,E.E. and Sternberg,P.W. (2004) Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biol*, **2**, e309.
38. Jonquet,C., Shah,N.H. and Musen,M.A. (2009) The open biomedical annotator. *Summit on Translat Bioinforma*, **2009**, 56–60.
39. Rzhetsky,A., Iossifov,I., Koike,T. *et al*. (2004) GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data. *J. Biomed. Inform.*, **37**, 43–53.
40. Domeo. http://annotationframework.org/ (14 March 2012, date last accessed).
41. Xuan,W., Dai,M., Mirel,B. *et al*. (2009) Open biomedical ontology-based Medline exploration. *BMC Bioinformatics*, **10 (**Suppl. 5), S6.
42. Yeh,A., Morgan,A., Colosimo,M. *et al*. (2005) BioCreAtIvE task 1A: gene mention finding evaluation. *BMC Bioinformatics*, **6 (**Suppl. 1), S2.
43. Hirschman,L., Colosimo,M., Morgan,A. *et al*. (2005) Overview of BioCreAtIvE task 1B: normalized gene lists. *BMC Bioinformatics*, **6 (**Suppl. 1), S11.
44. Magrane,M. and Consortium,U. (2011) UniProt Knowledgebase: a hub of integrated protein data. *Database*, Vol. 2011, Article ID bar009, doi:10.1093/database/bar009.
45. Krallinger,M., Leitner,F., Rodriguez-Penagos,C. *et al*. (2008) Overview of the protein-protein interaction annotation extraction task of BioCreative II. *Genome Biol.*, **9 (**Suppl. 2), S4.
46. Chatr-aryamontri,A., Kerrien,S., Khadake,J. *et al*. (2008) MINT and IntAct contribute to the Second BioCreative challenge: serving the text-mining community with high quality molecular interaction data. *Genome Biol.*, **9 (**Suppl. 2), S5.
47. Piao,S., McNaught,J. and Ananiadou,S. (2008) Clustering related terms with definitions. In: *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*. Marrakech, Morocco, pp. 2013–2019.
48. Wang,X., Rak,R., Restificar,A. *et al*. (2011) Detecting experimental techniques and selecting relevant documents for protein-protein interactions from biomedical literature. *BMC Bioinformatics*, **12 (**Suppl. 8), S11.
49. Pesquita,C., Faria,D., Falcao,A.O. *et al*. (2009) Semantic similarity in biomedical ontologies. *PLoS Comput. Biol.*, **5**, e1000443.
50. Fellbaum,C., Hahn,U. and Smith,B. (2006) Towards new information resources for public health–from WordNet to MedicalWordNet. *J. Biomed. Inform.*, **39**, 321–332.
51. Resnik,P. (1995) Using information content to evaluate semantic similarity in a taxonomy. In: *Proceedings of the 14th International Joint Conference on Artificial Intelligence*. Montréal, Canada, Vol. I, pp. 448–453.
52. Arighi,C.N., Lu,Z., Krallinger,M. *et al*. (2011) Overview of the BioCreative III Workshop. *BMC Bioinformatics*, **12 (**Suppl. 8), S1.
53. Krallinger,M., Vazquez,M., Leitner,F. *et al*. (2011) The protein-protein interaction tasks of BioCreative III: classification/ranking of articles and linking bio-ontology concepts to full text. *BMC Bioinformatics*, **12 (**Suppl. 8), S3.
54. Chatr-Aryamontri,A., Winter,A., Perfetto,L. *et al*. (2011) Benchmarking of the 2010 BioCreative Challenge III text-mining competition by the BioGRID and MINT interaction databases. *BMC Bioinformatics*, **12 (**Suppl. 8), S8.
55. Schneider,G., Clematide,S. and Rinaldi,F. (2011) Detection of interaction articles and experimental methods in biomedical literature. *BMC Bioinformatics*, **12 (**Suppl. 8), S13.
56. Lourenco,A., Conover,M., Wong,A. *et al*. (2011) A linear classifier based on entity recognition tools and a statistical approach to method extraction in the protein-protein interaction literature. *BMC Bioinformatics*, **12 (**Suppl. 8), S12.
57. Onto Norm framework. https://sourceforge.net/projects/ontonorm (14 March 2012, date last accessed).
58. Agarwal,S., Liu,F. and Yu,H. (2011) Simple and efficient machine learning frameworks for identifying protein-protein interaction relevant articles and experimental methods used to study the interactions. *BMC Bioinformatics*, **12 (**Suppl. 8), S10.
59. Beyond the PDF. http://sites.google.com/site/beyondthepdf/ (14 March 2012, date last accessed).
60. Kim,J.D., Ohta,T., Tateisi,Y. *et al*. (2003) GENIA corpus–semantically annotated corpus for bio-textmining. *Bioinformatics*, **19 (**Suppl. 1), i180–i182.
61. Bada,M., Hunter,L.E., Eckert,M. *et al*. (2010) An overview of the CRAFT concept annotation guidelines. In: *Proceedings of the Fourth Linguistic Annotation Workshop*. Uppsala Sweden, pp. 207–211.
62. Brewster,C., Jupp,S., Luciano,J. *et al*. (2009) Issues in learning an ontology from text. *BMC Bioinformatics*, **10 (**Suppl. 5), S1.