BRIEFINGS IN BIOINFORMATICS. VOL 9. NO 6. 466–478
Advance Access publication December 6, 2008
doi:10.1093/bib/bbn043

Facts from text: can text mining help to scale-up high-quality manual curation of gene products with ontologies?

Rainer Winnenburg, Thomas Wächter, Conrad Plake, Andreas Doms and Michael Schroeder

Submitted: 23rd May 2008; Received (in revised form): 10th September 2008

Abstract

The biomedical literature can be seen as a large integrated, but unstructured data repository. Extracting facts from literature and making them accessible is approached from two directions: manual curation efforts develop ontologies and vocabularies to annotate gene products based on statements in papers. Text mining aims to automatically identify entities and their relationships in text using information retrieval and natural language processing techniques. Manual curation is highly accurate but time consuming, and does not scale with the ever increasing growth of literature. Text mining as a high-throughput computational technique scales well, but is error-prone due to the complexity of natural language. How can both be married to combine scalability and accuracy? Here, we review the state-of-the-art text mining approaches that are relevant to annotation and discuss available online services analysing biomedical literature by means of text mining techniques, which could also be utilised by annotation projects. We then examine how far text mining has already been utilised in existing annotation projects and conclude how these techniques could be tightly integrated into the manual annotation process through novel authoring systems to scale-up high-quality manual curation.

Keywords: *text mining; data curation; ontology generation; entity recognition; GO annotation; authoring systems*

MOTIVATION

Controlled vocabularies allow scientists to communicate in a defined and unambiguous way, where all partners agree on the same usage of language to minimise the chances for misunderstanding leading to a better transfer of information. Ontologies [1, 2] additionally define relationships between the concepts used. Taxonomic relations enable scientists to communicate on different levels of granularity choosing the level suiting their purposes best. While one scientist might only refer to the 'cell' as a whole, others will specify the location as 'organelle' or even more specifically as 'endosome'. Other relations in ontologies allow to formulate complex statements such as temporal dependencies during development, causes of changes of state, or even periodically re-occurring events, which can be used to describe and evaluate data. The Gene Ontology (GO) [3], for example, provides concepts describing biological processes, molecular functions and cellular components, which are used to annotate gene products. Under the umbrella of the Open Biomedical

Corresponding author. Michael Schroeder, Biotechnology Center, Technische Universität Dresden, Tatzberg 47-49, 01307 Dresden, Germany. Tel: +49 351 463 40062; Fax: +49 351 463 40061; E-mail: ms@biotec.tu-dresden.de

Rainer Winnenburg is a PhD Student in the Bioinformatics Group at the BIOTEC of TU Dresden, Germany. His research interests are focused on protein interactions in regard to mutations and diseases.

Thomas Wächter is a PhD Student in the Bioinformatics Group at the BIOTEC of TU Dresden, Germany. His research interests are focused on automatic support for ontology development and semantic search technologies.

Conrad Plake is a PhD Student in the Bioinformatics Group at the BIOTEC of TU Dresden, Germany. His interest focuses on text mining, particularly named entity identification and information extraction with applications to biology and biomedicine.

Andreas Doms is a PhD Student in the Bioinformatics Group at the BIOTEC of TU Dresden, Germany. His principal interest focuses on the ontology-based literature search and text mining.

Michael Schroeder is a Professor in Bioinformatics at the BIOTEC of TU Dresden, Germany. He is interested in protein interactions and functional annotation based on text mining.

Ontologies (OBO) consortium [4], dozens of other ontologies have emerged, which follow shared design principles.

Many of the abovementioned ontologies are used to manually annotate gene products with ontology concepts based on evidence in literature, as, for example, done in the Gene Ontology Annotation (GOA) project [5]. Part of the GO consortium are numerous annotation projects for all important model organisms concerned with creating specific annotation repositories such as the Mouse Genome Informatics (MGI) [6], FlyBase for Drosophila [7], Wormbase for C. elegans [8], the Rice Genome Annotation project [9], or The Arabidopsis Information Resource TAIR [10]. Besides using and working on GO, many of these efforts also develop ontologies such as the Mammalian Phenotype Ontology at MGI, development and anatomy vocabularies in FlyBase, and the ontology used by Textpresso in Wormbase. Figure 1 describes such a typical annotation process and shows where automated text mining methods can support the human curator. As a prerequisite for the retrieval of relevant literature for genes of interest, e.g. of a certain species, genes and gene products have to be automatically identified in text. This task is equivalent to the question that most biomedical literature search engines are trying to answer. Thus, the underlying text mining approaches in both scenarios are the same. In addition, ontology concepts co-occurring in these texts can be recognised, which are then proposed to the human curator in relation to the identified genes. The human curator chooses from all proposed gene annotations those which she or he regards as correct and most precise, discards false predictions, and adds further annotations where appropriate. The curated annotations are typically entered to a database given the corresponding publication as reference. Linking gene products and ontology concepts supports users in finding gene products by querying the ontology [11] and allows for statistical analyses of large sets of gene products from high-throughput experiments to identify significantly enriched annotations [12, 13].

Although manual curation of gene products with ontology concepts ensures the highest possible quality, three problems arise: (i) Scalability. With progress in sequencing technology and the growth of literature, manual curation cannot keep pace with the growth of gene products requiring annotation. Baumgartner *et al.* [14] applied a software engineering metric for evaluating the curation progress and the

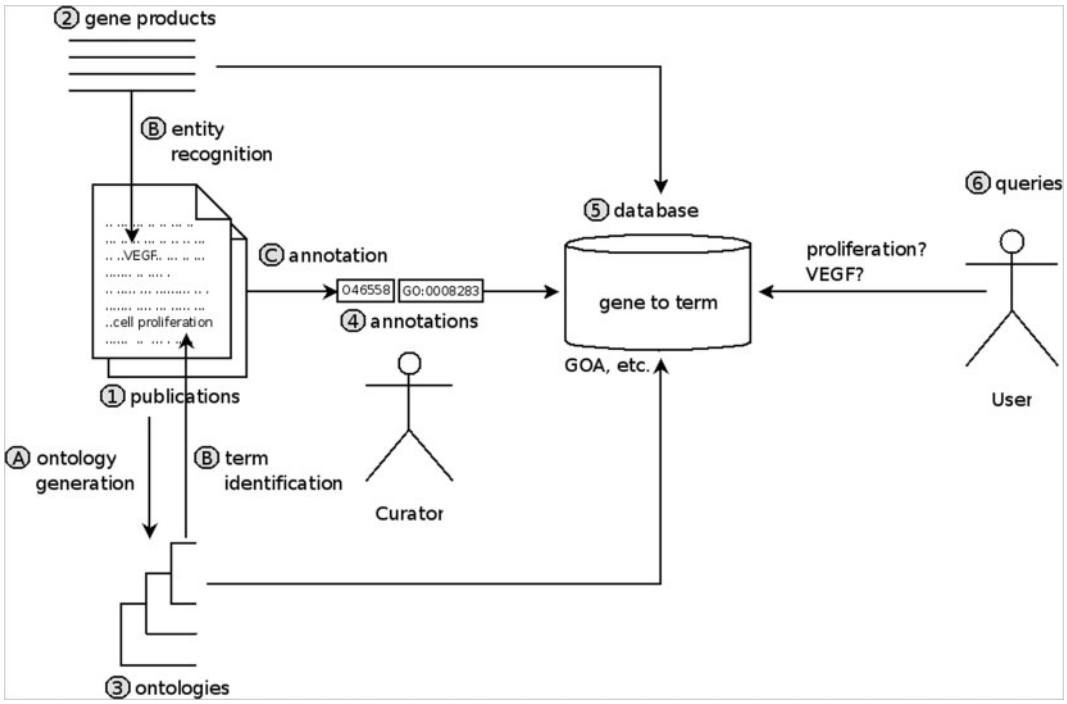

Figure 1: Integration of text mining and ontology development to curation process: the curator reads papers (1) and identifies gene products (2) and terms from ontologies (3), which have been proposed by text mining methods (A–C). Annotations (4) are formulated and added to a database (5), which can be queried by the end user (6).

completeness of biomedical knowledge bases, and concluded that completion time for reliable manual curation ranges up to decades for some organisms. (ii) Evolution. Ontologies change over time between several releases [15]. Concepts may be added, replaced, or moved in the graph, being assigned to new parent nodes. Concepts used in a former annotation may not be valid or specific enough any more. Annotation efforts like GOA ensure backward compatibility of annotations. However, they do not ensure completeness, especially with new concepts, and not all gene products may necessarily be annotated with all relevant concepts. To ensure completeness and the consistent use of old and new concepts, it is necessary to re-annotate all gene products with new concepts added. This is an important prerequisite for searching and clustering related genes in one organism or across species. Example (taken from [15]): since May 2005, node GO:0031399 (regulation of protein modification) exists, with two qualifying child nodes (positive/negative regulation) and one specialising node (regulation of protein ubiquitination). GO:0043538 (regulation of actin phosphorylation) was introduced as a child node to GO:0031399 half a year later. For some of the gene products, which were annotated with the general term GO:0031399, the newly introduced term might be more precise. However, contemplable gene products have to be identified, and a decision has to be made on an individual basis. (iii) Inter-annotator agreement. Manual annotation is subjective. Dependent on their scientific background or experience, different curators might choose more or less specific concepts for the annotation of a gene product or make their decision on different aspects mentioned in the literature. In an inter-annotator agreement study for manual GO curation of the GOA project, Camon *et al.* [16] showed that there is only a 39% chance of three curators selecting the same GO concept for a gene product.

Automated methods from natural language processing (NLP) and information retrieval (IR) do not suffer from these problems. They are fast and can be applied to large sets of text. Most problems concerning evolution can be overcome by re-computing annotations at any time, reporting changes in comparison to previous versions. The inter-annotator problem can be addresses by systems like the BioCreative MetaServer [17], which aims at unifying results from various automated text mining approaches for the annotation of PubMed abstracts. However, it has to be stated that automated systems do not yield as high quality results as manual annotation. In the remainder, we review how automated text mining methods can support manual curation with the goal of combining the quality of the latter with the scalability of the former. We follow the scheme depicted in Figure 1. First, we will review how text mining can help to build ontologies from text (Figure 1A), next how to identify gene products (Figure 1B), ontology concepts (Figure 1B), and their relations (Figure 1C) in text. Then, we briefly discuss online text mining tools, and we conclude by outlining how text mining integrated into authoring tools could pave the way to large-scale high-quality annotation.

ONTOLOGY LEARNING

Since designing an ontology is a cost- and labour-intensive process, automating parts of the design process is important. Ontology learning [18, 19] aims to solve this problem by supporting the discovery of terms, synonyms, concepts, and taxonomic and non-taxonomic relationships. By terms we refer to phrases from natural language which can be simple nouns as 'cell' or 'growth', or noun phrases like 'early endosome', 'epidermal growth factor', which are essentially single grammatical units containing a noun as a main word, and here, 'endosome' and 'factor'. More complex terms can be composed of several noun phrases like 'endosomal sorting complex required for transport proteins' or 'transcription factors involved in the regulation of endocytosis'. The concept, as used here, groups a number of terms and corresponding synonyms to a semantic unit, which can be referred to by all assigned terms. Concepts are defined by a natural language definition, and have a representative label (usually but not necessarily identical to one of the terms). (i) Terms. To find the vocabulary, automatic term recognition methods help to find single words and compound nouns, e.g. 'early endosome' and 'epidermal growth factor'. However, they fail to find multi-word phrases with definitional character, like 'hydrolase acting on ester bonds', 'endosomal sorting complex required for transport proteins' and 'chromosome migration to spindle pole during meiosis', which are commonly used throughout the GO. Fortunately, the majority of biomedical terms consist of compound nouns, e.g. almost 90% of the biomedical terms in the GENIA corpus [20] are

compounds [21, 22]. Therefore, many term recognition methods retrieve multi-word phrases as candidate terms using noun phrase chunking. The likelihood for terms being domain relevant is estimated using overall corpus frequencies [23, 24] or the internal structure of the terms themselves [24]. Apart from compounds, named entities such as gene or protein names certainly play an important role in biomedical terminology and are usually found using dictionaries [25, 26]. Here, disambiguation plays an important role, as [27] reported ambiguities from gene names to general English in the range from 2% to 32% depending on organisms and nomenclatures studied. A third and very specialised approach uses handcrafted syntactic rules to find terms. Manual creation of syntactic rules can work very well for small-scale examples. During the tasks of collecting cellular component specific or cell type specific terminology one could search for words ending with ''some'' aiming to find ''endosome'', ''lysosome'' etc., or for words ending with ''blast'' to find the terms ''osteoblasts'' and ''cytoblasts'' respectively. Nevertheless the creation of such patterns is usually very time consuming and lacks transferability to other domains. In a use case, Alexopoulou *et al*. [23] compared four different term generation tools and showed that they produce over 80% relevant terms within the first 50 and over 70% within the first 200 predicted candidate terms. In general, there will always be an upper limit to term recognition, as not all suitable terms will appear anywhere in text. As an example we found, that less than 20% of GO terms appear in PubMed abstracts and could hence be predicted. Out of the remaining terms, some could possibly be predicted because they have a definitional character, such as hydrolase acting on ester bonds, which comprises two noun phrases and a relation. For 53% of GO terms, the contained noun phrases appear in a sentence in PubMed. Currently, no mature methods exist for finding such composite terms from text. For the GO, methodologies are available for exploiting the structure of existing concepts to successfully propose new terms [28]. An overview about term recognition was given in [21]. (ii) Synonyms, abbreviations and definitions. Concepts are identified in text by finding associated terms (a preferred term, which acts as a representative label for the concept, synonyms, and abbreviations as a special class of synonyms). Synonyms are important as authors and annotators may use equivalent, but different terminology. For example, authors might refer to the concept fever in different ways. Some texts will mention the term fever itself, others the Latin name pyrexia. Furthermore, apoptosis and programmed cell death are used synonymously in literature. Often—as in these examples—the terms are not exact synonyms, but have slightly broader or narrower senses. A qualitatively good source for synonyms is WordNet [29], a lexical database of English where nouns, verbs, adjectives, and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Automatically finding such synonyms without such a resource is a difficult task. Mccrae and Collier [30] reported for a small-scale experiment on learning regular expression patterns for synonymy a very low recall of 7% at 100% precision when validating against WordNet. The same experimental results validated against the UMLS showed over 40% recall at 90% precision. For their own automatic pattern discovery method, the authors reported 29% recall at 73% precision, which corresponds to a study in [31] reaching 21–27% coverage at a precision greater than 70%. Unlike synonyms, abbreviations [32], e.g. the technique *RNAi* standing for RNA interference or more precisely for Ribonucleic acid interference, can be accurately identified: References [1, 33, 34] report all precision and recall above 90% and Okazaki et al. [35] reported of 78% and 85%, respectively. There is little work on the generation of definitions. However, Klavans and Muresan [36] developed a rule-based system extracting definitions from online resources. They report a precision and recall of 87% and 75%, respectively. We conclude that automatic methods can play an important role in finding synonyms, abbreviations or definitions and will be included in appropriate ontology engineering tools soon. (iii) Finding taxonomic relationships. Comparing annotations with different levels of detail requires the taxonomic structure of an ontology. There are lexico-syntactic and statistical methods to extract taxonomic relationships such as is-a and part-of from text. In general, the most challenging problem for such methods is the fact that many relationships are not made explicit in text. An example for a lexico-syntactic method is Hearst patterns [37], like 'X such as Y'. With these patterns one can infer, e.g. from the text fragments 'organelles such as mitochondria', that mitochondria are organelles. Pattern-based methods show typically high precision around 90%, but a low recall of 10%. An example for the statistical methods

is reported in [38]. Here, the decision on the existence of a relationship between two concepts depends on the measured co-occurrence of the concepts. The concept that shows more independence of the other will be suggested as parent in that relation. Another approach, which can generate ontologies from the concept usage in documents, is formal concept analysis [39, 18] reaching an *f*-measure of 39–45%. Generally, statistical approaches reach an *f*-measure of below 50% [40, 41].

The above success rates suggest that ontology learning methods are not and will not be able to create ontologies fully automatically as (1) the selection of relevant text is an equally hard problem as the learning process itself, (2) not all terms used in ontologies are contained in texts nor are intended to be contained, and (3) relationship extraction remains a hard problem. Nevertheless, ontology learning is able to improve the manual ontology creation process significantly as it is able to suggest terminology with direct evidence in literature and can discover rules or statistical relationships between concepts. The usefulness of an ontology learning method itself strongly depends on the final step - its integration in ontology editing tools. Tools such as Text2onto [42] already address many subtasks of the ontology learning process and can save a lot of time when serving as a starting point for subsequent manual refinements.

NAMED ENTITY AND CONCEPT IDENTIFICATION

Annotating gene products with ontology concepts based on evidence in literature requires the identification of entities such as the gene products and concepts from an ontology in text. Two problems can be distinguished: automatically recognising a text passage mentioning an entity or concept and identifying the entity/concept itself.

Gene name identification

Gene name recognition and identification are difficult, as there is an immense variety of gene names and naming conventions. For example, human genes have on average 5.55 different names [43]. Names are often abbreviations, database identifiers, or functional descriptions. Especially, abbreviations and functional descriptions lead to problems of ambiguity. Fundel and Zimmer [44] found that 2.4% of gene names in FlyBase are ambiguous as they are common English words such as 'and', 'the', or 'this'. Other examples of ambiguity arise from functional descriptions such as denoting a protein by its weight. The name p54 indicates that this protein has a peak at 54 kDa in a mass spectrum. However, this is the case for other proteins as well, and the name p54 refers in human alone to five different proteins [45]. Furthermore, usage patterns for protein names change over time. Tamames and Valencia [46] report a gene that is mostly referred to as PVR from the mid-1990s and as CD155 from 2000 onwards. Although there are standardisation bodies assigning official gene names, authors sometimes introduce their own names to emphasise a certain function of the gene. In [47], the authors refer to the yeast genes SRC1 and YDR458C as HEH1 and HEH2 to reflect their helix-extension-helix secondary structure. Besides variety and ambiguity of names, basic natural language processing problems arise such as conjunctions like freac-1 to freac-7, which mention two genes explicitly and five implicitly. Baumgartner *et al.* [48] found that about 8% of gene names in a representative dataset contained some form of conjunctions.

Recently, substantial progress has been made in the field of gene name recognition and identification. The BioCreative challenges [49, 50] defined benchmark data sets for both tasks in fruit fly, human, mouse, and yeast. The best results for gene name identification range from success rates of around 80% for mouse, human, and fruit fly to over 90% for yeast. For a simple problem of gene name recognition, results are around 87% [51–53].

Concept identification and gene annotation

Identifying ontology concepts by finding their associated terms in text is equally challenging as gene name identification, but for different reasons. As discussed in section Ontology learning, ontology terms often do not appear literally in text even if techniques such as stemming are applied. Searching in PubMed for an exact match of the GO term 'alkaline phosphatase activity' retrieves 10 times less documents than a search for alkaline phosphatase. Very long and descriptive terms such as for the GO term GO:0016706 'oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen, 2-oxoglutarate as one donor, and incorporation of one atom each of oxygen into both

donors' cannot be found literally in text. However, there are some approaches on identifying such ontology concepts, such as the following four: Doms and Schroeder [55] developed a method based on sequence alignment. Terms and text are seen as sequences of stemmed words, which have to be aligned. Matches of words are weighted by the words frequency in the ontology. A similar approach is used in [56] neglecting the sequence information and representing text and term as bag of words. For the task of finding GO annotations for proteins their system achieved precision rates between 7% and 15% on the BioCreative I task 2.1 data set. Another approach by Ruch [57] uses regular expressions over part-of-speech tags and a vector space model, which computes a cosine distance between a term and a text based on the word stems. The system achieved comparable results to other systems used for the GO annotation task of the first BioCreative challenge. Cakmak and Ozsoyoglu [54] combined textual patterns with the use of WordNet. The initial patterns have a left and a right end around a token or phrase that is part of the label of the GO term and occurs frequently in text. On a set of 40 GO terms, their system achieved a precision of 78% at a recall rate of 46% annotating genes in GenBank.

The integration of data from other resources can further improve precision of the automatic annotation process. Jaeger *et al.* [58] integrate conserved protein interactions to assign known functional annotations to yet uncharacterised orthologs. The service SherLoc integrates text and sequence-based features to predict the subcellular localisation of eukaryotic proteins [59].

The problem of ambiguity applies to terms as well. Examples are GO terms such as development, spindle, cell, envelope, which all have different senses. One approach to tackle such ambiguities uses co-occurring terms for disambiguation [60], which can achieve success rates of 80%.

Figure 2 summarises to which extend automated methods can fulfil tasks relevant for annotation work. With regard to ontology learning, relevant terms can be generated with acceptable accuracy, abbreviations can be found with high accuracy, whereas identification of synonyms and definitions, as well as finding taxonomic relationships are still open problems. The recognition of gene names in text can be solved with 80% and better success rates, thus approaching human inter-annotator agreement. Identifying concepts in text is more challenging since ontology concept labels or associated terms often do not appear literally or approximately in text. However, the integration of data from other resources can significantly help to solve specific tasks such as the annotation of gene products. These theoretical results indicate that text mining is suitable for practical tasks and systems. The next section summarises such a growing number of text mining-based web servers.

BIOMEDICAL SEARCH ENGINES

Information retrieval

Database curators face the problem of collecting the current state of knowledge about entities to be entered from peer-reviewed literature. Thus, the first step in this process is to retrieve all relevant documents, i.e. by querying a biomedical literature database by means of a literature search engine. In Figure 3, we compare currently available online biomedical search engines, which are all suitable to support this first step of the manual curation process.

Literature retrieval

Assuming that a database curator cannot know all synonyms of biological entities, there is a high probability that relevant publications will be missed when performing simple keyword searches on indexed documents. To overcome this problem most engines implicitly extend the tokens of the query by lexical variants, word stems, synonyms, and abbreviations, which is already supported by the standard PubMed web interface. Some systems

	Problem	Benchmark	F-measure
A	term generation	300 documents on lipoprotein metabolism best of four different term generation tools [26]	80% (in top 50)
	synonym detection	based on news articles [34]	13-55%
	abbreviation detection	abbreviation databases from Medline [36,37] and [38]	81-90%
	definitions detection	definition of biomedical terms from Medline [39]	>80% (precision)
	finding taxonomic relations	Hearst-patterns [40], statistical method [41], formal concept analysis [42,21]	<50%
B	gene name recognition	BioCreative I + II data sets on human, mouse, fruit fly, yeast [52,53]	80-90%
C	gene annotation	subset of genes and 40 GO concepts from GenBank [59]	60%

Figure 2: State-of-the-art text mining approaches relevant to annotation tasks (A) ontology generation, (B) named entity and term recognition, (C) annotation, and their current success rates.

	Information Retrieval							Knowledge Retrieval									
	Literature retrieval			Result processing				Semantic processing							Tools integration		
Biomedical Search Engines	Query transformation/ Query refinement	Explore controlled vocabulary	Link to original resultset	Evidence highlighting	Re-ranking	Connection to other related documents	Information organization	Entity recognition	Concurrence/ Relations	Disambiguation	Subsumption/ Reasoning	Summarization	User Dialog	Hypotheses generation	single citation processing	batch processing	batch export
PubMed	1	14, 15, 16	23, 24, 25			41											
ReleMed [68]	7	15	23, 25	31	**36!**												
PubMed PubReMiner	**8!**	15	26				**47!**										
ClusterMed	1	15	23, 24, 25	31, 34	36		**44!**										
BioMetaCluster	**9!**		28	31, 34	36		**44!**										
PubFocus [71]	1		23, 24		**38!**	42	**47!**									60	
HubMed [61]	7, 10		23, 24, 25, **29!**	32, 33	36	**41!**	**48!**	67, 68, 69							57	59	63, 64, 65
BioIE [67]	1		26, 27	31			**47**		72								
CiteXplore	2,**7!**		23,24, 25, 26	32, 33		43		67, 68	72	74					57		64, 65
iHOP [62]	2, 7	**18!**	25, 26, 27!	**32!**	**37, 38!**		46	67	**72!**			83.84				60	
Info-PubMed	12		26, 27!	**32!**			46, 47, 48	67	**72!**				90				
EBIMed [63]			24, 25, 26, 27				47	67.68	**72!**			83			57		
GoPubMed [55]	1, 2, 3, 8	15, 16, **17!**	23, 24, 25	31, **33!**	36, 37	41	**45!**	67, 68, 69	**71!**, 72	74, **76!**	**78!**	81, 82	89				62, 63, 64, 65
AliBaba [64]	1	18	25, 26, 27!	32			**46!**,49	67	72	**75!**			90			58	
XplorMed [69]	6	15	23, 25	34	**36!**		**44!**		72		78					58	
GOAnnotator [70]	12		23, 26		**38!**		47	**68!**	72							58, 60	
Textpresso [65]		17	23, 24, 25	34	36	41		67,**68!**	72								62, 65
Chilibot [66]	11		27	31, 32	39		46, 49	67	**72!**			**84!**	90	**92!**			

(1) PubMed query expansion/refinment: expands MeSH headings and additional vocabularies such as drugs or chemicals, citation metadata, (2) expands gene/protein names with synonyms, (3) offers narrowing/expanding with ontology concepts, (4) language translation of terms, (5) full natural language questions handled, (6) querying with other documents/database cross-references, (7) alternative full text index (Lucene/MySQL), (8) refinement based on metadata derived from initial resultset, (9) meta search in separate databases, (10) refinement based on keywords derived from initial resultset, (11) bypassed normal PubMed query expansion/special PubMed queries, (12) entity specific (genes/proteins)

(14) Search in UMLS, (15) Search in MeSH, (16) Search in Gene Ontology, (17) Browse within Taxonomy/Ontology hierarchy, (18) Browse within identified text occurences, (19) Query history, (20) Permanent profile, (21) Session clipboard

(23) Title, (24) Abstract, (25) external Links, (26) PMID, (27) Evidence sentence, (28) Text snippets, (29) Call external web services

(31) highlighted keywords from query, (32) highlighted biomedical entities/relations, (33) highlighted ontology concepts detected, (34) highlighted vocabular (cluster labels/significant

(36) Re-ranking based on concurrence of keywords, (37) Re-ranking based on concurrence of identified entities, (38) Re-ranking based on external database references or precomputed statistics, (39) Language structure (e.g. conclusive sentences)

(41) Cosin similarity based, (42) based on co-authorship, (43) via author name

(44) hierarchical classification based on distance metrics, (45) hierarchical classification using taxonomies/ontologies, (46) 2D concept graph, (47) tabular statistics, (48) Call external service

(50) graphical sliders, (51) email communication, (52) social tagging, (53) special query language, (54) batch processing, (55) drag&drop GUI

(57) external markup tool

(58) import literature references from external databases curations, (59) visualization using an external tool, (60) external large scale experimental metadata used

(62) XML, (63) RDF, (64) BibTex, (65) Endnote (RIS)

(67) biomedical entities (e.g. gene/proteins), (68) Taxonomy/Ontology terminology, (69) Wikipedia terminology

(71) within abstracts, (72) within sentences

(74) disambiguation for bio-entities, (75) disambiguation for taxonomy/ontology terminology, (76) disambiguation for authors

(78) is-a generalization

(80) significant strings, (81) significant taxonomy/ontology concepts, (82) expert profiles, (83) significant bio-entities, (84) textual synopsis

(87) explicit question answering

(89) question categories, (90) graphical interaction

(92) explicit hypothesis generation

Figure 3: Online available biomedical search engines for advanced information retrieval. The table does not intent to rank the systems according to the number of features.

allow the user to re-define the query according to controlled vocabulary, like GoPubMed [55] linking identified terminology to hierarchical background knowledge allowing for specialisation or generalisation of queries. However, not all identified entities can be browsed, e.g. gene/proteins are not part of a classification.

Result processing

The results to a query (e.g. for a gene) are presented within the web interface of the search engine. In case of the baseline system Pubmed, only the matching papers are subsequently listed. However, most tools specifically provide single textual evidences for detected tokens or providing a link out to the relevant text passage in the original publication (HubMed [61], iHOP [62], EBIMed [63], GoPubMed, AliBaba [64], Textpresso [65], Chilibot [66] and BioIE [67]). Some systems explicitly highlight those sentences of a publication, which provide evidences for identified entities or proposed relationships (iHOP, AliBaba and Chilibot), aiding the curator in immediately deciding whether a document is relevant or not, e.g. if the detected term is only a homonym to the term of interest. Besides presenting results to a query within their own web interface, most search engines also provide links out to the underlying publication, e.g. a link to the abstract in PubMed or to the full text publication via the corresponding Digital Object Identifier (DOI). This allows for easily continuing more detailed curation work on the full text probably in a different tool. The support of re-ranking documents based on identified terminologies like MeSH or GO (ReleMed [68], XplorMed [69], GoPubMed and GOAnnotator [70]) and external database references or pre-computed statistics (PubFocus [71], iHOP), allows for sorting the articles by relevance and level of confidence. Features aggregating information from document sets are well supported. Results are grouped based on distance metrics (ClusterMed, BioMetaCluster and XplorMed) or hierarchies (GoPubMed). Two-dimensional graphs visualise conceptual relationships (AliBaba, Chilibot) and tabular statistics reveal key aspects (PubMed, PubReMiner and PubFocus), which in summary give a first characterisation of the queried term. An ideal search engine would rank the retrieved documents according to the curation guidelines, but this exceeds the capability of standard biomedical search engines.

Knowledge retrieval

The information retrieval techniques described above are already sufficient to retrieve documents at a high recall but the advances in text mining have also sparked the development of novel biomedical search engines, aiding to collect even more relevant documents at a higher precision. Alex *et al*. [72] report how much NLP techniques help the curators during assisted curation and conclude that a system must adapt to the curator as well as to the text. Engines like Textpresso were specifically developed to support manual curation, but others also provide functionality, which exceeds pure literature retrieval, and are as such appropriate for curation tasks. In summary, the text mining driven semantic processing of the retrieved articles identifies genes and gene products from text and relates them explicitly to each other and to additionally identified ontology concepts.

Semantic processing

Biomedical entity recognition identifies gene/protein mentions (HubMed, iHOP, EBIMed, GoPubMed, AliBaba, Textpresso, Chilibot). Furthermore, taxonomy/ontology terminology is detected, (HubMed, EBIMed, GoPubMed, Textpresso) as well as community dictionaries such as Wikipedia entries (HubMed, GoPubMed). A group of biomedical search engines employ semantic processing, i.e. establishing links to formally defined terms, typically in ontologies, with entities from the retrieved documents and reason over them, to support Knowledge Retrieval. These proposed annotations and relations can be used as starting point for the manual curation process of the entities in focus. However, for the precise assignment of entities or ontology concepts to terms found in text disambiguation mechanisms are important (AliBaba, GoPubMed), e.g. for distinguishing between the development of an embryo and the development of an algorithm. Some tools make existing database curations available for further knowledge retrieval (AliBaba, XplorMed), which is a useful feature for bootstrapping the curation process.

Tools that employ some form of relation detection between biomedical entities within single sentences are iHOP, EBIMed, AliBaba, XplorMed, Textpresso and Chilibot. The used taxonomies/ontologies hierarchies are rather small (XplorMed, Textpresso) or sparsely identified (HubMed, EBIMed). With such limited linking to rich

biomedical ontologies reasoning support must be limited. However, Chilibot and iHOP offer textual synopses for detected relations and Chilibot supports hypothesis generation for indirect relations of bio-entities. Although the latter feature might be less interesting in standard curation scenarios where only wet lab supported facts are to be considered, it is nevertheless extremely valuable to extract such tentative or hypothesized information, which would be a starting point for follow up studies. Unfortunately, none of the systems offer argumentation mechanisms, e.g. incorporating external positive and negative evidence. Such mechanisms would be extremely useful for curators on the one hand to eliminate false positives and on the other hand to be aware of which facts are already known for the given entity.

Integration of external tools

More complex text mining methods may increase the automatically extracted information. As long running algorithms can not be applied on huge subsets of documents on-the-fly, external web services analyse single documents, which is a practicable procedure especially if the list of documents is already sorted. Alternatively, top ranked documents can be exported via batch export function (ReleMed, PubMed PubReMiner, ClusterMed, BioMetaCluster, HubMed, CiteXplore, GoPubMed, Textpresso) for the subsequent processing in an alternative environment.

To conclude, search engines can help to reveal unknown facts but it is up to the curator and the curation guidelines to select relevant facts to be stored in a structured schema. Although advanced search engines provide entity recognition, relation extraction and hypothesis generation, ideally supported by external evidences, the extracted knowledge is still error-prone and thus needs to be confirmed by the curators. According to the curation guidelines or depending on the experience of the individual curator, a system generating a high recall, for complex texts about new facts, or high precision for redundant texts about asserted facts might be advantageous. We assume combining the strengths of different search engines increases recall during the document retrieval phase of the curation process. Idealwise, the results of different tools could be combined with each other for further processing, e.g. via the export into various exchange formats.

EVALUATION OF ASSISTED CURATION

We define assisted curation as the transfer from unstructured information (typically text) into structured information (typically databases or ontologies) by human curators, who are assisted by computational methods based on text mining. There are three reasons, why so far only little text mining is used to support manual curation: first, although the results for gene name identification approach the levels of human inter-annotator agreement, relationship extraction still remains an open problem. Second, apart from directly extracting annotations from text, in many cases annotations can very effectively be inferred indirectly from information such as functional domains, motifs, and other supplementary information as discussed in [16]. In theses cases, the ontology term chosen for annotation will not appear in the text at all. Third, so far the focus of biomedical text mining has not been on including the user dynamically in the discovery process [73].

Nonetheless, some examples of systems that support the integration of text mining with manual annotation are Textpresso [65], PaperBrowser [11], GOAnnotator [70] and PreBIND [74]. They can provide the curators with decision support, such as highlighting entities of relevance in text, prepopulating curation front-end fields, controlled vocabularies and ontologies, on-the-fly error-correction, and the removal of redundancies. Textpresso, the ontology-based system for extracting and retrieving information from biomedical text is used by the curators of the model organism C. elegans project. Textpresso is both curation tool and search engine. It is designed to work on its own ontology, based on the GO and expanded by a controlled vocabulary for specific gene names, phenotypes, etc. PaperBrowser has been developed for the drosophila reference database FlyBase. It highlights identified gene names in full text as well as putative relations between noun phrases and these genes. The curator can decide if suggested tokens are actually gene names or not and by this allowing for active learning of the named entity recogniser. Besides GO, Flybase provides own ontologies, e.g. for the description of the anatomy of the fly or development. According to the model curation process (Figure 1), PaperBrowser can be seen as a typical curation tool. GOAnnotator was developed to provide textual evidences for gene products which

have already uncurated automatically generated annotations. The tool links the uncurated annotations to texts extracted from literature thus supporting GO curators in choosing the correct terms. GOAnnotator is utilising the hierarchical structure of GO and can also suggest alternative, i.e. more precise annotations. In contrast, PreBIND is dedicated to the extraction of protein–protein interaction data from the literature. The tool was developed to support the curation of the former Biomolecular Interaction Network Database (BIND) [75], now part of BOND.

The concluding question is how to measure the effectiveness of text mining support for database curation? A first intuitive way would be to evaluate the improvement of speed: [11] reported that FlyBase records were completed faster by about 20% when curators were working on an article in the interactive PaperBrowser, whereas [72] showed, that a maximum reduction of one-third in curation time can be expected, if text mining output is perfectly accurate. The PreBIND system is reported to have reduced the curation time for specific tasks by 70%.

Apart from speed, text mining has proven to be especially helpful in retrieving new relevant articles which would not have been detected by manual literature research using standard literature search engines. Thus systems are triggered to attain a high recall although this is only possible at the cost of a lower precision. Consequently, false positive documents will be presented, which the curator has to identify and remove from the list of candidate articles. A good cut-off between recall and precision has to be found if text mining results are supposed to be useful for curators. GOAnnotator was assessed on a set of 66 UniProt/SwissProt proteins and provided correct evidence text at 93% precision. Alex *et al*. [72] show quite diverse results for different curators, indicating that the usability depends on both the information in the text and the scientific background of the individual curator.

Several shortcomings have to be overcome in automated curation tools. The access to the full text of journal articles, which are critical for comprehensive database annotation by text mining [76], is still problematic for both technical and legal reasons. The incorporation of feedback from authors, which could be established in wiki-based solutions [77], is a reasonable but yet an extremely time-extensive approach. As part of the GeneWays project, a system of supervised machine learning algorithms was developed, which aims to imitate human curation of automatically extracted molecular interactions with a performance close to that of human curators [78]. However, the feasibility was only shown on one specific problem. While Baumgartner *et al*. [14] argue that current manual proteome annotation processes take far too long even for the most important model organisms, Burkhardt *et al*. [79] suggest that manual curation will always be necessary. As a conclusion, it seems that a reasonable and reliable approach for database annotation can only be established by the well-balanced combination of manual and automatic annotation methods, where the task of text mining methods is to speed up and standardise the curation process.

AUTHORING SYSTEMS

At present, journal articles are not well suited for automated information retrieval due to the complexity of natural language they are composed of. This shortcoming has a negative impact not only on in-silico analysis approaches relying on information from literature. It also influences the daily work of a researcher when using literature search engines. In order to make the content of articles machine readable, the technical editors of Royal Society of Chemistry (RSC) Publishing use text mining to annotate chemical compounds in texts and to formulate relationships between two biological entities as structured digital abstracts (SDA). The fact that subsequent information extraction from already published articles is time consuming and still error prone suggests changes regarding the structure of journal articles in the future [80]. In the so-called *FEBS Letters* experiment [81], which was set up in March 2008 in a collaboration between the editorial board of *FEBS Letters*, researchers from the MINT database [82] and text mining experts, the authors themselves were involved in adding additional structured summaries to their manuscripts, reporting protein interactions. The question is how software tools can aid authors of scientific papers in making the essential facts machine accessible.

Envisaged are interactive, semi-automated authoring systems, which incorporate computationally guided paper annotation as an integral part of the editorial process as proposed by Leitner and Valencia [83]. Text mining and data integration techniques can be used to create a first generation of electronically annotated information, which is

normalised and interlinked with data repository identifiers. In an iterative process, the authors correct generated annotations, and add missing facts, which were not found automatically. The modifications of the authors are checked on-the-fly against existing data repositories to ensure the correct usage of controlled vocabularies and identifiers. The direct feedback from the authors can be utilised to automatically further improve the incorporated text mining and machine learning techniques. Inconsistencies in data sources or missing terms, which could potentially be identified during the curation process, should be automatically reported to the correspondent developing communities. New entities like genes, proteins, or protein interactions should be uploaded to suitable databases on acceptance of the paper. Additional relevant information such as sequences or experimental evidence should be supplied by the authors.

If such a process is convenient enough, it might become mandatory for authors. This would eventually lead to a paradigm shift, awarding authors for the explicit retrieval of knowledge in addition to the full text publication itself. As an alternative to the author centred approach, semantic wikis could be the system of choice for removing noise and false positive data produced from automated extraction tools in a community-wide collaborative way. An example for controlled natural language is reported in [84], where a limited form of statements on protein interactions can be interpreted formally.

CONCLUSIONS

If the richness of data characterising scientific text will ever be completely captured in a structured format, then at least manual curation seems necessary. Human curators have the ability like no available software tool to intuitively distinguish relevant data from irrelevant data regarding a given area of interest. Thus, the combination of human expertise and automatic text mining systems has several advantages over purely automated information extraction and is indispensable for achieving best reliability. To achieve a good trade-off between the quality and the quantity, the guiding principle should be as much automated guidance as possible and as little manual curation as necessary. However, the chance that curators will accept automated tools aiding them in their annotation and ontology creation work depends heavily on their performance. In any case, it is of immense importance to process hidden information from literature to make it integrable with data from structured data sources by means of data integration. The more information can be made available from literature and the more new hypotheses can be created by modern bioinformatics methods on top of it.

Key Points

- Literature is a large integrated, but unstructured database. Manual curation and text mining are two independent approaches to access this database. Manual curation is accurate, but does not scale. Text mining scales, but is not accurate.
- Text mining can support the construction of ontologies by suggesting terms and inferring taxonomies. It cannot automate the process, as many terms do not appear literally in text.
- Text mining can identify gene names in text with success rates of over 80% across many species, but extracting relationships such as protein interactions has success rates of less than 50%.
- Biomedical search engines support assisted curation but lack requirements of large scale curation efforts.
- Manual curation of literature is necessary for high-quality annotation but can be supported by automated information retrieval techniques.
- Authors could annotate their papers on submission according to a standardised and guided curation methodology.

Acknowledgement

We kindly acknowledge funding of the EU project Sealife and Atif Iqbal for discussions on definition generation.

References

1. Yu AC. Methods in biomedical ontology. *J Biomed Inform* 2006;**39**:252–66.
2. Roberts P. Mining literature for systems biology. *Brief Bioinform* 2006;**7**:399.
3. Ashburner M, Ball C, Blake J, *et al*. Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nat genet*. 2000;**25**:25–9.
4. Smith B, Ashburner M, Rosse C, *et al*. The OBO foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol* 2007;**25**:1251–5.
5. Camon E, Magrane M, Barrell D, *et al*. The Gene Ontology Annotation (GOA) project: implementation of GO in SWISS-PROT, TrEMBL, and InterPro. *Genome Res* 2003; **13**:662–72.
6. Bult C, Eppig J, Kadin J, *et al*. The mouse genome database (MGD): mouse biology and model systems. *Nucleic Acids Res* 2007;**36**(Database issue):D724–28.
7. Crosby M, Goodman J, Strelets V, *et al*. FlyBase: genomes by the dozen. *Nucleic Acids Res* 2007;**35**:D486–91.
8. Chen N, Harris T, Antoshechkin I, *et al*. WormBase: a comprehensive data resource for Caenorhabditis biology and genomics. *Nucleic Acids Res* 2005;**33**:D383–9.

9. Ouyang S, Zhu W, Hamilton J, *et al*. The TIGR rice genome annotation resource: improvements and new features. *Nucleic Acids Res* 2007;**35**:D883–7.
10. Berardini T, Mundodi S, Reiser L, *et al*. Functional annotation of the arabidopsis genome using controlled vocabularies. *Plant Physiol* 2004;**135**:745–55.
11. Karamanis N, Lewin I, Seal R, *et al*. Integrating natural language processing with FlyBase curation. *Pac Symp Biocomput* 2007;245–56.
12. Masseroli M, Pinciroli F. Using Gene Ontology and genomic controlled vocabularies to analyze high-throughput gene lists: three tool comparison. *Comput Biol Med* 2006;**36**:731–47.
13. Yang D, Li Y, Xiao H, *et al*. Gaining confidence in biological interpretation of the microarray data: the functional consistence of the significant GO categories. *Bioinformatics* 2008;**24**:265–71.
14. Baumgartner JWA, Cohen KB, Fox LM, *et al*. Manual curation is not sufficient for annotation of genomic databases. *Bioinformatics* 2007;**23**:41–8.
15. Park J, Kim T-E, Park J. Monitoring the evolutionary aspect of the gene ontology to enhance predictability and usability. *BMC Bioinformatics* 2008;**9**:S7.
16. Camon EB, Barrell DG, Dimmer EC, *et al*. An evaluation of GO annotation retrieval for BioCreAtIvE and GOA. *BMC Bioinformatics* 2005;**6** (Suppl. 1):S17.
17. Leitner F, Krallinger M, Rodriguez-Penagos C. Introducing meta-services for biomedical information extraction. *Genome Biol* 2008;**9**:S6.
18. Cimiano P. *Ontology Learning and Population from Text: Algorithms, Evaluation and Applications*. New York, USA: Springer, Science+Business Media, LLC, 2006.
19. Zhou L. Ontology learning: state of the art and open issues. *Inf Technol Manag* 2007;**8**:241–52.
20. Ohta T, Tateisi Y, Kim J-D. The GENIA corpus: an annotated research abstract corpus in molecular biology domain. In: *Proceedings of the second international conference on Human Language Technology Research*. San Diego, California: Morgan Kaufmann Publishers Inc., 2002;82–6.
21. Krauthammer M, Nenadic G. Term identification in the biomedical literature. *J Biomed Inform* 2004;**37**:512–26.
22. Nenadic G, Spasic I, Ananiadou S. Mining biomedical abstracts: what is in a term? In: *Proceedings of International Joint Conference on NLP* 2004;247–54.
23. Alexopoulou D, Wächter T, Pickersgill L, *et al*. Terminologies for text-mining; an experiment in the lipoprotein metabolism domain. *BMC Bioinformatics* 2008; **9**(Suppl 4):S2.
24. Frantzi KT, Ananiadou S, Tsujii J. The C-value/NC-value method of automatic recognition for multi-word terms. In: *ECDL'98: Proceedings of the Second European Conference on Research and Advanced Technology for Digital Libraries*. London, UK: Springer-Verlag, 1998;585–604.
25. Hirschman L, Morgan AA, Yeh AS. Rutabaga by any other name: extracting biological names. *J Biomed Informatics* 2002; **35**:247–59.
26. Tanabe L, Wilbur WJ. Tagging gene and protein names in full text articles. In: *Proceedings of the ACL-02 Workshop on Natural Language Processing in the Biomedical Domain*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, 2002;9–13.
27. Tuason O, Chen L, Liu H, *et al*. Biological nomenclatures: a source of lexical knowledge and ambiguity. *Pac Symp Biocomput* 2004;238–49.
28. Lee J-B, Kim J-J, Park JC. Automatic extension of gene ontology with flexible identification of candidate terms. *Bioinformatics* 2006;**22**:665–70.
29. Fellbaum C. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. Cambridge, USA: The MIT Press, 1998.
30. Mccrae J, Collier N. Synonym set extraction from the biomedical literature by lexical pattern discovery. *BMC Bioinformatics* 2008;**9**:159+.
31. Shimohata M, Sumita E. Acquiring synonyms from monolingual comparable texts. In: *IJCNLP* 2005;233–44.
32. Wren JD, Chang JT, Pustejovsky J, *et al*. Biomedical term mapping databases. *Nucleic Acids Res* 2005;**33**:D289–93.
33. Gaudan S, Kirsch H, Rebholz-Schuhmann D. Resolving abbreviations to their senses in medline. *Bioinformatics* 2005; **21**:3658–64.
34. Zhou W, Torvik VI, Smalheiser NR. ADAM: another database of abbreviations in MEDLINE. *Bioinformatics* 2006; **22**:2813–18.
35. Okazaki N, Ananiadou S. Building an abbreviation dictionary using a term recognition approach. *Bioinformatics* 2006;**22**:3089–95.
36. Klavans J, Muresan S. Evaluation of the DEFINDER system for fully automatic glossary construction. *Proc AMIA Symp* 2001;324–8.
37. Hearst MA. Automatic acquisition of Hyponyms from large text corpora. In: *Proceedings of the Fourteenth International Conference on Computational Linguistic*. France: Nantes, 1992.
38. Heymann P, Garcia-Molina H. *Collaborative Creation of Communal Hierarchical Taxonomies in Social Tagging Systems*. Stanford University, 2006.
39. Ganter B, Stumme G, Wille R. *Formal Concept Analysis: Foundations and Applications*. Berlin Heidelberg, Germany: Springer-Verlag, 2005.
40. Ryu P-M, Choi K-S. Taxonomy learning using term specificity and similarity. In: *Proceedings of the 2nd Workshop on Ontology Learning and Population: Bridging the Gap between Text and Knowledge*. Sidney, Australia: Association for Computational Linguistics, 2006;41–8.
41. Snow R, Jurafsky D, Ng AY. Semantic taxonomy induction from heterogeneous evidence. In: *ACL'06: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL*. Sidney, Australia: Association for Computational Linguistics, 2006;801–8.
42. Cimiano P, Völker J. Text2Onto – a framework for ontology learning and data-driven change discovery. In: *Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems(NLDB' 2005)*. Spain: Alicante, 2005.
43. Wilbur J, Smith L, Tanabe L. BioCreative 2. Gene mention task. In: *Proceedings of the Second BioCreative Challenge Evaluation Workshop*. Madrid, Spain: Fundacion CNIO Carlos III, 2007;7–16.
44. Fundel K, Zimmer R. Gene and protein nomenclature in public databases. *BMC Bioinformatics* 2006;**7**:372.
45. Hakenberg J, Plake C, Leaman R, *et al*. Inter-species normalization of gene mentions with GNAT. *Bioinformatics* 2008;**24**(16):i126–32.

46. Tamames J, Valencia A. The success (or not) of HUGO nomenclature. *Genome Biol* 2006;**7**:402.
47. King M, Lusk C, Blobel G. Karyopherin-mediated import of integral inner nuclear membrane proteins. *Nature* 2006; **442**:1003–7.
48. Baumgartner WA, Lu Z, Johnson HL, *et al*. An integrated approach to concept recognition in biomedical text. In: *Proceedings of the Second BioCreative Challenge Evaluation Workshop*. Madrid, Spain: Fundacion CNIO Carlos III, 2007;257–271.
49. 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.
50. Morgan A, Hirschman L. Overview of biocreative II gene normalization. In: *Proceedings of the Second BioCreative Challenge Evaluation Workshop*. Madrid, Spain: Fundacion CNIO Carlos III, 2007;17–27.
51. Huang H-S, Lin Y-S, Lin K-T, *et al*. High-recall gene mention recognition by unification of multiple backward parsing models. In: *Proceedings of the Second BioCreative Challenge Evaluation Workshop*. Madrid, Spain: Fundacion CNIO Carlos III, 2007;109–111.
52. Kuo C-J, Chang Y-M, Huang H-S, *et al*. Rich feature set, unification of bidirectional parsing and dictionary filtering for high F-score gene mention tagging. In: *Proceedings of the Second BioCreative Challenge Evaluation Workshop*. Madrid, Spain: Fundacion CNIO Carlos III, 2007;105–107.
53. Ando RK. BioCreative II gene mention tagging system at IBM Watson. In: *Proceedings of the Second BioCreative Challenge Evaluation Workshop*. Madrid, Spain: Fundacion CNIO Carlos III, 2007;101–103.
54. Cakmak A, Ozsoyoglu G. Annotating genes using textual patterns. *Pac Symp Biocomput* 2007;221–32.
55. Doms A, Schroeder M. GoPubMed: exploring PubMed with the gene ontology. *Nucleic Acids Res* 2005;**33**:W783–6.
56. Couto FM, Silva MJ, Coutinho PM. Finding genomic ontology terms in text using evidence content. *BMC Bioinformatics* 2005;**6** (Suppl 1):S21.
57. Ruch P. Automatic assignment of biomedical categories: toward a generic approach. *Bioinformatics* 2005;**22**: 658–664.
58. Jaeger S, Gaudan S, Leser U, *et al*. Integrating protein–protein interactions and text mining for protein function prediction. *BMC Bioinformatics* 2008;**9** (Suppl 8):S2.
59. Shatkay H, Höglund A, Brady S, *et al*. SherLoc: High-accuracy prediction of protein subcellular localization by integrating text and protein sequence data. *Bioinformatics* 2007;**23**(11):1410–17.
60. Andreopoulos B, Alexopoulou D, Schroeder M. Word sense disambiguation in biomedical ontologies with term co-occurrence analysis and document clustering. *Int J Data Min Bioinform* 2008;**2**(3):193–215.
61. Eaton A. HubMed: a web-based biomedical literature search interface. *Nucleic Acids Res* 2006;**34**:W745–7.
62. Hoffmann R, Valencia A. A gene network for navigating the literature. *Nat genet* 2004;**36**:664.
63. Rebholz-Schuhmann D, Kirsch H, Arregui M, *et al*. EBIMed – text crunching to gather facts for proteins from Medline. *Bioinformatics* 2007;**23**:e237–44.
64. Plake C, Schiemann T, Pankalla M, *et al*. AliBaba: PubMed as a graph. *Bioinformatics* 2006;**22**:2444–5.
65. Müller H-M, Kenny EE, Sternberg PW. Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biol* 2004;**2**:e309.
66. Chen H, Sharp B. Content-rich biological network constructed by mining PubMed abstracts. *BMC Bioinformatics* 2004;**5**:147.
67. Divoli A, Attwood T. BioIE: extracting informative sentences from the biomedical literature. *Bioinformatics* 2005;**21**:2138–9.
68. Siadaty M, Shu J, Knaus W. Relemed: sentence-level search engine with relevance score for the MEDLINE database of biomedical articles. *BMC Med Inform Decis Mak* 2007;**7**:1.
69. Perez-Iratxeta C, Pérez A, Bork P, *et al*. Update on XplorMed: A web server for exploring scientific literature. *Nucleic Acids Res* 2003;**31**:3866–8.
70. Couto F, Silva M, Lee V, *et al*. GOAnnotator: linking protein GO annotations to evidence text. *J Biomed Discov Collab* 2006;**1**:19.
71. Plikus M, Zhang Z, Chuong C. PubFocus: semantic MEDLINE/PubMed citations analytics through integration of controlled biomedical dictionaries and ranking algorithm. *BMC Bioinformatics* 2006;**7**:424.
72. Alex B, Grover C, Haddow B, *et al*. Assisted curation: does text mining really help? *Pac Symp Biocomput* 2008;556–67.
73. Hoffmann R, Valencia A. Implementing the iHOP concept for navigation of biomedical literature. *Bioinformatics* 2005; **21** (Suppl 2):ii252–ii258.
74. Donaldson I, Martin J, Bruijn B, *et al*. PreBIND and Textomy – mining the biomedical literature for protein–protein interactions using a support vector machine. *BMC Bioinformatics* 2003;**4**:11.
75. Bader GD, Donaldson I, Wolting C, *et al*. BIND – The biomolecular interaction network database. *Nucleic Acids Res* 2001;**29**:242–245.
76. Caporaso JG, Deshpande N, Fink J.L, *et al*. Intrinsic evaluation of text mining tools may not predict performance on realistic tasks. *Pac Symp Biocomput* 2008;640–51.
77. Salzberg S. Genome re-annotation: a wiki solution? *Genome Biol* 2007;**8**:102.
78. Rodriguez-Esteban R, Iossifov I, Rzhetsky A. Imitating manual curation of text-mined facts in biomedicine. *PLoS Computat Biol* 2006;**2**:e118.
79. Burkhardt K, Schneider B, Ory J. A biocurator perspective: annotation at the research collaboratory for structural bioinformatics protein data bank. *PLoS Comput Biol* 2006; **2**:e99.
80. Gerstein M, Seringhaus M, Fields S. Structured digital abstract makes text mining easy. *Nature* 2007;**447**:142.
81. Ceol A, Chatr-Aryamontri A, Licata L, *et al*. Linking entries in protein interaction database to structured text: The FEBS letters experiment. *FEBS Lett* 2008;**582**:1171–1177.
82. Chatr-aryamontri A, Ceol A, Palazzi L, *et al*. MINT: the Molecular INTeraction database. *Nucleic Acids Res* 2007;**35**: D572–4.
83. Leitner F, Valencia A. A text-mining perspective on the requirements for electronically annotated abstracts. *FEBS Lett* 2008;**582**:1178–1181.
84. Kuhn T, Royer L, Fuchs NE, *et al*. Improving text mining with controlled natural language: a case study for protein interactions. In: *Proc DILS*. Berlin Heidelberg, Germany: Springer-Verlag, 2006;66–81.