BIOINFORMATICS ORIGINAL PAPER
Vol. 21 no. 9 2005, pages 2059–2066
doi:10.1093/bioinformatics/bti284

Data and text mining

Concept-based annotation of enzyme classes

Oliver Hofmann* and Dietmar Schomburg

Department of Biochemistry, University of Cologne, Zuelpicher Strasse 47, 50674 Cologne, Germany

Received on October 25, 2004; revised on January 17, 2005; accepted on January 18, 2005
Advance Access publication January 20, 2005

ABSTRACT

Motivation: Given the explosive growth of biomedical data as well as the literature describing results and findings, it is getting increasingly difficult to keep up to date with new information. Keeping databases synchronized with current knowledge is a time-consuming and expensive task—one which can be alleviated by automatically gathering findings from the literature using linguistic approaches. We describe a method to automatically annotate enzyme classes with disease-related information extracted from the biomedical literature for inclusion in such a database.

Results: Enzyme names for the 3901 enzyme classes in the BRENDA database, a repository for quantitative and qualitative enzyme information, were identified in more than 100 000 abstracts retrieved from the PubMed literature database. Phrases in the abstracts were assigned to concepts from the Unified Medical Language System (UMLS) utilizing the MetaMap program, allowing for the identification of disease-related concepts by their semantic fields in the UMLS ontology. Assignments between enzyme classes and diseases were created based on their co-occurrence within a single sentence. False positives could be removed by a variety of filters including minimum number of co-occurrences, removal of sentences containing a negation and the classification of sentences based on their semantic fields by a Support Vector Machine. Verification of the assignments with a manually annotated set of 1500 sentences yielded favorable results of 92% precision at 50% recall, sufficient for inclusion in a high-quality database.

Availability: Source code is available from the author upon request.

Contact: o.hofmann@smail.uni-koeln.de

Supplementary information: ftp.uni-koeln.de/institute/biochemie/pub/brenda/info/diseaseSupp.pdf

INTRODUCTION

During the past years biology has gradually changed from a hypothesis-driven science to a data-driven one. With high-throughput methods in fields like genome analysis, proteomics and system biology arises the need to make sense of the ever increasing amount of raw data.

Automatic evaluation and annotation of this data usually requires the utilization of several of the literally hundred biomedical databases currently accessible online via the World Wide Web (Galperin, 2004), covering genomic data (Benson *et al.*, 2003), protein sequences (Camon *et al.*, 2004) and structures (Berman *et al.*, 2000), metabolic pathways (Kanehisa *et al.*, 2002) or disease-related information (McKusick, 2000). Frequently, essential information needs to be manually retrieved from the literature to facilitate the analysis since few of the available data repositories are current, exhaustive and machine-readable—all of which are required to facilitate the automatic evaluation of experimental results.

Several problems contribute to this situation, chief among them being the cost of manually extracting information from scientific publications, as exemplified by the BRENDA database—a manually curated repository for functional enzyme information. The current version contains qualitative and quantitative data for about 4200 enzyme classes representing >80 000 different enzyme molecules (Schomburg *et al.*, 2004). The data include, among other fields, information on nomenclature, catalyzed reaction and specificity, occurrence and application, all stored in a relational database system with free access to the academic community (Access to the BRENDA database at http://www.brenda-uni-koeln.de). All parameters were collected from >50 000 articles by experts. Given this significant investment of resources, the rapid growth of new knowledge leads to a widening gap between available, manually curated data and already published current knowledge.

This growth is reflected by the PubMed literature database. Containing more than 12 million abstracts of biomedical publications and growing at a rate of ∼40 000 entries per month it exemplifies the need to enhance the manual annotation process by other means (de Bruijn and Martin, 2002; Wren and Garner, 2004).

Biomedical information extraction

Automatic annotation of biological entities by extracting relevant information from the biomedical literature offers an opportunity to keep up with the pace of published knowledge. Linguistic methods have seen increasing popularity in the field of bioinformatics during recent years, being used for a variety of tasks ranging from information retrieval (Yang, 1999) to sequence annotation based on the description of related sequences (Dobrokhotov *et al.*, 2003) to enhancing existing algorithms like PSI-BLAST (Chang *et al.*, 2001).

The task of automatic extraction and identification of relationships or associations between entities in the biomedical literature has received a particularly strong focus. Research ranges from identifying gene–gene associations (Tao and Leibel, 2002) to protein–protein interactions (Donaldson *et al.*, 2003) to more complex tasks like recognizing inhibitors (Pustejovsky *et al.*, 2002) and building metabolic networks from enzyme information (Humphreys *et al.*, 2000) (for a recent review of computational linguistics in the field of biology see Hirschman *et al.*, 2002 and Yeh *et al.*, 2003). Although the specificity and sensitivity achieved by automatic methods usually do not rival the precision of human experts it is more than sufficient to complement experimental data, which is often associated with similar margins of error.

*To whom correspondence should be addressed.

Unfortunately, there is a lack of freely available, modular systems which can be tailored effortlessly to extract information from a variety of domains while covering all the different aspects of an automatic annotation system: the creation of a text corpus, the identification of biological entities and their potential association and finally the presentation of the extracted data.

Here, we describe a system which allows for the annotation of enzyme classes with different properties according to the need of the user. In our case—the addition of the automatically acquired annotation to a curated database—an emphasis was placed on minimizing the number of false positive annotations.

The benefits of concepts

The presented prototype system is based on the co-occurrence of enzyme names and concepts within abstracts retrieved from the PubMed literature database—an approach that has been used successfully to annotate gene (Stapley and Benoit, 2000; Jenssen *et al.*, 2001) and protein interactions (Blaschke *et al.*, 1999; Ono *et al.*, 2001; Marcotte *et al.*, 2001). It identifies enzyme names extracted from the BRENDA database using a dictionary-based approach and analyzes their co-occurrence concepts describing human diseases. A concept within this context is an abstract entity that describes the meaning of an event, a relationship or a class of entities. Using concepts entails several benefits:

- Differing term representations of the same concept (e. g. 'high blood pressure' and 'hypertension') can be collapsed to the same idea.
- Compound terms can be maintained, i. e. the concept 'high blood pressure' differs from the three concepts 'high', 'blood' and 'pressure' which might appear independently of each other within a single document. A majority of the biomedical vocabulary has been described to consist of such compound terminology (Bodenreider *et al.*, 2002).
- The mapping of synonyms and variants to one concept results in a controlled vocabulary and simplifies a comparison and evaluation of the annotation.

Despite these benefits the majority of information extraction systems in the biomedical field analyze words and combinations of words only (de Bruijn and Martin, 2002), mostly due to the lack of a concise, exhaustive and freely available ontology and the difficult task of mapping phrases to the correct concepts. Given the recent integration of the Gene Ontology (GO) into the Unified Medical Language System (UMLS) the first challenge is no longer an issue. The 13th edition of this system contains more than 1.5 million terms gathered from 60 dictionaries and grouped to about 775 000 concepts (Bodenreider, 2004). In addition, assigned semantic fields describe the properties of each concept, allowing for the distinction of concepts like 'rat' with the semantic field 'mammal' and 'polymerase chain reaction', marked as a 'laboratory procedure'. Semantic filters based on these fields facilitate the annotation of enzyme classes with different properties, in agreement with the underlying idea of being able to easily enhance the annotation process by including different knowledge domains.

Matching text segments to concepts can be achieved either through a simple string comparison or preferably by the use of a specialized program like MetaMap (Aronson, 2001) which breaks the text into phrases, assigns part-of-speech tags (the markup of words with their corresponding parts of speech) and generates variants before evaluating candidate concepts by a variety of different criteria. It has been applied successfully to a number of tasks ranging from information retrieval (Aronson and Rindflesch, 1997) to text mining (Weeber *et al.*, 2001).

Table 1. Date sources used in this publication

Database	Available at
BRENDA	http://www.brenda.uni-koeln.de
UMLS	http://www.nlm.nih.gov/research/umls/
PubMed	http://www.pubmed.gov
MESH	http://www.nlm.nih.gov/mesh/MBrowser.html

Annotation by concept-based co-occurrence

Creating relationships between objects based on their co-occurrence within different text segments has been used successfully in the past (Weeber *et al.*, 2001; Wren *et al.*, 2004). While more complex systems using statistical methods to detect patterns or those implementing rule-based algorithms to extract the relevant keywords and their dependencies (Temkin and Gilder, 2003) can identify associations with high precision, they are often difficult to maintain and adapt to different tasks.

In accordance with the aim of implementing an easily maintainable and adoptable system we focused on co-occurring enzyme classes and relevant concepts while trying to retain a high precision by concentrating on sentences instead of larger text segments (Ding *et al.*, 2002). Additionally, extracting links from sentences simplifies the evaluation task for the user in comparison to full-length abstracts since the proposed association is immediately available for verification. This review process is assisted by presenting the network of enzymes and diseases as an interactively explorable graph (see Supplementary material for a sample visualization).

SYSTEMS AND METHODS

The code was written in Python 2.2 and tested on a Linux (Kernel 2.4.20), Solaris (5.8) and Windows 2000 system. All data were stored in a MySQL database (version 3.23.45) and accessed using the mySQLdb extension for Python. Table 1 lists the data sources used to create the text corpus and enzyme dictionary. Entries processed in this publication were gathered during January–March 2002. Access to the UMLS requires a license freely available to academic users. Concepts were assigned using the MetaMap software package (version 2.2a), which is covered by the same license as that of the UMLS. The resulting network was inspected visually using TouchGraph (TouchGraph, 2002).

All system parts and filters were implemented in an object-oriented, modular way, so that individual components such as the enzyme name recognition or negation detection could be replaced with more sophisticated algorithms in a straightforward manner if so required.

ALGORITHM AND IMPLEMENTATION

Figure 1 outlines the system's workflow and can be divided into four basic parts: information retrieval and processing, entity recognition, linking the entities and finally visualizing the created network.

Abstracts were collected from the PubMed database utilizing the Medical Subject Headings (MESH). The MESH terms represent a hierarchically organized, controlled vocabulary used to manually

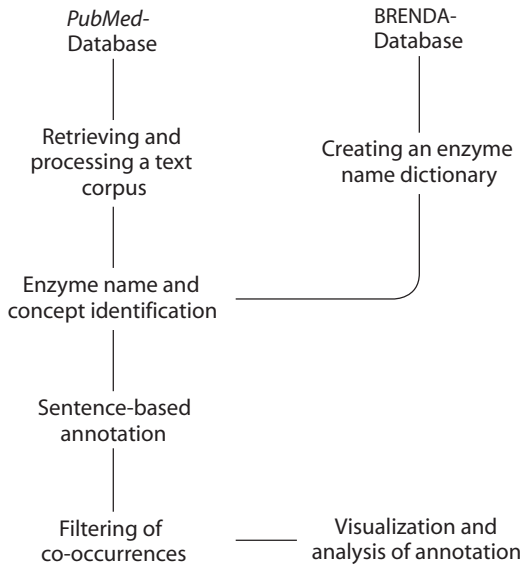

Fig. 1. Basic workflow for the information retrieval, co-occurrence analysis and evaluation.

index PubMed entries. Queries included all combinations of MESH terms from the disease-related branches C1–C21 of the MESH tree with recommended enzyme names and enzyme class identifiers stored in the BRENDA database.

Collected entries were required to have full-length English abstracts and limited to the MESH-keyword 'human'. Queries returning less than ten hits for a single enzyme class were expanded by including synonyms taken from the BRENDA-database. Individual queries with >1000 hits were successively limited to more recent publications until <1000 relevant documents were found. Entries were split into sentences; special chars, hyphens and brackets were replaced by whitespace. Any alphanumeric sequence bracketed by whitespace was defined as a token and compared to the enzyme dictionary.

Detecting enzymes and diseases

Enzyme names in the BRENDA database (recommended names, synonyms, EC numbers and systematic names) were processed in the same way as the collected documents and expanded by adding more names: all numerals were augmented by their roman (or arabic) and spelled-out versions. After removing names also appearing in a common English dictionary and those with a length of fewer than four characters, entries in the dictionary were matched against the collected abstracts. In the case of overlapping matches the enzyme name with the largest coverage took precedence. Ambiguous names could be partially resolved by keeping track of unambiguous matches in the same abstract and discarding ambiguous names that could not be detected elsewhere.

Concept identification by MetaMap used default parameters except for a limitation to unique abbreviations and acronyms only (option -u) to improve precision (Table 2 shows a sample sentence and its concept representation). Additionally the strict model was used while preprocessing the UMLS metathesaurus (option -A), resulting in the removal of terms for which an identification was unlikely due to their internal structure (Aronson, 2002).

Table 2. Sample sentence and its representation after concept identification using MetaMap

Word	Concept	Semantic field	Score
Deficiency	Deficiency	Functional concept	1000
Ubiquinone	Ubiquinone	Biologically active substance, organic chemical	645
Cytochrome c reductase	Cytochrome c reductase	Amino acid, peptide, or protein, enzyme	923
Patient	Patient	Patient or disabled group	632
Mitochondrial myopathy	**Mitochondrial myopathy**	Disease or syndrome	827

Disease-related concepts are marked in bold.
'Deficiency in ubiquinone cytochrome c reductase in a patient with mitochondrial myopathy.'

Phrases, assigned concepts and positional information were stored in the database and the preferred concept term in the UMLS was used to represent each concept. The semantic field 'disease or syndrome' (UMLS concept identifier 'T047') marked disease-related concepts and could be used for studying co-occurring enzyme names and diseases. Removal of very generic concepts was implemented by a simple frequency analysis filtering out disease-related concepts occurring in more than 5% of the mapped sentences.

Guilty by association: co-occurrence based annotation

Annotation of enzyme classes with disease-related concept was based on the co-occurrence of both terms within single sentences. Sentences fulfilling those conditions were filtered by four criteria and the resulting change in precision and recall was studied:

(1) The confidence in the mapped disease-related concept: The MetaMap program scores matches of candidate concepts to phrases by their centrality, coverage, cohesion and the number of necessary variations [see Aronson (2001) for details]. A higher score reflects a better match and, therefore, higher confidence in the assignment.

(2) The optional detection of sentences containing a negation: according to Mutalik *et al*. (2001) six words ('no', 'denies/denied', 'not', 'none' and 'without') cover 93% of all negations in medical text. Removing sentences containing such a negation should remove most false positive annotations that would have been assigned due to negative findings.

(3) The semantic context of the detected enzyme names and diseases: during manual inspection of several hundred sentences it became obvious that they could be grouped into several categories, i. e. description of diseases, a patient's case or a laboratory method. Therefore, we classified sentences into disease-related and unrelated sentences based on the semantic fields associated with any concept identified within the sentence by MetaMap.

Similar semantic context analysis has been used successfully for word sense disambiguation based on the semantic properties of the surrounding text using statistical methods like a Support Vector Machine (SVM) (Cabezas *et al*., 2001). Being developed in the field of machine learning, an SVM

creates a binary classifier returning either class +1 or −1 for each feature vector. Classes are separated by a hyperplane—each feature vector's class being determined by the side of the space separated by the hyperplane it is located on. Training the SVM with an annotated dataset means finding the optimal hyperplane, i.e. the one with the maximum distance between itself and the feature vectors closest to the plane—the support vectors (Kazama *et al.*, 2002).

The SVM classifier was implemented using the ORANGE machine learning library (Demsar and Zupan, 2004) with a training and test set of 1000 manually annotated sentences, 400 of which described an association between an enzyme class and a disease. The feature vectors consisted of the semantic fields found in each sentence.

(4) The minimum number of detected enzyme and disease co-occurrences within sentences before accepting an association between both entities. This filter was applied after the application of the previous three filters.

Links created in this manner were stored and the network of enzyme classes and disease concepts visualized using the TouchGraph library, each object being represented by a node with edges connecting associated enzymes and diseases. Enzyme nodes point to the original entries in the BRENDA database. Disease nodes present definitions and summaries automatically generated from the sentences containing the disease.

RESULTS

The enzyme dictionary collected from the BRENDA database contained 3901 recommended names and 17 530 synonyms, amounting to more than six names per enzyme on average, including the enzyme class number. Some enzyme classes like the protein kinases (EC 2.7.1.37) list well above 60 different names, while the diverse class of Type II site-specific deoxyribonucleases (EC 3.1.21.4) includes 180 different enzymes.

Based on this enzyme name dictionary a text corpus of 105 897 PubMed documents was retrieved, containing ∼200 000 identified enzyme names and synonyms. While the majority of documents only mentioned 1–3 enzymes, >10 000 abstracts listed from 4 to as many as 25 different enzyme classes, reconfirming the need to extract links from smaller text segments. Manual annotation of 120 enzyme names in a sample of 100 abstracts revealed a recall of 97%, missing enzyme names mostly being a result of an incomplete enzyme name dictionary and errors during tokenization. Given that the same enzyme names were part of the original queries to retrieve documents for the text corpus those numbers should not be surprising.

Concept identification reduced the dictionary from 300 000 different tokens to a mere 50 000 concepts, while the average abstract was reduced from >200 tokens to ∼40 concepts. A manually reviewed sample of 100 sentences showed 67% of the tokens were assigned to concepts with a precision of >90%, in agreement with other studies (Pratt and Yetisgen-Yildiz, 2003). Phrases without a concept representation were either missing from the UMLS or consisted of compound phrases and artifacts like '12-year-old' or '3alpha'. Errors mostly occurred due to incorrectly mapped acronyms and the inability to distinguish between homonyms.

To evaluate precision and recall an additional 1500 sentences with at least one identified enzyme class were manually annotated. A total of 273 sentences was found to be describing 430 associations between enzyme classes and diseases, of which 384 were unique; 80 sentences mentioned both, one or more, enzyme classes and disease-related concepts without any discernible interrelation between both. With the task being the automatic enhancement of databases it was deemed more important to correctly extract the links themselves, as compared to identifying each true co-occurrence. A true positive link, therefore, represents a correctly identified association between an enzyme class and a disease-related concept, regardless of where it was identified. Conversely, a true negative is a co-occurrence of both entities with no discernible relationship that was correctly filtered out during the extraction process. Using these definitions 84.8% of all links could be identified by evaluating the co-occurring concepts and enzyme names, while precision was at 82.1% (Table 3).

Table 3. Precision and recall in percent for a manually annotated set of 1500 sentences

Filter	Precision	Recall
Basic co-occurrence	82.1	84.8
MetaMap assignment score ⩾800	85.2	70.0
negated sentences removed	87.5	65.5
more than one co-occurrence	90.7	52.9
filtering semantic context	94.8	38.9

Filters are applied successively as described in the algorithm and implementation section. A true positive is defined as the correct identification of an association between an enzyme class and a disease-related concept, independent of the number of correctly identified instances of that association.

More stringent criteria

Requiring a higher MetaMap confidence score before accepting a disease-related concept—and thus its co-occurrence with an enzyme name—gradually increases precision up to 3% at the cost of ∼15% recall (Fig. 2a). Raising the minimum confidence score beyond 800 removes more correct than incorrect concept assignments.

Changing the minimum number of co-occurrences to more than one increased the precision by another 4% for an additional decrease of recall by 14% (Fig. 2b). In both cases precision benefitted from the removal of sentences containing a negation prior to the evaluation of co-occurrences—precision increased by ∼3% for the same loss in recall.

Finally, at up to 50% recall the semantic context filter improves precision by about 4% independently of the minimum confidence score and number of co-occurrences. Tables outlining all precision/recall values at the different threshold levels can be found in the supplementary material.

Enzymes and their diseases

Using stringent parameters (MetaMap score of 800, no negated sentences, using the SVM-filter and requiring at least two co-occurrences) a total of 1409 disease-related concepts could be assigned to 524 enzyme classes and added to the BRENDA database. On average, each enzyme class was associated with 10.3 different diseases. The hydrolases exhibited significantly more links to diseases—averaging more than 13 links per enzyme class—than the oxidoreductases and transferases, despite similar numbers of enzyme classes. In contrast, the ligases with the least number of enzyme classes also exhibited the lowest connectivity.

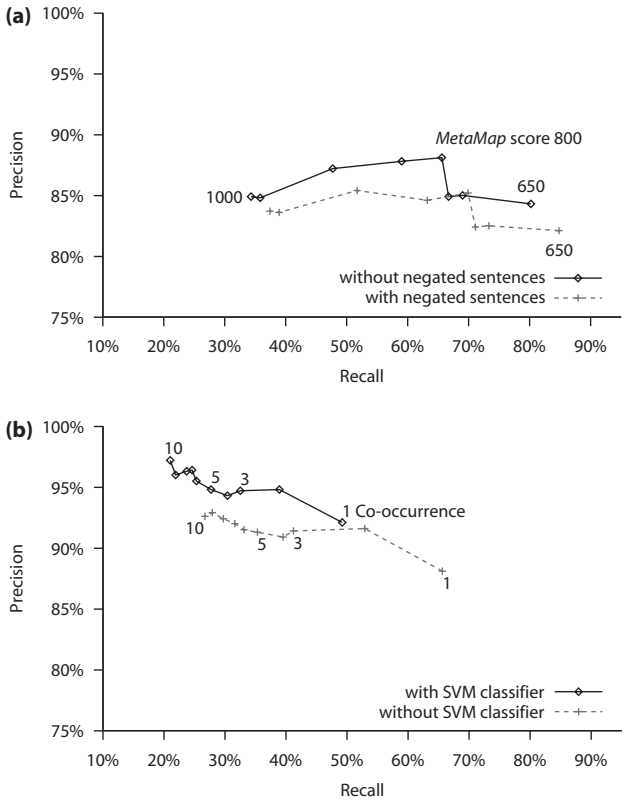

Fig. 2. Recall and precision using different parameters. (**a**) shows the MetaMap confidence score in combination with the negation detection (no minimum number of co-occurrences required); (**b**) reflects the influence of the SVM classifier along with the minimum number of co-occurrences (no negated sentences allowed, MetaMap confidence score $\geqslant$800).

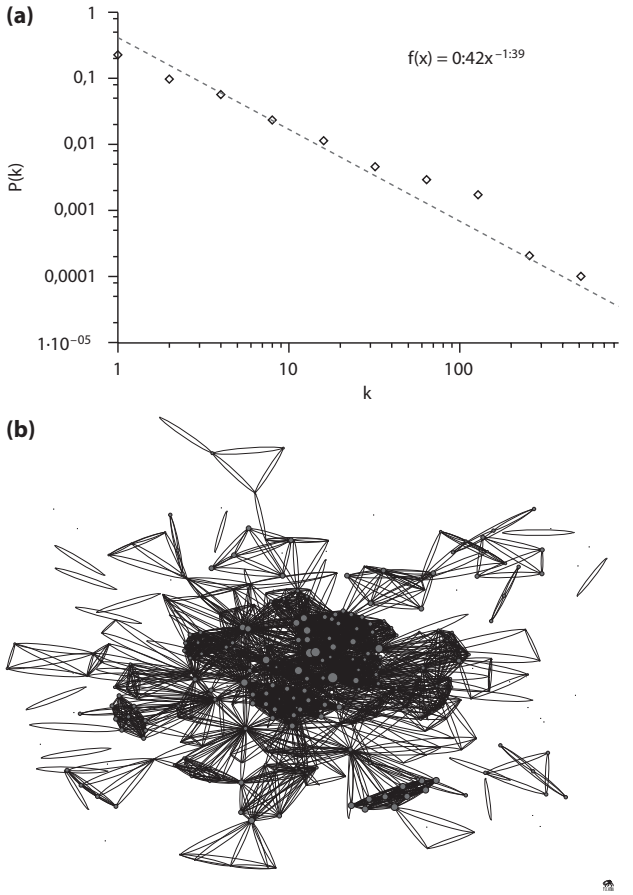

Fig. 3. (**a**) Connectivity distribution for the disease graph showing the probability $P(k)$ and the logarithmically binned node's degree k following a power law. (**b**) Overview of the disease graph. Node size represents the node's clustering coefficient. The graph is based on only 10% of the text corpus for easier visualization.

The bipartite network of diseases and enzymes can be represented as two distinct networks: an enzyme graph with edges connecting enzymes that are associated with the same disease and a disease graph with nodes being connected due to shared enzymes. A comparison of those graphs with random networks of equal size and average connectivity revealed a scale-free, small-world structure for both networks (graph statistics, comparisons to random networks of equal size and connectivity distributions are available in the Supplementary information). The enzyme graph in particular is three times more clustered than a random graph of equal size while retaining a similar diameter, typical for the easily traversed small-world network architecture.

Node connectivity in both random and small-world networks follows a poisson distribution, resulting in homogenous networks and generally lacking nodes with a high degree of connectivity. In the analyzed enzyme and disease graphs, however, the connectivity distribution is best described by a power law, classifying the graphs as scale-free networks (For the connectivity distribution and an overview of the disease graph, see Fig. 3).

DISCUSSION

A comparison of the results with published approaches taking similar advantage of the concepts in the UMLS proves difficult. They do not state the achieved precision and recall values (Rindflesch *et al.*, 2000b), do use different evaluation standards (Feldman *et al.*, 2003) nor tackle a different information extraction problem. Rindflesch *et al.* (2000a) report a precision of 73% for the interaction of biological molecules at a recall of 53%, Wren and Garner (2004) assume an 83% probability of an interrelation between two genes co-occurring within one sentence. The later assumption is similar to the precision attained in this work using no additional filters.

The four different filters all improve the precision of the extracted annotation at the cost of lowering recall while covering different aspects of the extracted links. Raising the minimum confidence score for the concept assignment filters associations based on partial or badly matched concepts (Table 4), but is bound to fail in cases where concepts in the UMLS are labelled with an inappropriate semantic field (example 3). While filtering concepts occurring too frequently did help to remove general concepts (example 4), the strategy did not help in every case.

Sentences that are being removed during the SVM classification step often describe laboratory methods or experimental strategies and use different concepts—and thus semantic fields—in comparison to disease-related sentences. In contrast to filtering for the minimum number of co-occurring enzymes and diseases, this approach allows

Table 4. Sample sentences and the confidence score of disease-related concepts (emphasized in *italics*)

	Sentence	Concept and score
1	Immunohistochemical staining of *lung tissue* with anti human neutrophil elastase…	Lung diseases (694)
2	Three bk virus mutants forming clear large *plaques* like those…	Dental plaques (670)
3	The patients had normal or *increased activities* of…	Increased activities (900)
4	…in inactive *tb* patients reveals the quiescent stage of the *disease*	Disease (1000), Tuberculosis (660)

for the removal of systematic errors during the annotation process like the following example:

> 'three bk virus mutants . . . capable of transforming *rat* cells were derived from the recombinant virus carrying the *hind iii* c segment'.

The term RAT is being recognized as an acronym for recurrent acute tonsilitis, a bacterial infection. As a result, all experiments dealing with rats link this disease to a multitude of different enzymes, most of which occur more frequently than a still reasonable co-occurrence filter could remove. The latter, however, is helpful in eliminating rare, incorrect co-occurrences that frequently happen due to the wrong usage of enzyme classes in publications, lineups and comparisons of diseases or undetected negations.

While negated sentences occur infrequently—only 3% of the manually annotated sentences used a negation—their removal still results in a slight improvement in precision, despite sentences similar to the following:

> 'There was no *alpha-d-mannosidase* activity in the hair roots of the patient with *mannosidosis*.'

In this particular case, the lack of enzyme activity is the actual cause for a metabolic disease and the removal leads to a lower recall score. As the number of required co-occurrences increases the benefit of this filter decreases, since bad annotations created from negated sentences tend to be mentioned only rarely within the text corpus. At lower co-occurrence thresholds the negation filter remains useful, nevertheless.

Limitations and alternatives

Using enzyme names collected from a manually curated data source like the BRENDA database results in a precise name recognition, but recall suffers due to the rapid growth and constant changes in biological terminology. Keeping up with current literature and name changes is crucial, as names having not yet been annotated or assigned to an enzyme class cannot be identified and, therefore, won't contribute to automatically extracted links. Unfortunately, there is no reliable way of identifying enzyme names in the studied documents, as well as correctly assigning them to the appropriate enzyme class. Recommended names and synonyms of enzyme classes neither follow an exploitable rigid nomenclature nor are homogenous enough to utilize edit distance comparisons. For the time being a dictionary-based enzyme name recognition, enhanced by various methods to 'boost' the number of recognized name variations, seems to be the best approach when focussing on high precision results.

While the majority of missed annotations are caused by errors during the concept assignment, missing concepts in the UMLS or incorrect filtering, wrong annotations have a number of different causes. Sentences emphasizing one disease distinguishing features by comparing it with other diseases frequently result in an error. Text fragments describing laboratory methods while dealing with patient material or clinical studies are difficult to distinguish from true positive co-occurrences based on their semantic properties alone. Finally, lists of enzymes being tested for their causal relationship to a given disease are another common source of errors.

Despite the achieved precision the high connectivity of the enzyme-disease network poses a problem. Major factors contributing to this connectivity are the assignment of related or very general concepts to the same enzyme class. In principle, the semantic network of the UMLS could be used to merge such concepts and filter basic concepts. However, due to the UMLS being constructed from a large number of different lexica and ontologies with different levels of detail, distance information is not a reliable criterion for filtering. Additionally, several inconsistencies in the semantic network (Pisanelli *et al*., 1998; Liu *et al*., 2002) complicate the merging of similar concepts. Joining concepts based on the similarity of their assigned preferred term seems counterintuitive (Pratt and Yetisgen-Yildiz, 2003), but seems to be the best approach while the semantic network of the UMLS keeps improving with each released version.

A hierarchy-based filter utilizing the semantic network of the UMLS would be preferable, but is difficult to implement due to the heterogeneity of the dictionaries used to construct the ontology (Liu *et al*., 2002).

A small world of diseases and enzymes

Assuming that the structure of a network influences its function (Strogatz, 2001) the question remains whether the small-world, scale-free architecture of both the enzyme and disease graph is a coincidence or is caused by biological reasons similar to those influencing the network structure of neurons (Watts and Strogatz, 1998), metabolic pathways (Fell and Wagner, 2000) or social interactions (Newman, 2001). Similar diseases are linked by the same enzymes, causing a higher clustering coefficient than would be expected of a random network. At the same time, highly connected common ailments like diabetes, arthritis and chronic kidney failure provide the short-cuts that are required within a small-world network. The same analysis holds true for the enzyme network, with large enzyme families like the protein kinases serving as central hubs connecting the different subgraphs and being responsible for the scale-free properties of the network.

The observed topology of both the enzyme and disease graph is likely to be distorted by non-biological properties. Semantic networks are known for their scale-free properties—a feature based on the way a vocabulary is created (Steyvers and Tenenbaum, 2005). Older concepts are augmented by newer ones and general terms expanded by specialized ones. Ideas and general concepts connect distinct parts of the network, similar to central metabolites like ATP connecting the metabolic pathways (Fell and Wagner, 2000; Jeong *et al*., 2000). The same features can be observed for the UMLS—older disease-related concepts, i. e. those with a lower concept identifier tend to have a higher connectivity than more recent ones

(Supplementary material). Well-known diseases might have been investigated more thoroughly than the recently added concepts for new diseases. Additionally, common ailments with a large number of studies and publications are likely to be among the first concepts to be added to a newly created ontology. An objective study of the disease network would require an ontology without this scale-free topology, which is not available.

Similar conclusions can be drawn for the enzyme graph. The existence of central nodes or enzymes in metabolic networks can be explained by the stability of the evolving networks. As long as no central enzyme is knocked out due to a mutation, scale-free networks remain highly resistant to random perturbations (Lemke *et al*., 2004; Jeong *et al*., 2001). The same consideration could be applied to the analyzed enzyme graph: evolutionary older enzymes play a more central role in the metabolism by connecting different metabolic pathways (Fell and Wagner, 2000) and, therefore, might be associated with a larger variety of diseases. Due to the grouping of individual enzymes into enzyme classes, however, this hypothesis is difficult to prove. Few publications reference the enzyme sequence that is being studied, hampering the task of annotating individual enzymes with disease-related concepts.

Given these non-biological influences an evaluation of the network topologies seems difficult and doubts about the structural analysis of other literature-derived networks. Regardless of the biological relevance, the connectivity structure needs to be kept in mind when developing a scoring function based on the frequency of co-occurring biological terms.

Conclusion and further work

A sentence-based evaluation of co-occurring enzyme names and concepts offers an easy and—in combination with several filters—a precise way to annotate biological entities. The process can be tailored to match the task at hand: high precision for immediate database enhancement, or rather high recall for manual evaluation and exploration of the created networks.

The disease-related annotation of enzyme classes generated by the workflow described in this publication was added to the BRENDA database without further manual intervention; it is freely available to the academic users and can provide valuable insights while exploring the related enzyme information. In addition, the flexibility of the presented method allows for the adaption to different knowledge domains by simply exchanging the annotated semantic fields. Initial trials annotating the enzyme classes with associated pharmaceutical compounds and symptoms proved promising, requiring only another small labeled set of sentences and a manual review of the concepts with the highest frequency.

Finally, the approach is able to grow along with the used ontology. Recent upgrades of the Unified Medical Language System include Gene Ontology and the clinical SNOMED vocabulary (Bodenreider, 2004), which are likely to improve the overall effectiveness of the annotation process. Future work includes the implementation of a scoring function taking the described network structure into account. This is likely to require some sort of concept merging prior to assessing each annotations significance.

ACKNOWLEDGEMENTS

The authors would like to thank the German Human Genome Project (DHGP) for funding this project. We are particularly thankful to the curators of the BRENDA database for their assistance during the implementation and evaluation of this system. Finally, we would like to thank the Semantic Knowledge Representation Group for making the MetaMap software available and for their help with any questions.

REFERENCES

Aronson,A. (2001) Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc. AMIA Symp.*, 17–21.

Aronson,A.R. (2002) Filtering the UMLS Metathesaurus for MetaMap. Technical report, Semantic Knowledge Representation Group.

Aronson,A.R. and Rindflesch,T.C. (1997) Query expansion using the UMLS Metathesaurus. *Proc. AMIA Annu. Fall Symp.*, pp. 485–489.

Benson,D.A. *et al*. (2003) GenBank. *Nucleic Acids Res.*, **31**, 23–27.

Berman,H. *et al*. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Blaschke,C. *et al*. (1999) Automatic extraction of biological information from scientific text: protein–protein interactions. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 60–67.

Bodenreider,O. (2004) The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.*, **32** (Database issue), D267–D270.

Bodenreider,O., Rindflesch,T. and Burgun,A. (2002) Unsupervised, corpus-based method for extending a biomedical terminology. In *Proceedings of the Workshop on Natural Language Processing in the Biomedical Domain*, Association for Computational Linguistics, Philadelphia, PA, pp. 53–60.

Cabezas,C., Resnik,P. and Stevens,J. (2001) Supervised sense tagging using support vector machines. In *Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems*, Toulouse, France, 59–62.

Camon,E. *et al*. (2004) The Gene ontology annotation (GOA) database: sharing knowledge in Uniprot with gene ontology. *Nucleic Acids Res.*, **32** (Database issue), D262-D266.

Chang,J. *et al*. (2001) Including biological literature improves homology search. *Pac. Symp. Biocomput.*, 374–383.

de Bruijn,B. and Martin,J. (2002) Literature mining in molecular biology. In *Proceedings of the EFMI Workshop on Natural Language Processing in Biomedical Applications*, Baud R. and Ruch P. (eds), Nicosia, Cyprus, pp. 1–5.

Demsar,J. and Zupan,B. (2004) Orange: from experimental machine learning to interactive data mining. White paper, Faculty of Computer and Information Science, University of Ljubljana.

Ding,J. *et al*. (2002) Mining MedLine: abstracts, sentences, or phrases? *Pac. Symp. Biocomput.*, 326–337.

Dobrokhotov,P.B. *et al*. (2003) Combining NLP and probabilistic categorisation for document and term selection for Swiss-Prot medical annotation. *Bioinformatics*, **19** (Suppl 1), i91–i94.

Donaldson,I. *et al*. (2003) PreBIND and Textomy—mining the biomedical literature for protein–protein interactions using a support vector machine. *BMC Bioinformatics*, **4**, 11.

Feldman,R. *et al*. (2003) Mining the biomedical literature using semantic analysis of natural language processing techniques. *Biosilico*, **1**, 69–80.

Fell,D.A. and Wagner,A. (2000) The small world of metabolism. *Nat. Biotechnol.*, **18**, 1121–1122.

Galperin,M.Y. (2004) The molecular biology database collection: 2004 update. *Nucleic Acids Res.*, **32** (Database issue), D3–D22.

Hirschman,L. *et al*. (2002) Accomplishments and challenges in literature data mining for biology. *Bioinformatics*, **18**, 1553–1561.

Humphreys,K. *et al*. (2000) Two applications of information extraction to biological science journal articles: enzyme interactions and protein structures. *Pac. Symp. Biocomput.*, 505–516.

Jenssen,T. *et al*. (2001) A literature network of human genes for high-throughput analysis of gene expression. *Nat. Genet.*, **28**, 21–28.

Jeong,H. *et al*. (2001) Lethality and centrality in protein networks. *Nature*, **411**, 41–42.

Jeong,H. *et al*. (2000) The large-scale organization of metabolic networks. *Nature*, **407**, 651–654.

Kanehisa,M. *et al*. (2002) The KEGG databases at GenomeNet. *Nucleic Acids Res.*, **30**, 42–46.

Kazama,J., Makino,T., Ohta,Y. and Tsujii,J. (2002) Tuning support vector machines for biomedical named entity recognition. In *Proceedings of the Workshop on Natural Language Processing in the Biomedical Domain*, Association for Computational Linguistics, Philadelphia, PA, pp. 1–8.

Lemke,N. *et al*. (2004) Essentiality and damage in metabolic networks. *Bioinformatics*, **20**, 115–119.

Liu,H. *et al*. (2002) Automatic resolution of ambiguous terms based on machine learning and conceptual relations in the UMLS. *J. Am. Med. Inform. Assoc.*, **9**, 621–636.

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

McKusick,V. (2000) Online mendelian inheritance in man OMIM (TM). Technical report, McKusick-Nathans Institute for Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD).

Mutalik,P. *et al*. (2001) Use of general-purpose negation detection to augment concept indexing of medical documents: a quantitative study using the UMLS. *J. Am. Med. Inform. Assoc.*, **8**, 598–609.

Newman,M.E. (2001) The structure of scientific collaboration networks. *Proc. Natl Acad. Sci. USA*, **98**, 404–409.

Ono,T. *et al*. (2001) Automated extraction of information on protein–protein interactions from the biological literature. *Bioinformatics*, **17**, 155–161.

Pisanelli,D.M. *et al*. (1998) An ontological analysis of the UMLS Methathesaurus. *Proc. AMIA Symp.*, 810–814.

Pratt,W. and Yetisgen-Yildiz,M. (2003) LitLinker: capturing connections across the biomedical literature. In *Proceedings of the International Conference on Knowledge Capture*, ACM Press pp. 105–112. ISBN 1-58113-583-1.

Pustejovsky,J. *et al*. (2002) Robust relational parsing over biomedical literature: extracting inhibit relations. *Pac. Symp. Biocomput.*, 362–373.

Rindflesch,T.C., Rajan,J.V. and Lawrence,H. (2000a) Extracting molecular binding relationships from biomedical text. In *Applied Natural Language Processing* Conference, Seattle, WA, pp. 188–195.

Rindflesch,T.C. *et al*. (2000b) EDGAR: extraction of drugs, genes and relations from the biomedical literature. *Pac. Symp. Biocomput.*, 517–528.

Schomburg,I. *et al*. (2004) BRENDA, the enzyme database: updates and major new developments. *Nucleic Acids Res.*, **32** (Database issue), D431–D433.

Stapley, B. and Benoit, G. (2000) Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in MedLine abstracts. *Pac. Symp. Biocomput.*, 529–540.

Steyvers,M. and Tenenbaum,J.B. (2005) The large-scale structure of semantic networks: Statistical analysis and a model of semantic growth. *Cognitive Science*, **29**.

Strogatz,S.H. (2001) Exploring complex networks. *Nature*, **410**, 268–276.

Tao,Y.-C. and Leibel,R. (2002) Identifying functional relationships among human genes by systematic analysis of biological literature. *BMC Bioinformatics*, **3**, 16.

Temkin,J.M. and Gilder,M.R. (2003) Extraction of protein interaction information from unstructured text using a context-free grammar. *Bioinformatics*, **19**, 2046–2053.

TouchGraph (2002) TouchGraph LLC. Technical report, Open Source.

Watts,D.J. and Strogatz,S.H. (1998) Collective dynamics of 'small-world' networks. *Nature*, **393**, 440–442.

Weeber,M. *et al*. (2001) Using concepts in literature-based discovery: Simulating Swanson's Raynaud–fish oil and migraine–magnesium discoveries. *JASIST*, **52**, 548–577.

Wren,J.D. and Garner,H.R. (2004) Shared relationship analysis: ranking set cohesion and commonalities within a literature-derived relationship network. *Bioinformatics*, **20**, 191–198.

Wren,J.D. *et al*. (2004) Knowledge discovery by automated identification and ranking of implicit relationships. *Bioinformatics*, **20**, 389–398.

Yang,Y. (1999) An evaluation of statistical approaches to text categorization. *Information Retrieval*, **1**, 69–90.

Yeh,A.S. *et al*. (2003) Evaluation of text data mining for database curation: lessons learned from the KDD Challenge Cup. *Bioinformatics*, **19** (Suppl 1), I331–I339.