Validating Associations in Biological Databases

Francisco M. Couto
Informatics Department
University of Lisboa, Faculty of Sciences
fcouto@di.fc.ul.pt

Mário J. Silva
Informatics Department
University of Lisboa, Faculty of Sciences
mjs@di.fc.ul.pt

Pedro M. Coutinho
Architecture et Fonction des Macromolécules Biologiques
Centre National de la Recherche Scientifique
pedro.coutinho@afmb.cnrs-mrs.fr

ABSTRACT

To cope with the large amount of biological sequences being produced, a significant number of genes and proteins have been annotated by automated tools. A protein annotation is an association between a protein and a term describing its role. These tools have produced a significant number of misannotations that are now present in biological databases. This paper proposes a new method for automatically scoring associations by comparing them to preexisting curated associations. An association is a pair that links two entities. The score can be used to filter incorrect or uncommon associations.

We evaluated the method using the automated protein annotations submitted to BioCreAtIvE, an international evaluation of state-of-the-art text-mining systems in Biology. The method scored each of these annotations and those scored below a certain threshold were discarded. The results have shown a small trade-off in recall for a large improvement in precision. For example, we were able to discard 44.6%, 66.8% and 81% of the misannotations, maintaining 96.9%, 84.2%, and 47.8% of the correct annotations, respectively. Moreover, we were able to outperform each individual submission to BioCreAtIvE by proper adjustment of the threshold.

Categories and Subject Descriptors

H.2.8 [**Information Systems**]: Database ManagementDatabases applications[Data Mining]; J.3 [**Life and Medical Sciences**]: Biology and genetics

General Terms

Algorithms, Experimentation

Keywords

Knowledge management, Biological databases, Filtering associations

CIKM'06, November 5–11, 2006, Arlington, Virginia, USA.

1. INTRODUCTION

The large amount of data available nowadays has transformed the traditional way of conducting scientific work. However, the data generated is not always accurate and detecting errors and inconsistencies in the databases is an expensive and arduous task. For example, traditional functional characterisation of genes and proteins cannot cope with the large amount of sequences being produced. Therefore, a significant number of genes and proteins have been functionally characterised by automated tools, which extrapolate functional annotations from similar sequences. However, these tools have also produced a significant number of misannotations that are now present in the databases [16]. Some of these tools have been extrapolating new annotations from misannotations and are therefore spreading the errors. This happens because most databases do not distinguish between extrapolated and curated annotations. Functional characterisation is not normally linked to the experimental evidence that substantiates it, which makes it difficult to judge if it is correct.

This paper proposes a new approach to validate uncurated associations. An association is a pair that links two members of two entities. The proposed approach compares each member of the association with the members of known curated associations. The underlying intuition is that uncurated associations having similar curated associations should also be correct. The intuition is motivated by the observation of the manual annotation technique adopted by biological curators, which consists in using preexisting curated information as a guide to evaluate uncurated biological data [8].

We applied the proposed approach to automatically filter protein misannotations by developing CAC (Correlate the Annotations' Components), a novel heuristic method that scores uncurated annotations. A protein annotation is an association between a protein and a term describing its role. CAC requires minimal human intervention, since it takes advantage of publicly available domain knowledge, i.e. previously curated annotations, to score each uncurated annotation. CAC avoids the complexities of creating rules and patterns covering all possible cases or creating training sets that are too specific to be extended to new domains [29]. Besides avoiding direct human intervention, automatically collected domain knowledge is usually much larger than manually generated domain knowledge and does not become outdated, since public databases can be tracked for updates as they evolve [10].

An example scenario where automated annotation systems produced a significant number of misannotations was BioCreAtIvE (Critical Assessment of Information Extraction systems in Biology) [22]. We applied CAC to all the annotations submitted to BioCreAtIvE, and CAC was able to obtain a good accuracy by discarding a significant number of these misannotations. The results obtained by CAC in this task demonstrate the efficiency and feasibility of the proposed approach.

The remainder of this paper is organised as follows. Section 2 introduces the Gene Ontology. Section 3 describes BioCreAtIvE and discusses the results obtained by its participants. Section 4 describes CAC in detail. Section 5 presents the experimental evaluation of CAC using the annotations submitted to BioCreAtIvE. Section 6 discusses the obtained results. Finally, Section 7 expresses our main conclusions.

2. GENE ONTOLOGY

Biological databases annotate genes or proteins with statements that describe their biological role. Sometimes, these annotations are stored as ambiguous statements that are domain specific and context dependent. To cope with this, the research community is developing and using BioOntologies to annotate genes and proteins [30]. Using a BioOntology to annotate genes or proteins avoids ambiguous statements that are domain specific and context dependent.

For example, the GO (Gene Ontology) is a well-established structured vocabulary that for example has been successfully used for gene annotation of different species [18]. The GO project is one of the major efforts in Molecular Biology, for constructing a BioOntology of broad scope and wide applicability. GO provides a structured controlled vocabulary of gene and protein biological roles, which can be applied to different species. GO comprised 20,069 distinct terms in December 2005. Since the activity or function of a protein can be defined at different levels, GO has three different aspects: *molecular function, biological process* and *cellular component.* Each protein has elementary molecular functions that normally are independent of the environment, such as catalytic or binding activities. Sets of proteins interact and are involved in cellular processes, such as metabolism, signal transduction or RNA processing. Proteins can act in different cellular localisations, such as the nucleus or membrane.

GO organises the concepts as a DAG (Directed Acyclic Graph), one for each aspect. Each node of the graph represents a concept, and the edges represent the links between concepts (see example in Figure 1). Links can represent two relationship types: *is-a* and *part-of.* GO is a dynamic hierarchy: its content changes every month with the publication of a new release. Any user can request modifications to GO, which is maintained by a group of curators who add, remove and change terms and their relationships in response to modification requests. This prevents GO from becoming outdated and from providing incorrect information.

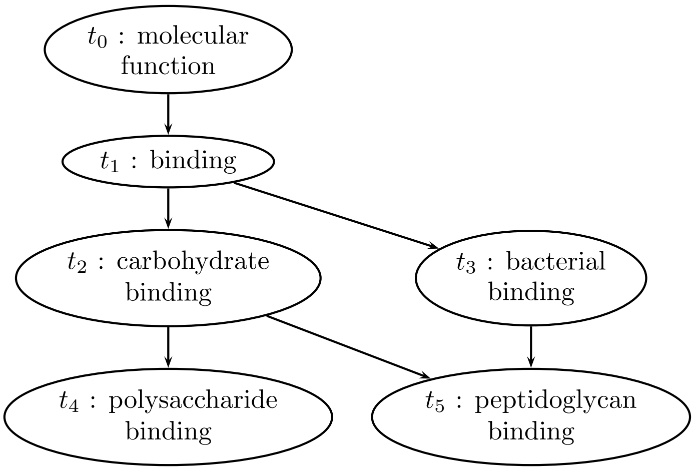

t_0 : molecular function
t_1 : binding
t_2 : carbohydrate binding
t_3 : bacterial binding
t_4 : polysaccharide binding
t_5 : peptidoglycan binding

Figure 1: Sub-graph of GO.

GO started by adding generic terms and simple relationships to provide a complete coverage of the Molecular Biology domain. Thus, the main limitation of GO is the lack of specific terms that, for example, represent precise biochemical reactions. However, as different research communities understand the importance of adding their domain knowledge to GO, it will acquire more specific terms and relationships and therefore overcome this limitation.

Many databases are using GO terms to annotate their proteins. For example, the GOA (Gene Ontology Annotation) database provides GO annotations to supplement the UniProt (Universal Protein Resource) [8]. UniProt is a universal repository of protein sequence and functional data [2]. GOA provides high-quality manual GO annotations, but manual curation is a time-consuming task that currently covers less than 5% of UniProt. The manual processing capacity for gene and protein characterisation is overloaded by the increasingly larger amounts of literature to analyse. Thus, the GOA database mainly consists of uncurated annotations that have a lower quality than manual annotations.

3. BIOCREATIVE

A large amount of the information discovered in Molecular Biology has been mainly published in BioLiterature (a shorter designation for the biological and biomedical scientific literature). Analysing and identifying information in a large collection of unstructured texts is a painful and hard task, even to an expert. To improve the access to the information, most researchers also deposit their findings in a structured form in databases, such as UniProt, which collect and distribute biological information. However, the management of these databases also became a complex problem, and most of them contain a significant number of errors. Moreover, most facts are only valid in a specific biological setting, and should not be directly extrapolated to other cases. Therefore, researchers cannot only rely in the facts available in these databases, they also need the evidence substantiating the facts, which is normally present in the BioLiterature. The evidence text can be the description of the biological setting where the experiment was conducted or the subsequent discussion of the results. In addition, different research communities have different needs and requirements at a given period in time. As these constraints evolve, its man-

Participant	Approach
Ehrler et al. [17]	Sequentially applied finite state automata
Couto et al. [13]	Information content of terms
Verspoor et al. [31]	Word proximity networks
Rice et al. [28]	Term-based SVM
Ray et al. [25]	Statistical learning and Naïve Bayes method
Chiang et al. [9]	Pattern matching and Sentence classification

Table 1: Participants of the subtask 2.2 of BioCreAtIvE and their approaches.

agement becomes harder to fulfil by databases, which have a static structure. Thus, researchers tend to use databases as an additional source to store and find facts, but the evidence substantiating them is still described as unstructured text, given its higher flexibility. As a consequence, a large amount of the knowledge acquired in Molecular Biology can only be found in the BioLiterature.

An approach to improve the access to the knowledge published in BioLiterature is to use Text Mining, which aims at automatically extracting knowledge from natural language text [20]. The application of text-mining tools to BioLiterature started just a few years ago [1]. Since then, the interest in the topic has been steadily increasing, motivated by the vast amount of documents that curators have to read to update biological databases, or simply to help researchers keep up with progress in a specific area [11]. Thus, bioinformatics tools are increasingly using Text Mining to collect more information about the concepts they analyse. Text-mining tools have mainly been used to identify: entities, such as genes, proteins and cellular components; relationships, such as protein localisation or protein interactions; events, such as experimental methods used to discover protein interactions.

An important application of text-mining tools is the automatic annotation of genes and proteins. A gene or protein annotation consists of a pair composed by the gene or protein and a description of its biological role. The biological role is often a concept from a BioOntology (e.g. GO). Using a BioOntology to annotate genes or proteins avoids ambiguous statements that are domain specific and context dependent. To understand the activity of a gene or protein, it is also important to know the biological entities that interact with it. Thus, the annotation of a gene or protein also involves identifying interacting chemical substances, drugs, genes and proteins.

Most of the manual annotation process done by the GOA team involves analysing the literature, which is a painful and hard task, even to an expert. Thus, the GOA team accepted to take part in BioCreAtIvE, to access the ability of text mining-systems for assisting curators in the annotation of UniProt proteins to GO terms. BioCreAtIvE was a challenging evaluation that compared the performance of different text-mining systems in solving common tasks using the same corpus. The tasks addressed meaningful challenges for text-mining systems and at the same time real problems of Biology. The biologically realistic scenarios posed additional difficulties for the participants, which resulted in less successful performances than to the ones obtained in the Genomics TREC 2004, a similar challenging evaluation [21].

The subtask 2.2 of BioCreAtIvE aimed at predicting GO annotations to human proteins from 200 new full-text articles from the *Journal of Biological Chemistry.* Table 1 shows the participants of this subtask and the approaches used. Each participant could submit three different sets of predictions to test the parameters of his system. Overall, there were 18 sets of submitted annotations that were individually evaluated.

For each scientific article, the participants had to submit the list of annotations predicted by their system and evidence text for each annotation. Three curators of the GOA team manually evaluated each predicted annotation and respective evidence [7]. They evaluated if the predicted GO term assignment was correct, or close to what a curator would choose. Sometimes, the GO term was in the correct lineage, but the curators considered it as incorrect because it was too generic or too specific. The GOA team considered a submission correct when it contained both a correct annotation and a valid evidence text substantiating it.

The predictions submitted to this subtask achieved unacceptable levels of accuracy. The participant with the best accuracy identified 6% of all the correct annotations found by all the participants, and only 35% of his predictions were correct. The task addressed by BioCreAtIvE is representative of the complexities that have to be faced in real biological research environments. Without improvements, such automated systems are unhelpful to curators [26]. Therefore, techniques that could achieve good solutions to validate the automated annotations and improve their accuracy are much needed.

4. CAC

CAC assumes that an annotation is correct when there is at least a preexisting curated annotation composed by a similar gene (or protein) and a similar property. CAC considers an annotation as a pair (g, p), where g is a gene (or a protein) and p a biological property. For example, the annotations submitted to BioCreAtIvE were composed by a UniProt protein and a GO term that are instances of gene and property, respectively.

Algorithm 1 outlines CAC, which assigns a confidence score to $a_{predicted}$, an annotation predicted by an automated system given as input. CAC also receives as input $\mathcal{A}_{curated}$, a set of preexisting curated annotations collected from public databases, e.g. GOA.

CAC starts by assigning a zero confidence score to the predicted annotation (line 1). Next, CAC collects all the genes in the set of curated annotations (line 3). For each curated

Algorithm 1 CAC

```
Input: $a_{predicted}$, an uncurated annotation predicted by an automated system;
    $\mathcal{A}_{curated}$, set of previously curated annotations.
Output: $confidence \in [0, +\infty]$, confidence score of the predicted annotation.
 1: $confidence(a_{predicted}) = 0$
 2: $(g_{predicted}, p_{predicted}) = a_{predicted}$
 3: $\mathcal{G}_{curated} = \{g : \exists p \ (g, p) \in \mathcal{A}_{curated}\}$
 4: for all $g_{curated} \in \mathcal{G}_{curated}$ do
 5:     $\mathcal{P}_{curated} = \{p : (g_{curated}, p) \in \mathcal{A}_{curated}\}$
 6:     $geneSim = geneSim(g_{predicted}, g_{curated})$
 7:     $propSim = \sum_{p_{curated} \in \mathcal{P}_{curated}} propSim(p_{predicted}, p_{curated})$
 8:     $confidence(a_{predicted}) += geneSim \times propSim$
 9: end for
10: $SG = similarGenes(g_{predicted}, \mathcal{G}_{curated})$
11: $confidence(a_{predicted}) = \frac{confidence(a_{predicted})}{SG}$
```

gene, CAC collects the properties annotated to it (line 5). Next, CAC calculates the similarity between the curated and the predicted genes (line 6), and calculates the similarity between the predicted property and each property annotated to the curated gene (line 7). CAC increments the confidence of the predicted annotation by the product of the gene similarity and the sum of all property similarities (line 8). Thus, the confidence only increases if both the gene similarity and at least one property similarity are larger than zero, i.e., if they are similar genes and have been annotated with at least one similar property.

However, the $\mathcal{A}_{curated}$ set can contain groups of similar genes that are over-represented. In this case, the predicted annotations that contain genes with a large number of similar curated genes will tend to have higher confidence scores. To overcome this problem, CAC calculates the number of curated genes similar to the predicted gene (line 10), and employs it as a damping factor (line 11). This factor reduces the effect of the amount of similar curated genes in the confidence score calculation.

CAC returns a confidence score of $a_{predicted}$ being correct. To filter the annotations predicted by an automated system, CAC scores each predicted annotation and discards those scored below a confidence threshold (CT). CAC is able to trade precision against recall by manipulating CT. Raising CT increases precision and decreases recall, lowering CT has the opposite effect.

CAC cannot score annotations without similar curated annotations. When the given predicted annotation has no similar curated genes ($SG = 0$), CAC assigns a confidence score of $+\infty$ to it. This means that the predicted annotation will never be filtered independently of the threshold used. Therefore, CAC does not discard new knowledge; instead, it gives the curators the opportunity to manually verify these potentially novel annotations.

4.1 Gene Similarity

The most popular way to calculate the similarity between two genes is by comparing their sequence [3]. However, sequence similarity is not the only kind of structural similarity that can be computed between two proteins. Family similarity is also a structural similarity of a higher level than sequence similarity. Each family describes a set of related proteins, which can have identical molecular functions, are involved in the same process, or act in the same cellular location. Classifying proteins in families has been a common technique to organise them according to their biological role. For example, the most successful large-scale effort for increasing the coverage of GO annotations within the UniProt database is based on the exploitation of family annotations [8]. Unlike standard sequence similarity methods, family categorisation is normally based on experimental results about protein domains, which represent some evolutionarily conserved structure and have implications on the protein's biological role.

geneSim was calculated from the number of shared Pfam families. Pfam is a database that provides a set of protein domains and families [4]. These families are constructed semi-automatically using hidden Markov models (HMMs). Each family describes a set of related proteins that can have identical molecular functions, are involved in the same process, or act in the same cellular location. This database contained 8183 families in December 2005. The UniProt database provides family assignments, where each protein is assigned to a set of Pfam families. This calculation can be improved by taking in account the sequence related to each Pfam family. For example, the length of the sequence and the percentage of similarity may constitute important factors to calculate the *geneSim* function. Apart from the sequence the *geneSim* could also use other type of information, e.g. gene expression profiles and evolutionary profiles.

4.2 Property Similarity

CAC assumes that two properties are similar if one of them subsumes the other or if they have a common parent in the functional classification scheme, e.g. GO. To calculate the degree of similarity between properties, CAC can use a semantic similarity measure that combines the structure and content of a BioOntology with statistical information from corpus [27]. Recent projects investigated the use of semantic similarity measures over GO [14, 15, 24]. Their results demonstrated the feasibility of a semantic similarity measure in a biological setting.

propSim was calculated using the measure proposed by Jiang&Conrath which is one of the most efficient semantic similarity measures [6, 23]. Jiang&Conrath defined the semantic distance of two concepts in a corpus as the difference between their information content and the information content of their most informative common ancestor. The information content of a concept is inversely proportional to its frequency in the corpus. Concepts that are frequent in the corpus have low information content. For example, the stop words (such as *the*) that occur almost everywhere in the text normally provide little semantic information. The information content of a GO term was calculated as the number of proteins annotated with it. The ancestor of two GO terms having the largest information content was considered the most informative common ancestor of both terms.

4.3 Computational Performance

Set	*#annotations*	*#proteins*	$max(SG)$	$min(SG)$	$\overline{SG}$
Set-1	1135	30	583	5	223.7841
Set-2	1101	25	1762	613	1077.7221
Set-3	1049	22	11605	1855	3098.9790

Table 2: Statistics of the three sets of annotations created according to the number of similar curated proteins per annotation (SG). The statistics include the number of annotations, the number of distinct predicted proteins, and the maximum, minimum and average of SG for each set.

We implemented CAC as a Java/MySQL application [19]. The execution time of CAC is linearly proportional to the number of curated annotations used, which makes it scalable. The performance of CAC is directly linked to the time spent on the calculation of both *geneSim* and *propSim*. *geneSim* can be implemented as a simple SQL query counting the number of shared families, and therefore it is not computational expensive. On the other hand, the calculation of *propSim* is more complex but it is also not computational expensive as it is demonstrated by FuSSiMeG (Functional Semantic Similarity Measure between Gene-Products), a web tool that measures the functional similarity between proteins based on the semantic similarity of the GO terms annotated to them [12]. FuSSiMeG is available on the Web[1], affording the similarity calculation on the fly.

4.4 Example

In the subtask 2.2 of BioCreAtIvE, the participants annotated the protein *Lipid phosphate phosphohydrolase 1* to the GO terms *membrane* and *mRNA metabolism* [5]. However, only the assignment of *membrane* is correct. Below the results obtained by CAC for these two annotations are described.

The protein *Lipid phosphate phosphohydrolase 1* belongs to the *PF01569* family. For the annotation of this protein to *membrane,* CAC found 91 curated proteins from the *PF01569* family ($geneSim = 1$) that were annotated to similar GO terms ($propSim > 0$) in GOA. From these 91 proteins, 21 were annotated to the same term. For example, the protein *Lipid phosphate phosphohydrolase 2* belongs to the *PF01569* family ($geneSim = 1$) and is annotated to *membrane* and *integral to membrane,* which results in $propSim = 1.445297776$. The confidence score resulted from these 91 proteins is 53.09, but since the *PF01569* family contains 630 proteins ($SG = 629$), CAC returned $\frac{53.09}{639} \approx 0.08$.

On the other hand, for the annotation of the protein *Lipid phosphate phosphohydrolase 1* to *mRNA metabolism,* CAC only found one curated protein (*HH1165*) from the *PF01569* family ($geneSim = 1$) that was annotated to a similar GO term (*metabolism*) ($propSim = 0.1$) in GOA. Thus, in this case CAC returned $\frac{0.1}{639} \approx 0.0002$.

5. ASSESSMENT

We tested CAC to find how effectively it could discard the misannotations submitted to BioCreAtIvE independently of their evidence text. CAC scored each submitted annotation individually ($a_{predicted}$), using the GOA annotations as the curated set of annotations ($\mathcal{A}_{curated}$). The annotations submitted to BioCreAtIvE[2] and the GOA[3] annotations are both publicly available on the Web. However, in the publicly available information there is no reference to the author of each annotation submitted to BioCreAtIvE. It is not even possible to know which annotations were submitted by the same system.

We decided not to increase the confidence of a predicted annotation based on curated annotations to the same protein, i.e., the protein $g_{predicted}$ was discarded from $\mathcal{G}_{curated}$. This way, CAC was restricted to score each predicted annotation based only on curated annotations to similar but distinct proteins. This restriction ensures a fair evaluation of CAC by checking if CAC copes with proteins having no previously curated annotations.

The restriction increased the number of proteins for which it was not possible to obtain similar proteins, i.e., having $SG = 0$. However, only 455 out of the 3740 predicted annotations did not have a similar protein in the December 2004 release of GOA. These novel annotations have a precision of 7%, i.e., only 32 of them were correct. The assumption that supports CAC is not applicable to these novel annotations, thus scoring these annotations is out of CAC objectives. CAC does not discard these annotations, since it assigns an infinite score to them. Therefore, in the first part of the evaluation these annotations were disregarded, but they were included in the end to show the overall impact of CAC on the curation process.

The 3285 annotations having $SG > 0$ assign 1239 distinct GO terms to 77 UniProt proteins. The 77 proteins were assigned to 87 distinct Pfam families with an average of 1.6 families per protein. These 87 families contained 64863 distinct proteins. Thus, each protein had $\frac{64863}{87} \times 1.6 = 1192.9$ similar curated proteins on average.

To compare the performance of CAC when applied to over-annotated or under-annotated proteins, the 3285 annotations were divided in three different sets (*Set-1*, *Set-2* and *Set-3*) according to the number of similar curated proteins (SG). Table 2 shows statistical information about each set.

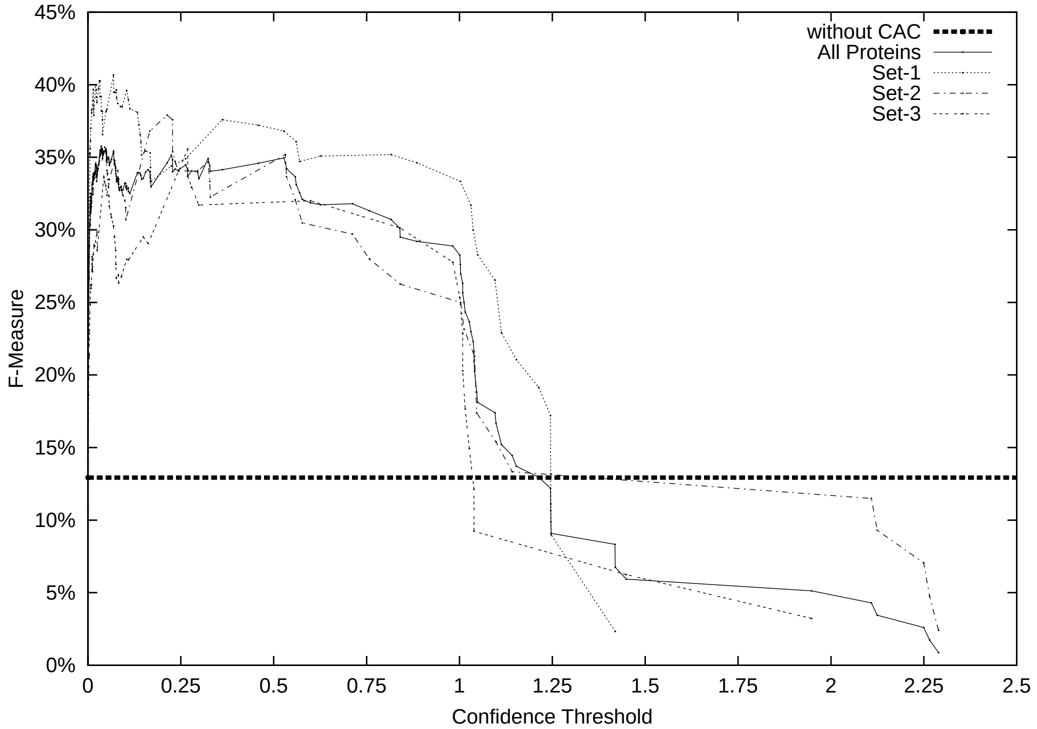
5.1 Results

Each distinct confidence score was used as a confidence threshold to obtain different subsets of the 3285 predicted annotations. For each confidence threshold, the resulting subset contains all the annotations with a confidence score not below the threshold. For a zero confidence threshold, the

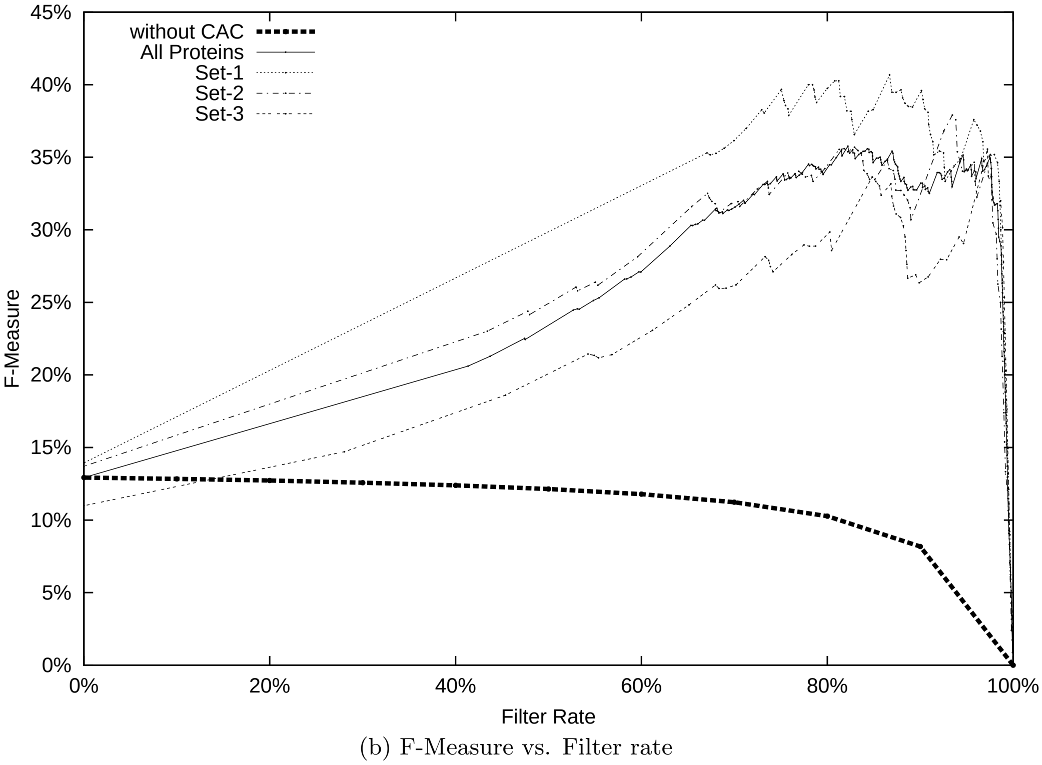
[1] http://xldb.fc.ul.pt/rebil/tools/ssm/

[2] http://www.pdg.cnb.uam.es/BioLINK/workshop_BioCreative_04/results/data/

[3] ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/UNIPROT/gene_association.goa_uniprot.gz

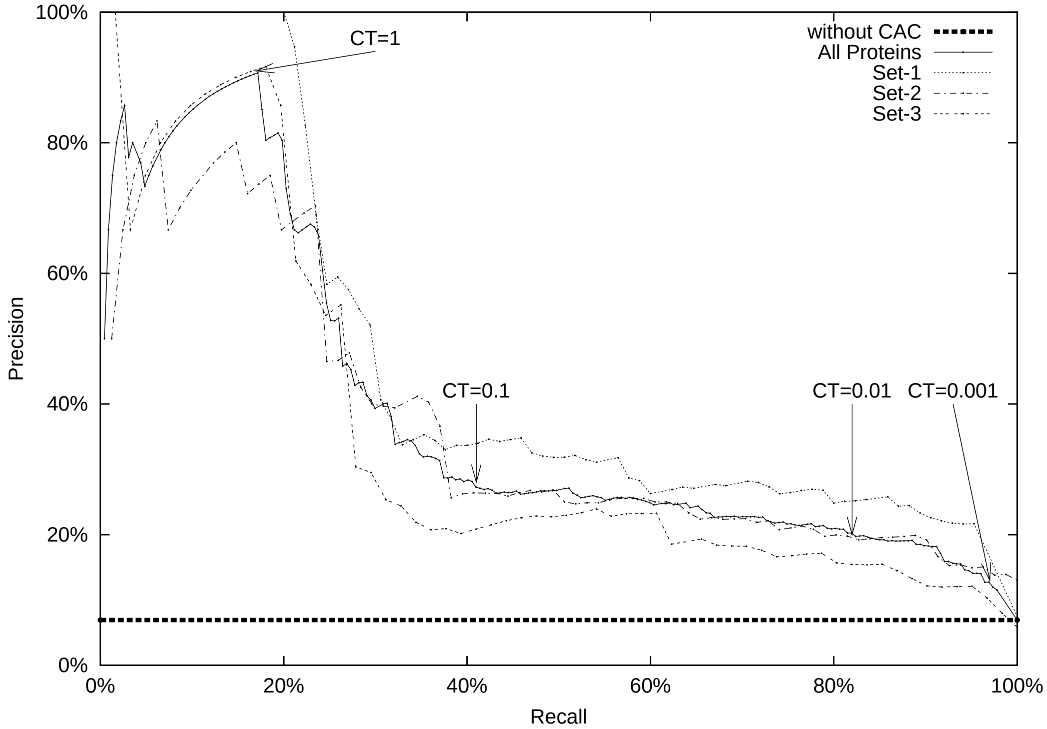


(a) F-Measure vs. CT



(b) F-Measure vs. Filter rate

subset contains all the predicted annotations, since none of them is discarded. As the confidence threshold increases, the size of the subset decreases. For each subset, it was calculated: the precision, representing the fraction of correct annotations in the subset; the recall, representing the number of correct annotations in the subset over the number of correct annotations in the original set; and the F-measure $= \frac{2 \times precision \times recall}{precision + recall}$, representing the trade-off between pre-



(c) Precision vs. Recall

Figure 2: Accuracy of the annotations retained by different confidence thresholds (CT) after running CAC. The *All Proteins* lines represent all the 3285 annotations. The *Set 1* and *Set 3* lines represent the annotations with the smallest and the largest number of similar curated proteins, respectively. The *Set-2* lines represent all the other annotations not present in *Set 1* and *Set 3*. The *without CAC* baselines represent the original annotations without using CAC. In chart (a), the baseline shows the F-Measure when none of annotations is filtered. In the other charts, the baselines assume a random model to filter the annotations, i.e., having a constant precision for any filter rate.

cision and recall. Note that if we replace CAC by a random model to filter the annotations, the precision would remain constant. For instance, if we select at random 25% of the annotations in the original set, it is predictable that the selected annotations also contain 25% of the correct annotations in the original set.

Only 227 out of the 3285 annotations submitted to BioCreAtIvE were considered correct, a precision of 6.9%. The real recall is unknown, since the organisation of BioCreAtIvE did not measure it. Thus, we can assume a recall of 100% for the original set of annotations. Note that CAC cannot increase recall. As a filter, it does not generate new annotations.

Figure 2(a) shows the F-measure for different confidence thresholds. For confidence thresholds smaller than one, the chart shows that the use of CAC to discard annotations is beneficial by achieving a substantial improvement in F-measure. The F-measure achieves its maximum value when the confidence threshold is around 0.1. Figure 2(c) shows the precision and recall obtained for different confidence thresholds. With a few exceptions, we have a steadily increase in precision as we increase the confidence threshold.

Table 3 shows the accuracy of the predicted annotations when not using CAC ($CT = 0$), and the accuracy of the subsets of annotations retained by different confidence thresholds. Besides the precision, recall and F-measure, the Table shows the number of correct and incorrect annotations that were not discarded by CAC, and the percentage of misannotations discarded by CAC from the original set. For example, by using $CT = 0.001$ CAC discarded 50.8% ($\frac{3058-1506}{3058}$) of the misannotations, maintaining 96.5% ($\frac{219}{227}$) of the correct annotations.

The confidence threshold has no biological meaning to curators. They simply would like to discard a given amount of annotations to speedup the curation process without loosing a significant part of valuable information. This can be done by increasing CT until a defined filter rate is reached. The filter rate means the percentage of annotations that are discarded by CAC from the original set. For example, a filter rate of 90% means that only 10% of the original annotations were retained. Figure 2(b) shows the F-measure obtained by CAC for different filter rates. The chart shows that the use of CAC to discard annotations is beneficial by achieving a steady improvement in F-measure as we increase the filter rate, except for filter rates larger than 99% ($CT > 1$). Table 4 shows the precision and the recall of the different sets of annotations over different filter rates, together with the

CT	Filter Rate	#correct	#incorrect	Precision	Recall	F-measure	Misannotations Discarded
0	0%	227	3058	6.9%	100%	12.9%	0%
0.001	47.5%	219	1506	12.7%	96.5%	22.4%	50.8%
0.01	72%	186	733	20.2%	81.9%	32.5%	76%
0.1	90%	92	235	28.1%	40.5%	33.2%	92.3%
1	98.7%	39	4	90.7%	17.2%	28.9%	99.9%

Table 3: Results obtained by filtering the 3285 annotations using different confidence thresholds.

	All Proteins		
Filter Rate	Precision	Recall	CT
0%	6.9%	100%	0
70%	19.3%	84.6%	0.008
80%	22.6%	67%	0.025
90%	27.3%	41%	0.094
95%	40.6%	29.5%	0.235

	Set-1			Set-2			Set-3		
Filter Rate	Precision	Recall	CT	Precision	Recall	CT	Precision	Recall	CT
0%	7.5%	100%	0	7.4%	100%	0	5.8%	100%	0
70%	22.6%	90.6%	0.007	19.8%	81.5%	0.008	15.4%	82%	0.008
80%	27.3%	72.9%	0.028	23.3%	64.2%	0.028	18.4%	67.2%	0.018
90%	32.5%	47.1%	0.091	25.6%	38.3%	0.102	20.8%	36.1%	0.083
95%	40.6%	30.6%	0.263	40%	29.6%	0.243	30.4%	27.9%	0.162

Table 4: Results obtained by filtering the 3285 annotations using different filter rates.

selected CT in each set. The standard deviation of both recall and precision is always smaller than 5% for the same filter rate, even with a standard deviation of 0.8% in precision in the original sets. The selected CT is almost the same in all sets, except in the *Set-3* where in some cases CT is about 1/3 smaller.

6. DISCUSSION

The increase in precision is already a positive result to GOA curators, since they primarily require high precision in an automated annotation system. In this experiment, CAC increased precision at the cost of a low decrease in recall. The trade-off between precision and recall is worth it, as it is shown by the increase in the F-measure. This is always true except for filter rates larger than 99% ($CT > 1$), because recall decreases and precision is not improved. For such high confidence thresholds, there are still some misannotations not discarded. For example, CAC assigned a high confidence score to the annotation that assigns the GO term *kinase activity* to the protein *Sulfate transporter 1.2,* but this annotation is not in GOA. However, the GO term *protein kinase activity* is annotated to the same protein in GOA. Since the term *kinase activity* is a generalisation of *protein kinase activity,* the predicted annotation is correct but still not of interest to curators.

From 3058 misannotations, four remain with a confidence threshold of one. These four annotations are not defined in GOA because all of them assign generic GO terms to proteins. This does not mean that they represent incorrect assignments, they are only too generic to be of interest to curators. Since in reality these generic annotations are correct, CAC does not discard these annotations even with large confidence thresholds. This explains the sharp drop in precision when recall is close to zero in Figure 2(c), because these generic annotations were considered incorrect in our assessment. Thus, by considering generic annotations as correct, the performance of CAC would increase, but this would not reflect the curators' interest for precise and specific annotations. Nevertheless, it is undesirable to discard these generic annotations, since the evidence substantiating them may be of interest to curators.

The participant of BioCreAtIvE who achieved the largest precision predicted 41 annotations, 14 of which were correct. Using a confidence threshold of 1, CAC selected 43 annotations, 39 of which were correct. On the other hand, the participant who achieved the largest recall predicted 661 annotations, 78 of which were correct. Using a confidence threshold of 0.1, CAC selected 327 annotations, 92 of which were correct. Therefore, by proper adjustment of the confidence threshold we can use CAC to outperform each individual submission to BioCreAtIvE.

For a small decrease in recall, CAC was able to obtain a large improvement in precision, since annotations that clearly do not satisfy the correlation between structure and function are normally incorrect. Unfortunately, there are exceptions. Using a confidence threshold of 0.001, CAC discarded 8 out of 227 correct annotations. For these eight annotations, CAC could not find similar annotations mainly because of the restriction that discarded curated annotations to similar but distinct proteins. When CAC was tested without this restriction, 47% of the misannotations were discarded maintaining all the correct annotations, i.e., a two-fold increase in precision maintaining 100% recall. This restriction was applied to ensure a fair evaluation of CAC. However, in a real application setting, this restriction would not be applied and therefore obtain a higher performance. It is expected that, as the scientific community produces better classifica-

CT	Filter Rate	#correct	#incorrect	Precision	Recall	F-measure	Misannotations Discarded
0	0%	259	3481	6.9%	100%	13%	0%
0.001	41.7%	251	1929	11.5%	96.9%	20.6%	44.6%
0.01	63.3%	218	1156	15.9%	84.2%	26.7%	66.8%
0.1	79.1%	124	658	15.9%	47.8%	23.8%	81.0%
1	86.7%	71	427	14.3%	27.4%	18.8%	87.7%

Table 5: Results obtained by filtering all the 3740 annotations using different confidence thresholds.

tion schemes, CAC will also improve its performance.

The results of the three different sets of annotations show that CAC is not biased toward proteins with a large number of similar curated proteins. In Figure 2, the results of these sets were uniform over all the confidence thresholds. The small differences are due to different precision values of each original set. The Set 1 of under-annotated proteins has the highest precision (7.5%) and the Set 3 of over-annotated proteins has the lowest precision (5.8%). The Set 1 achieves a precision of 100% for a recall larger than 20%, because any correct annotation to under-annotated proteins is of interest to curators, i.e., the problem of generic annotations described above is not applicable to these proteins.

The results show that the performance obtained by a given filter rate is preserved when applied to different sets of annotations. Therefore, curators can expect to obtain similar performances in different sets of annotations by using similar filter rates. Using different sets of curated and uncurated annotations may imply different CT for obtaining the same filter rate. For example, the uncurated annotations in *Set-3* have more similar curated annotations, thus it is also expected to have larger confidence scores. However, curators can easily adjust CT to obtain a required filter rate.

CAC does not discard new knowledge, but it does not discard the misannotations to under-annotated proteins either. To measure the real impact of using CAC on the curation process it should take into account the 455 novel annotations. CAC never discards these annotations, leaving the decision to the curator by assigning an infinite confidence score to them. Table 5 shows that including these novel annotations has a small effect on the performance of CAC. For example, by using a filter rate of 41.7% ($CT = 0.001$) the curator only has to verify 58.3% (100%-41.7%) of the original annotations only loosing 3.1% (100%-96.9%) of the correct annotations. However, the precision for large filter rates is constrained by the precision of the novel annotations. Since CAC does not discard any of the 455 novel annotations, the precision converges to 7% (32 out of 455 annotations are correct) as CT increases. Nevertheless, CAC can overcome this limitation and contribute toward adding new knowledge. Nowadays, there are automated systems that predict generic annotations with high precision. If these generic annotations were considered, CAC would use them to score specific annotations, which is what curators really want. CAC can also be used to crosscheck annotations predicted by different automated systems. For example, CAC can score annotations predicted by a text-mining system based on annotations predicted by sequence similarity.

7. CONCLUSIONS

This paper proposed the use of curated associations as domain knowledge for scoring uncurated associations. To demonstrate its feasibility and efficiency, we developed and evaluated CAC, which scores uncurated annotations based on similar curated annotations. The results obtained in a realistic scenario show that CAC can effectively be used to speed up the curation process by discarding a large amount of misannotations without loosing a significant amount of correct annotations. Thus, CAC can be used by any automated annotation system to improve the accuracy and to reduce the effort of curators. Its main advantage is that it requires minimal human intervention, since CAC uses extensive domain knowledge automatically collected from public databases.

The precision/recall trade-off is tunable by a method's confidence threshold, which can be adjusted to obtain different filter rates according to the curator's requirements. The results obtained by similar filter rates were consistent for different subsets of the annotations, so the effectiveness of CAC is predictable as we change a single tuning parameter.

One may argue that the proposed approach is only effective when there is a substantial amount of curated information available. However, the amount of curated information will tend to increase as more genes are characterised, especially for model organisms (e.g. human) whose characterisation has a great fundamental economical and social impact. On the other hand, the percentage of curated genes will tend to decrease, since the manual characterisation efforts are powerless to overcome the huge amount of data being generated by high-throughput analysis tools. Therefore, automatic methods, such as CAC, are much required to help the characterisation efforts.

CAC can be easily adapted to score associations between other objects than genes and biological properties. All it requires is a similarity measure for each kind of object used and a set of curated associations.

8. REFERENCES

[1] M. Andrade and P. Bork. Automated extraction of information in Molecular Biology. *FEBS Letters*, 476:12–17, 2000.

[2] R. Apweiler, A. Bairoch, C. Wu, W. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. Martin, D. Natale, C. O'Donovan, N. Redaschi, and L. Yeh. UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, 32(Database issue):D115–D119, 2004.

[3] T. Attwood and D. Parry-Smith. *Introduction to Bioinformatics*. Longman Higher Education, 1999.

[4] A. Bateman, L. Coin, R. Durbin, R. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E. Sonnhammer, D. Studholme, C. Yeats, and S. Eddy. The Pfam protein families database. *Nucleic Acids Research*, 32(Database issue):D138–D141, 2004.

[5] C. Blaschke, E. Leon, M. Krallinger, and A. Valencia. Evaluation of BioCreAtIvE assessment of task 2. *BMC Bioinformatics*, 6(Suppl 1):S16, 2005.

[6] A. Budanitsky and G. Hirst. Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In *Proc. of the Workshop on WordNet and Other Lexical Resources co-located with the 2nd North American Chapter of the Association for Computational Linguistics*, June 2001.

[7] E. Camon, D. Barrell, E. Dimmer, V. Lee, M. Magrane, J. Maslen, D. Binns, and R. Apweiler. An evaluation of GO annotation retrieval for BioCreAtIvE and GOA. *BMC Bioinformatics*, 6(Suppl 1):S17, 2005.

[8] E. Camon, M. Magrane, D. Barrell, V. Lee, E. Dimmer, J. Maslen, D. Binns, N. Harte, R. Lopez, and R. Apweiler. The Gene Ontology Annotations (GOA) database: sharing knowledge in UniProt with Gene Ontology. *Nucleic Acids Research*, 32:262–266, 2004.

[9] J. Chiang and H. Yu. Extracting functional annotations of proteins based on hybrid text mining approaches. In *Proc. of the BioCreAtIvE Challenge Evaluation Workshop*, 2004.

[10] F. Couto, B. Martins, and M. Silva. Classifying biological articles using web resources. In *Proc. of the 2004 ACM Symposium on Applied Computing*, 2004.

[11] F. Couto and M. Silva. *Advanced Data Mining Techonologies in Bioinformatics*, chapter Mining the BioLiterature: towards automatic annotation of genes and proteins. Idea Group Inc., 2006.

[12] F. Couto, M. Silva, and P. Coutinho. Implementation of a functional semantic similarity measure between gene-products. DI/FCUL TR 03–29, Department of Informatics, University of Lisbon, November 2003.

[13] F. Couto, M. Silva, and P. Coutinho. Finding genomic ontology terms in text using evidence content. *BMC Bioinformatics*, 6(S1):S21, 2005.

[14] F. Couto, M. Silva, and P. Coutinho. Semantic similarity over the gene ontology: Family correlation and selecting disjunctive ancestors. In *Proc. of the ACM Conference in Information and Knowledge Management as a short paper*, 2005.

[15] F. Couto, M. Silva, and P. Coutinho. Measuring semantic similarity between gene ontology terms. *DKE - Data and Knowledge Engineering, Elsevier Science (in press)*, 2006.

[16] D. Devos and A. Valencia. Intrinsic errors in genome annotation. *Trends Genetics*, 17(8):429–431, 2001.

[17] F. Ehrler, A. Jimeno, and P. Ruch. Data-poor categorization and passage retrieval for Gene Ontology annotation in Swiss-Prot. *BMC Bioinformatics*, 6(Suppl 1):S23, 2005.

[18] GO-Consortium. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research*, 32(Database issue):D258–D261, 2004.

[19] M. Grand. *Java Language Reference*. O'Reilly, 1997.

[20] M. Hearst. Untangling text data mining. In *Proc. of the 37th Annual Meeting of the Association for Computational Linguistics*, 1999.

[21] W. Hersh, R. Bhuptiraju, L. Ross, P. Johnson, A. Cohen, and D. Kraemer. TREC 2004 genomics track overview. In *Proc. of the 13th Text REtrieval Conference*, 2004.

[22] L. Hirschman, A. Yeh, C. Blaschke, and A. Valencia. Overview of BioCreAtIvE: critical assessment of information extraction for biology. *BMC Bioinformatics*, 6(Suppl 1):S1, 2005.

[23] J. Jiang and D. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proc. of the 10th International Conference on Research on Computational Linguistics*, 1997.

[24] P. Lord, R. Stevens, A. Brass, and C. Goble. Semantic similarity measures as tools for exploring the Gene Ontology. In *Proc. of the 8th Pacific Symposium on Biocomputing*, 2003.

[25] S. Ray and M. Craven. Learning statistical models for annotating proteins with function information using biomedical text. *BMC Bioinformatics*, 6(Suppl 1):S18, 2005.

[26] D. Rebholz-Schuhmann, H. Kirsch, and F. Couto. Facts from text - is text mining ready to deliver? *PLoS Biology*, 3(2):e65, 2005.

[27] P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *Proc. of the 14th International Joint Conference on Artificial Intelligence*, 1995.

[28] S. Rice, G. Nenadic, and B. Stapley. Mining protein functions from text using term-based support vector machines. *BMC Bioinformatics*, 6(Suppl 1):S22, 2005.

[29] H. Shatkay and R. Feldman. Mining the biomedical literature in the genomic era: An overview. *Journal of Computational Biology*, 10(6):821–855, 2003.

[30] R. Stevens, C. Wroe, P. Lord, and C. Goble. *Handbook on Ontologies*, chapter Ontologies in Bioinformatics. Springer, 2003.

[31] K. Verspoor, J. Cohn, C. Joslyn, S. Mniszewski, A. Rechtsteiner, L. Rocha, and T. Simas. Protein annotation as term categorization in the Gene Ontology using word proximity networks. *BMC Bioinformatics*, 6(Suppl 1):S20, 2005.