# Verification of Uncurated Protein Annotations

Francisco M Couto, Mário J Silva[1], Vivian Lee[2], Emily Dimmer[2], Evelyn Camon[2], Rolf Apweiler[2], Harald Kirsch[2], Dietrich Rebholz-Schuhmann[2]

(1)Departamento de Informática, Faculdade de Ciências, Universidade de Lisboa, Portugal
(2)European Bioinformatics Institute, Hinxton, Cambridge, UK

# Abstract

Molecular Biology research projects produced vast amounts of data, part of which has been preserved in a variety of public databases. However, a large portion of the data contains a significant number of errors and therefore requires careful verification by curators, a painful and costly task, before being reliable enough to derive valid conclusions from it. On the other hand, research in biomedical information retrieval and information extraction are nowadays delivering Text Mining solutions that can support curators to improve the efficiency of their work to deliver better data resources.

Over the past decades, automatic text processing systems have successfully exploited biomedical scientific literature to reduce the researchers' efforts to keep up to date, but many of these systems still rely on domain knowledge that is integrated manually leading to unnecessary overheads and restrictions in its use. A more efficient approach would acquire the domain knowledge automatically from publicly available biological sources, such as BioOntologies, rather than using manually inserted domain knowledge. An example of this approach is GOAnnotator, a tool that assists the verification of uncurated protein annotations. It provided correct evidence text at 93% precision to the curators and thus achieved promising results. GOAnnotator was implemented as a web tool that is freely available at http://xldb.di.fc.ul.pt/rebil/tools/goa/.

**Keywords:**

Information Retrieval, Information Extraction, BioOntologies, Text Mining, BioLiterature, Biomedical Annotation

# Introduction

A large portion of publicly available data provided in biomedical databases is still incomplete and incoherent (Devos and Valencia, 2001). This means that most of the data has to be handled with care and further validated by curators before we can use it to automatically draw valid conclusions from it. However, biomedical curators are overwhelmed by the amount of information that is published every day and are unable to verify all the data available. As a consequence, curators have verified only a small fraction of the available data. Moreover, this fraction tends to be even smaller given that the rate of data being produced is higher than the rate of data that curators are able to verify.

In this scenario, tools that could make the curators' task more efficient are much required. Biomedical information retrieval and extraction solutions are well established to provide support to curators by reducing the amount of information they have to seek manually. Such tools automatically identify evidence from the text that substantiates the data that curators need to verify. The evidence can, for example, be pieces of text published in BioLiterature (a shorter designation for the biological and biomedical scientific literature) describing experimental results supporting the data. As part of this process, it is not mandatory that the tools deliver high accuracy to be effective, since it is the task of the curators to verify the evidence given by the tool to ensure data quality. The main advantage of integrated text mining solutions lies in the fact that curators save time by filtering the retrieved evidence texts in comparison to scanning the full amount of available information. If the IT solution in addition provides the data in conjunction with the evidence supporting the data and if the solutions enable the curators to decide on their relevance and accuracy, it would surely make the task of curators more effective and efficient.

A real working scenario is given in the GOA (GO Annotation) project. The main objective of GOA is to provide high-quality GO (Gene Ontology) annotations to proteins that are kept in the UniProt Knowledgebase (Apweiler et al., 2004; Camon et al., 2004; GO-Consortium, 2004). Manual GO annotation produces high-quality and detailed GO term assignments (i.e. high granularity), but tends to be slow. As a result, currently less than 3% of UniProtKb has been confirmed by manual curation. For better coverage, the GOA team integrates uncurated GO annotations deduced from automatic mappings between UniProtKb and other manually curated databases (e.g. Enzyme Commission numbers or InterPro domains). Although these assignments have high accuracy, the GOA curators still have to verify them by extracting experimental results from peer-reviewed papers, which is time-consuming. This motivated the development of GOAnnotator, a tool for assisting the GO annotation of UniProtKb entries by linking the GO terms present in the uncurated annotations with evidence text automatically extracted from the documents linked to UniProtKb entries.

The remainder of this chapter starts by giving an overview on relevant research in biomedical information retrieval and extraction and exposing which contribution GOAnnotator brings to the current state-of-the-art. Afterwards, the chapter describes the main concepts of GOAnnotator and discusses its outcome and its main limitations, as well as proposals how the limitations can be solved in the future. Subsequently, the chapter provides insight into how approaches like GOAnnotator can change the way curators verify their data in the future, making the delivered data more complete and more reliable. For achieving this, the chapter will describe the issues that need to be addressed leading into valuable future research opportunities. The chapter ends by giving an overview of what was discussed and by presenting the concluding remarks.

## Background

A large amount of the information discovered in Molecular Biology has been mainly published in BioLiterature. However, analysing and identifying information in a large collection of unstructured texts is a painful and hard task, even to an expert.

## BioLiterature

The notion of BioLiterature includes any type of scientific text related to Molecular Biology. The text is mainly available in the following formats:

**Statement:** a short piece of text that is normally a remark or an evidence for a fact stored in a database.

**Abstract:** a short summary of a scientific document.

**Full-text:** the full-text of a scientific document including scattered text such as figure labels and footnotes.

Statements contain more syntactic and semantic errors than abstracts, since they are not peer-reviewed, but they are directly linked to the facts stored in the databases. The main advantage of using statements or abstracts is the brief and succinct format on which the information is expressed. However, usually this brief description is insufficient to draw a solid conclusion, since the authors have to skip some important details given the text size constraint. These details can only be found in the full-text of a document, which contains a complete description of the results obtained. For example, important details are sometimes only present in figure labels.

The full-text document contains the complete information for the presented research. Unfortunately, full-text documents are not yet readily available, since up to now publishers' licensing agreements restrict access to most of the full-text content. In addition, the formats and structures of the full-text document tend to vary according to the needs of the journal in where the document has been published leading to unnecessary complications in processing the documents. Furthermore, processing of full-text documents in comparison to document summaries also increases the complexity for text-mining solutions, leading to the result that the availability of more information is not necessarily all-beneficial to text-mining tools. Some of the information may even induce the production of novel errors, for example, the value of a fact reported in the Results section has to be interpreted differently in comparison to a fact that has been reported in the Related Work section. Therefore, the use of full-text will also create several problems regarding the quality of information extracted (Shah et al., 2004).

Access to BioLiterature is mainly achieved through the PubMed Web portal, which in 2008 delivered more than 17 million[1] references to biomedical documents dating from the 1950s (Wheeler et al., 2003). It is the main task of PubMed to make it easier for the general public to find scientific results in the BioLiterature. The users can search for citations by author name, journal title or keywords. PubMed also includes links to full-text documents and other related resources. More than 96% of the citations available through PubMed are from MEDLINE, a large repository of citations to the BioLiterature indexed with NLM's controlled vocabulary, the Medical Subject Headings (MeSH). Besides the bibliographic citations, PubMed also provides the abstracts of most documents, especially of the newer ones. The articles from 1950 through 1965 are in OLDMEDLINE, which contains approximately 1.7 million citations (Demsey et al., 2003). These old

[1] http://www.nlm.nih.gov/bsd/licensee/baselinestats.html

citations do not contain the abstract and certain fields may contain outdated or erroneous data.

MEDLINE was designed to deal with printed documents, but nowadays many journals provide the electronic version of their documents. Moreover, some of them became Open Access Publications, which means that their documents are freely available and can be processed and displayed without any restrictions. These documents have been exploited by tools, such as Google Scholar[2], Scirus[3] or EBSCO[4], which can be used to search and locate scientific documents. One of the major free digital archives of life sciences full-text documents is PMC (PubMed Central), which aims at preserving and maintaining access to this new generation of electronic documents. Presently, PMC includes over 1.3 million documents. The availability of full-text documents offers new opportunities for research to text-mining tool providers, who were up to now often restricted to analysing only the abstracts of scientific documents.

## Text Mining

An approach to improve the access to the knowledge published in BioLiterature is to use Text Mining, which aims at automatically retrieving and extracting knowledge from natural language text (Hearst, 1999). The application of text-mining tools to BioLiterature started just a few years ago (Andrade and Bork, 2000). 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 (Couto and Silva, 2006). 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.

One of the most important applications 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

[2] http://scholar.google.com/
[3] http://www.scirus.com/
[4] http://www.epnet.com/

description of its biological role. The biological role is often a concept from a BioOntology, which organises and describes biological concepts and their relationships. Using a BioOntology to annotate genes or proteins avoids ambiguous statements that are domain specific and context dependent. The best-known example is the gene ontology (GO) that is a well-established structured vocabulary that has been successfully designed and applied for gene annotation of different species (GO-Consortium, 2004). 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.

Very early on the text-mining system AbXtract was developed to identify keywords from MEDLINE abstracts and to score their relevance for a protein family (Andrade and Valencia, 1998). Other systems have been developed in recent years to identify GO terms from the text: MeKE identified potential GO terms based on sequence alignment (Chiang and Yu, 2003) and BioIE uses syntactic dependencies to select GO terms from the text (Kim and Park, 2004). Furthermore, other approaches use IT solutions where GO terminology is applied as a dictionary (Koike et al., 2005; Müller et al., 2004; Pérez et al., 2004; Rebholz-Schuhmann et al., 2007). However, none of these systems have been integrated into the GOA curation process. Moreover, only Perez et al. make use of the hierarchical structure of GO to measure the distance between two terms based on the number of edges that separate them (path length). However, incorrect annotations can be caused by neglecting the semantics of the hierarchical structure of GO causes. For example, if a large number of GO terms from the leaves or the deep levels in GO are assigned then the system tends to generate over-predictions, and if general GO terms from the top levels of the hierarchy are produced then the annotations tend to be useless because they are not meaningful.

The performance of state-of-the-art text-mining tools for automatic annotation of genes or proteins is still not acceptable by curators, since gene or protein annotation is more subjective and requires more expertise than simply finding relevant documents and recognising biological entities in texts. To improve their performance, state-of-the-art text-mining tools use domain knowledge manually inserted by curators (Yeh et al., 2003). This knowledge consists of rules inferred from patterns identified in the text, or on predefined sets of previously annotated texts. The integration of domain knowledge improves overall the precision of predictions, but it cannot be easily extended to work on other domains and demands an extra effort to keep the knowledge updated as BioLiterature evolves.

The selection of pieces of text that mention a GO term was assessed as part of the first BioCreAtIvE competition (Hirschman et al., 2005). This competition enabled the assessment of different text mining approaches and their ability to assist curators. The system with the best precision predicted 41 annotations, but 27 were not correct, which lead to a 35% precision (14 out of 41) (Chiang and Yu, 2004). The main problem is that the terms denoting GO concepts were never designed to support text-mining solutions. Terms in the vocabulary are ambiguous and could not be easily deciphered by automatic processing and sometimes even by humans (Camon et al., 2005). Without improvements to the precision, such automatic extractions are unhelpful to curators. This reflects the importance of designing more efficient tools to aid in the curation effort.

GOAnnotator uses publicly available biological data sources as domain knowledge for improving the retrieval and extraction tasks requiring minimal human intervention, since it 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 (Shatkay and Feldman, 2003). Apart from avoiding direct human intervention, automatically collected domain knowledge is usually more extensive than manually generated domain knowledge and does not become outdated as easily, if the originating public databases can be automatically tracked for updates as they evolve. The most important data resource used by GOAnnotator is GO.

## Gene Ontology (GO)

The GO project is one of the long-lasting and successful resource building efforts in Molecular Biology constructing a BioOntology of broad scope and wide applicability (Bada et al., 2004). GO provides a structured controlled vocabulary denoting the diversity of biological roles of genes and proteins in a species-independent way (GO-Consortium, 2004). GO comprised 24,397 distinct terms in September 2007. Since the activity or function of a protein can be defined at different levels, GO is composed of 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.

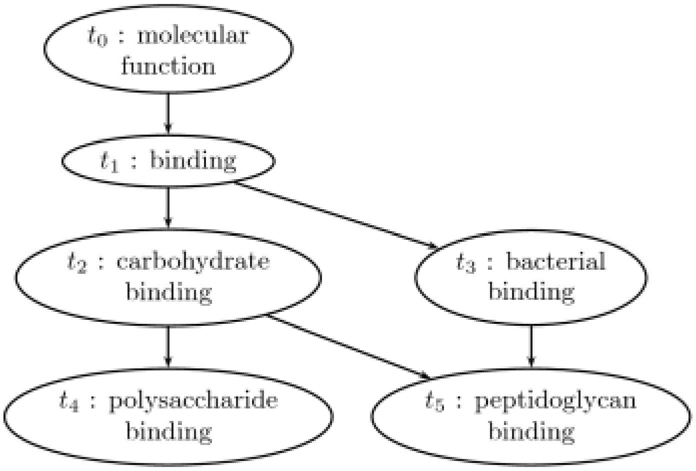
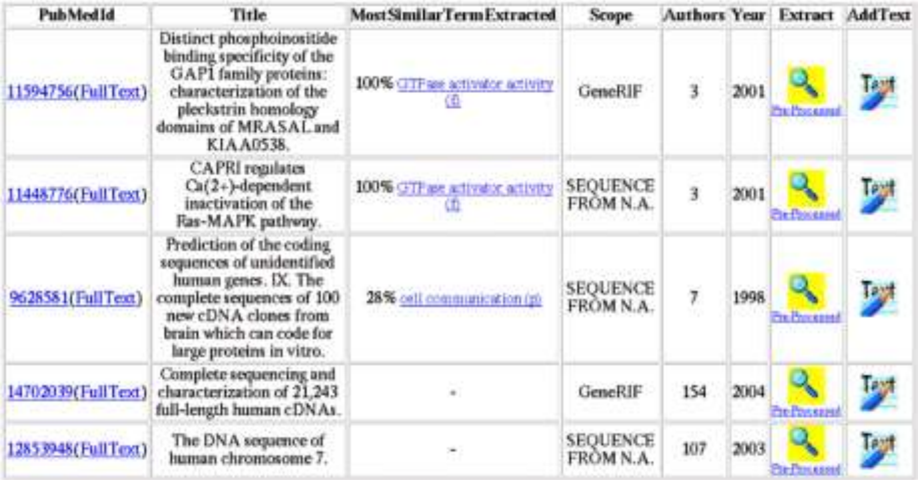

$t_0$ : molecular function
$t_1$ : binding
$t_2$ : carbohydrate binding
$t_3$ : bacterial binding
$t_4$ : polysaccharide binding
$t_5$ : peptidoglycan binding

Figure 1: Sub-graph of GO

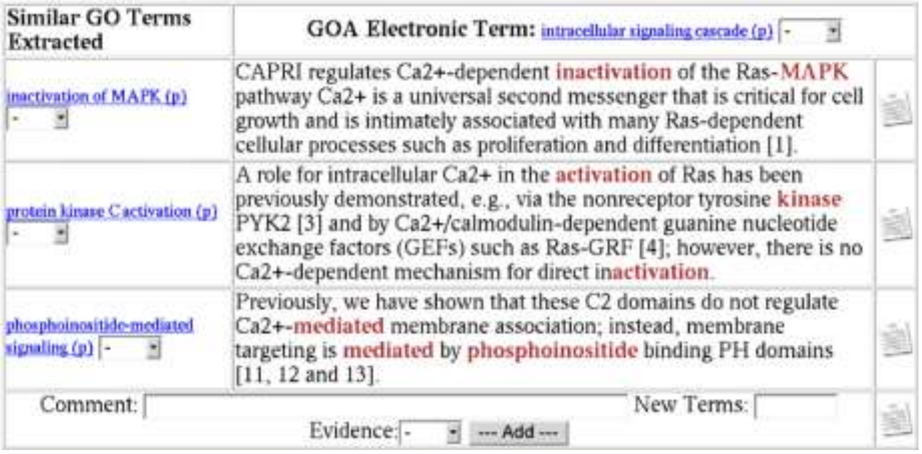
GO organises the concepts as a DAG (Directed Acyclic Graph), one for each aspect. Each node of the graph represents a concept, and the edges represent the links between concepts (see example in Figure 1). Links can represent two relationship types: *is-a* and *part-of.* The content of GO is still evolving dynamically: 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.

# GOAnnotator



| PubMedId | Title | MostSimilarTermExtracted | Scope | Authors | Year | Extract | AddText |
|---|---|---|---|---|---|---|---|
| 11594756(FullText) | Distinct phosphoinositide binding specificity of the GAP1 family proteins: characterization of the pleckstrin homology domains of MRASAL and KIAA0538. | 100% GTPase activator activity (f) | GeneRIF | 3 | 2001 | PreProcessed | Text |
| 11448776(FullText) | CAPRI regulates Ca(2+)-dependent inactivation of the Ras-MAPK pathway. | 100% GTPase activator activity (f) | SEQUENCE FROM N.A. | 3 | 2001 | PreProcessed | Text |
| 9628581(FullText) | Prediction of the coding sequences of unidentified human genes. IX. The complete sequences of 100 new cDNA clones from brain which can code for large proteins in vitro. | 28% cell communication (p) | SEQUENCE FROM N.A. | 7 | 1998 | PreProcessed | Text |
| 14702039(FullText) | Complete sequencing and characterization of 21,243 full-length human cDNAs. | - | GeneRIF | 154 | 2004 | PreProcessed | Text |
| 12853948(FullText) | The DNA sequence of human chromosome 7. | - | SEQUENCE FROM N.A. | 107 | 2003 | PreProcessed | Text |

Figure 2: List of documents related with a given protein. The list is sorted by the most similar term extracted from each document. The curator can use the *Extract* option to see the extracted terms together with the evidence text. By default GOAnnotator uses only the abstract, but the curator can use the *AddText* option to replace or insert text.



| Similar GO Terms Extracted | GOA Electronic Term: intracellular signaling cascade (p) - | |
|---|---|---|
| inactivation of MAPK (p)<br>- | CAPRI regulates Ca2+-dependent inactivation of the Ras-MAPK pathway Ca2+ is a universal second messenger that is critical for cell growth and is intimately associated with many Ras-dependent cellular processes such as proliferation and differentiation [1]. | |
| protein kinase C activation (p)<br>- | A role for intracellular Ca2+ in the activation of Ras has been previously demonstrated, e.g., via the nonreceptor tyrosine kinase PYK2 [3] and by Ca2+/calmodulin-dependent guanine nucleotide exchange factors (GEFs) such as Ras-GRF [4]; however, there is no Ca2+-dependent mechanism for direct inactivation. | |
| phosphoinositide-mediated signaling (p) - | Previously, we have shown that these C2 domains do not regulate Ca2+-mediated membrane association; instead, membrane targeting is mediated by phosphoinositide binding PH domains [11, 12 and 13]. | |
| Comment: | New Terms:<br>Evidence: - --- Add --- | |

Figure 3: GO terms extracted. For each uncurated annotation, GOAnnotator shows the similar GO terms extracted from a sentence of the selected document. If any of the sentences provides correct evidence for the uncurated annotation, or if the evidence supports a GO term similar to that present in the uncurated annotation, the curator can use the *Add* option to store the annotation together with the document reference, the evidence codes and any comments.

GOAnnotator is a tool for assisting the GO annotation of UniProtKb entries by linking the GO terms present in the uncurated annotations with evidence text automatically extracted from the documents linked to UniProtKb entries. Initially, the curator provides a UniProtKb accession number to GOAnnotator. GOAnnotator follows the bibliographic links found in the UniProtKb database and retrieves the documents. Additional documents are retrieved from the GeneRIF database or curators can add any other text (Mitchell et al., 2003). GOAnnotator prioritizes the documents according to the extracted GO terms from the text and their similarity to the GO terms present in the protein uncurated annotations (see Figure 2). Any extracted GO term is an indication for the topic of the document, which is also taken from the UniProtKb entry. The curator uses the topic as a hint to potential GO annotation.

The extraction of GO terms is based on FiGO, a method used for the BioCreAtIvE competition (Couto et al., 2005). FiGO receives a piece of text and returns the GO terms that were detected in the given text. To each selected GO term, FiGO assigns a confidence value that represents the terms' likelihood of being mentioned in the text. The confidence value is the ratio of two parameters. The first parameter is called local evidence context and is used to measure the likelihood that words in the text are part of a given GO term. The second parameter is a correction parameter, which increases the confidence value when the words detected in the text are infrequent in GO. In BioCreAtIvE, FiGO predicted 673 annotations but 615 were not correct, which lead to a 8.6% precision (58 of 673).

GO terms are considered to be similar if they are in the same lineage or if they share a common parent in the GO hierarchy. To calculate a similarity value between two GO terms, we decided to implement a semantic similarity measure. Research on Information Theory proposed many semantic similarity measures. Some of them calculate maximum likelihood estimates for each concept using the corpora, and then calculate the similarity between probability distributions. Semantic similarity measures take into consideration a combination of parameters linked to the structure of an ontology as well as information content based on statistical data from corpora (Rada et al., 1989). The information content of a concept is inversely proportional to its frequency in the corpora. Concepts that are frequent in the corpora have low information content. In case of GO the corpora used to derive the statistical information is the annotations provided by GO, i.e. the information content of a GO term is calculated based on the number of proteins annotated to it. For example, GO terms annotated to most of the proteins normally provide little semantic information.

Many semantic similarity measures applied to ontologies have been developed. We implemented a measure based on the ratio between the information content of the most informative common ancestor and the information content of both concepts (Lin, 1998). Recent studies have explored on the effectiveness of semantic similarity measures over the GO (Couto et al., 2006; Lord et al., 2003; Gaudan et al., 2008). The results have shown that GO similarity is correlated with sequence and family similarity, i.e., they demonstrated the feasibility of using semantic similarity measures in a biological setting.

GOAnnotator displays a table for each uncurated annotation with the GO terms that were extracted from a document and were similar to the GO term present in the uncurated annotation (see Figure 3). The sentences from which the GO terms were extracted are also displayed. Words that have contributed to the extraction of the GO terms are highlighted. GOAnnotator gives the curators the opportunity to manipulate the confidence and similarity thresholds to modify the number of predictions.

| GO Aspect | GO Terms |
|---|---|
| molecular function | 54 |
| biological process | 18 |
| cellular component | 6 |
| total | 78 |

Table 1: Distribution of the GO terms from the selected uncurated annotations through the different aspects of GO.

| Evidence Evaluation | Extracted Annotations |
|---|---|
| correct | 83 |
| incorrect | 6 |
| total | 89 |

Table 2: Evaluation of the evidence text substantiating uncurated annotations provided by the GOAnnotator.

| GO Terms | Extracted Annotations |
|---|---|
| exact | 65 |
| same lineage | 15 |
| different lineage | 3 |
| total | 83 |

Table 3: Comparison between the extracted GO terms with correct evidence text and the GO terms from the uncurated annotations.

# Results

The GOA team agreed to curate about 3% proteins from a list of 1,953 uncurated UniProtKb/SwissProt proteins. As a consequence the similarity and confidence thresholds of GOAnnotator were adapted until GOAnnotator generated this percentage of predictions. 66 proteins were selected at the similarity threshold of 40% and confidence threshold of 50%. In other words, GOAnnotator identified evidence texts with at least 40% similarity and 50% confidence to selected GO terms for all 66 proteins. For 80 uncurated annotations to these proteins, GOAnnotator extracted 89 similar annotations and their evidence text from 118 MEDLINE abstracts. The 80 uncurated annotations included 78 terms from different domains of GO (see Table 1). After analyzing the 89 evidence texts, GOA curators found that 83 were valid to substantiate 77 distinct uncurated annotations (see Table 2), i.e. 93% precision.

Table 3 shows that 78% (65 out of 83) of the correct evidence texts confirmed the uncurated annotations, i.e. the extracted annotation and the uncurated annotation contained the same GO identifier. In cases where the evidence text was correct, it did not always contain exactly any of the known variations of the extracted GO term. In the other cases the extracted GO term was similar: in 15 cases the extracted GO term was in the same lineage of the GO term in the uncurated annotation; in 3 cases the extracted GO term was in a different lineage, but both terms were similar (share a parent). In general, we can expect GOAnnotator to confirm the uncurated annotation using the findings from the scientific literature, but it is also obvious that GOAnnotator can propose new GO terms.

## Examples

GOAnnotator provided correct evidence for the uncurated annotation of the protein “Human Complement factor B precursor” (P00751) with the term “complement activation, alternative pathway” (GO:0006957). The evidence is the following sentence from the document with the PubMed identifier 8225386: “The human complement factor B is a centrally important component of the alternative pathway activation of the complement system.”

GOAnnotator provided a correct evidence for the uncurated annotation of the protein “U4/U6 small nuclear ribonucleoprotein Prp3” (O43395) with the term “nuclear mRNA splicing, via spliceosome” (GO:0000398). From the evidence the tool extracted the child term “regulation of nuclear mRNA splicing, via spliceosome” (GO:0048024). The evidence is the following sentence from the document with the PubMed identifier 9328476: “Nuclear RNA splicing occurs in an RNA-protein complex, termed the spliceosome.” However, this sentence does not provide enough evidence on its own, the curator had to analyze other parts of the document to draw a conclusion.

GOAnnotator provided a correct evidence for the uncurated annotation of the protein “Agmatinase” (Q9BSE5) with the term “agmatinase activity” (GO:0008783). From the evidence the tool extracted the term “arginase activity” (GO:0004053) that shares a common parent. The evidence was provided by the following sentence from the document with the PubMed identifier 11804860: “Residues required for binding of Mn(2+) at the active site in bacterial agmatinase and other members of the arginase superfamily are fully conserved in human agmatinase.” However, the annotation only received a NAS (Non-traceable author statement) evidence code, as the sentence does not provide direct experimental evidence of arginase activity. Papers containing direct experimental evidence for the function/subcellular location of a protein are more valuable to GO curators.

GOAnnotator provided a correct evidence for the uncurated annotation of the protein “3’-5’ exonuclease ERI1” (Q8IV48) with the term “exonuclease activity” (GO:0004527). The evidence is the following sentence from the document with the PubMed identifier 14536070: “Using RNA affinity purification, we identified a second protein, designated 3’hExo, which contains a SAP and a 3’ exonuclease domain and binds the same sequence.” However, the term “exonuclease activity” is too high level, and a more precise annotation should be “3’-5’ exonuclease activity” (GO:0008408).

## Discussion

Researchers need more than facts, they need the source from which the facts derive (Rebholz-Schuhmann et al., 2005). GOAnnotator provides not only facts but also their evidence, since it links existing annotations to scientific literature. GOAnnotator uses text-mining methods to extract GO terms from scientific papers and provides this information together with a GO term from an uncurated annotation. In general, we can expect GOAnnotator to confirm the uncurated annotation using the findings from the scientific literature, but it is obvious as well that GOAnnotator can propose new GO terms. In both cases, the curator profits from the integration of both approaches into a single interface. By comparing both results, the curator gets convenient support to take a decision for a curation item based on the evidence from the different data resources.

GOAnnotator provided correct evidence text at 93% precision, and in 78% of these cases the GO term present in the uncurated annotation was confirmed. These results were obtained for a small subset of the total number of uncurated annotations, but it represents already a significant set for curators. Notice that manual GO annotation covers less than 3% of UniProtKb. Over time, proteins tend to be annotated with more accurate uncurated terms and bibliography. Thus, the percentage of uncurated proteins satisfying the 40% similarity and 50% confidence thresholds will grow, and therefore make GOAnnotator even more effective.

Sometimes, the displayed sentence from the abstract of a document did not contain enough information for the curators to evaluate an evidence text with sufficient confidence. Apart from the association between a protein and a GO term, the curator needs additional information, such as the type of experiments performed and the species from which the protein originates. Unfortunately, quite often this information is only available in the full text of the scientific publication. GOAnnotator can automatically retrieve the abstracts, but in the case of the full text the curator has to copy and paste the text into the GOAnnotator interface, which only works for a limited number of documents. BioRAT solves this problem by retrieving full text documents from the Internet (Corney et al., 2004). In addition, the list of documents cited in the UniProtKb database was not sufficient for the curation process. In most cases, the curators found additional sources of information in PubMed. In the future, GOAnnotator should be able to automatically query PubMed using the protein's names to provide a more complete list of documents.

GOAnnotator ensures high accuracy, since all GO terms that did not have similar GO terms in the uncurated annotations were rejected. Using this 40% similarity threshold may filter out meaningful potential annotations that are not similar to known curated annotations. However, without this restriction the results returned by the text mining method would contain too much noise to be of any use to curators, as it was demonstrated in the BioCreAtIvE competition. GOAnnotator meets the GOA team's need for tools with high precision in preference to those with high recall, and explains the strong restriction for the similarity of two GO terms: only those that were from the same lineage or had a shared parent were accepted. Thus, GOAnnotator not only predicted the exact uncurated annotation but also more specific GO annotations, which was of strong interest to the curators. MeKE selected a significant number of general terms from the GO hierarchy (Chiang and Yu, 2003). Others distinguished between gene and family names to deal with general terms (Koike et al., 2005). GOAnnotator takes advantage of uncurated annotations to avoid general terms by extracting only similar terms, i.e. popular proteins tend to be annotated to specific terms and therefore GOAnnotator will also extract specific annotations to them.

The applied text-mining method FiGO was designed for recognizing terms and not for extracting annotations, i.e. sometimes the GO term is correctly extracted but is irrelevant to the actual protein of interest. The method also generated mispredictions in the instances where all the words of a GO term appeared in disparate locations of a sentence or in an unfortunate order. Improvements can result from the incorporation of better syntactical analysis into the identification of GO terms similar to the techniques used by BioIE (Kim and Park, 2004). For example, a reduction of the window size of FiGO or the identification of noun phrases can further increase precision. In the future, GOAnnotator can also use other type of text-mining methods that prove to be more efficient for extracting annotations.

## Future Trends

Recent publications on the improvements by using information retrieval and extraction tools are promising and encourage the research community to make an effort to improve their quality and expand their scope. However, the performance of most tools is still highly dependent on domain knowledge provided through experts. Integration of the expert knowledge is time-consuming and imposes limitations whenever services have to be extended to other domains with different user requirements. On the other side, the domain of molecular biology draws profits from publicly available databases containing a significant amount of information. In our opinion, better use of such domain knowledge and automatic integration of the data from these biological information resources will be the key to develop more efficient tools and will thus contribute to their wider acceptance among curators in the biological domain. Apart from avoiding direct human intervention, automatic collection of domain relevant information is usually more comprehensive than any manually generated representation of domain knowledge and does not become outdated, since public databases can be automatically tracked for updates as they evolve.

Domain knowledge is only available thanks to the research community efforts in developing accurate and valuable data resources and by making them publicly available. These data resources are continually being updated with more information. However, they are still too incomplete, too inconsistent and/or too morpho-syntactically inflexible to efficiently be used by automatic tools. For example, 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 like EC numbers. However, as different research communities understand the importance of adding their domain knowledge to GO, it will expand its coverage and improve its interoperability with other data sources. While BioOntologies are traditionally used mainly for annotation purposes, their ultimate goal should be to accurately represent the domain knowledge so as to allow automated reasoning and support knowledge extraction. The establishment of guiding principles, as in OBO, to guide the development of new BioOntologies is a step in this direction, by promoting formality, enforcing orthogonality, and proposing a common syntax that facilitates mapping between BioOntologies. This not only improves the quality of individual BioOntologies, but also enables a more effective use of them by information retrieval and extraction tools.

## Conclusions

This document introduced the biomedical information retrieval and extraction research topics and how their solutions can help curators to improve the effectiveness and efficiency of their tasks. It gives an overview on three important aspects of these research topics: BioLiterature, Text Mining and BioOntologies.

The document presented GOAnnotator, a system that automatically identifies evidence text in literature for GO annotation of UniProtKb/SwissProt proteins. GOAnnotator provided evidence text at high precision (93%, 66 sample proteins) taking advantage of existing uncurated annotations and the GO hierarchy. GOAnnotator assists the curation process by allowing fast verification of uncurated annotations from evidence texts, which can also be the source for novel annotations. This document discusses the results obtained by GOAnnotator pointing out its main limitations. This document ends by providing insight into the issues that need to be addressed in this research area and represent good future research opportunities.

The approach presented in this document constitutes a small and relatively early contribution to the advance of biomedical information retrieval and extraction topics, but the main idea presented here seems promising and the results encourage further study. Still, despite all the limitations presented here, many relevant biological discoveries in the future will certainly result from an efficient exploitation of the existing and newly generated data by tools like GOAnnotator.

## Acknowledgements

This work was supported by the Marie Curie Training Sites scheme of the European Commission's Quality of Life Programme (Contract no. QLRI-1999-50595) and by FCT through the Multiannual Funding Programme.

## Bibliography

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

Andrade, M. and Valencia, A. (1998). Automatic extraction of keywords from scientific text: Application to the knowledge domain of protein families. *Bioinformatics*, 14(7):600–607.

Apweiler, R., Bairoch, A., Wu, C., Barker, W., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M., Natale, D., O'Donovan, C., Redaschi, N., and Yeh, L. (2004).

UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, 32(Database issue):D115–D119.

Bada, M., Stevens, R., Goble, C., Gil, Y., Ashburner, M., Blake, J., Cherry, J., Harris, M., and Lewis, S. (2004). A short study on the success of the gene ontology. *Journal of Web Semantics*, 1(1):235–240.

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

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

Chiang, J. and Yu, H. (2003). MeKE: discovering the functions of gene products from biomedical literature via sentence alignment. *Bioinformatics*, 19(11):1417–1422.

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

Corney, D., Buxton, B., Langdon, W., and Jones, D. (2004). BioRAT: Extracting biological information from full-length papers. *Bioinformatics*, 20(17):3206–3213.

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

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

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

Demsey, A., Nahin, A., and Braunsberg, S. V. (2003). Oldmedline citations join pubmed. *NLM Technical Bulletin*, 334(e2).

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

Gaudan, S., Jimeno, A., Lee, V., and Rebholz-Schuhmann, D. (2008) Combining evidence, specificity and proximity towards the normalization of Gene Ontology terms in text. EURASIP JBSB (accepted)

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

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

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

Kim, J. and Park, J. (2004). BioIE: retargetable information extraction and ontological annotation of biological interactions from literature. *Journal of Bioinformatics and Computational Biology*, 2(3):551–568.

Koike, A., Niwa, Y., and Takagi, T. (2005). Automatic extraction of gene/protein biological functions from biomedical text. *Bioinformatics*, 21(7):1227–1236.

Lin, D. (1998). An information-theoretic definition of similarity. In *Proc. of the 15th International Conference on Machine Learning*.

Lord, P., Stevens, R., Brass, A., and Goble, C. (2003). Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics*, 19(10):1275–1283.

Mitchell, J., Aronson, A., Mork, J., Folk, L., Humphrey, S., and Ward, J. (2003). Gene indexing: characterization and analysis of NLM's GeneRIFs. In *Proc. of the AMIA 2003 Annual Symposium*.

Müller, H., Kenny, E., and Sternberg, P. (2004). Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLOS Biology*, 2(11):E309.

Pérez, A., Perez-Iratxeta, C., Bork, P., Thode, G., and Andrade, M. (2004). Gene annotation from scientific literature using mappings between keyword systems. *Bioinformatics*, 20(13):2084–2091.

Rada, R., Mili, H., Bicknell, E., and Blettner, M. (1989). Development and application of a metric on semantic nets. *IEEE Transactions on Systems*, 19(1):17–30.

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

Rebholz-Schuhmann, D., Kirsch, H., Gaudan, M. A. S., Rynbeek, M., and Stoehr, P. (2007). EBIMed - text crunching to gather facts for proteins from Medline. *Bioinformatics*, 23(2):e237–44.

Shah, P., Perez-Iratxeta, C., Bork, P., and Andrade, M. (2004). Information extraction from full text scientific articles: Where are the keywords? *BMC Bioinformatics*, 4(20).

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

Wheeler, D., Church, D., Federhen, S., Lash, A., Madden, T., Pontius, J., Schuler, G., Schriml, L., Sequeira, E., Tatusova, T., and Wagner, L. (2003). Database resources of the national center for biotechnology. *Nucleic Acids Research*, 31(1):28–33.

Yeh, A., Hirschman, L., and Morgan, A. (2003). Evaluation of text data mining for database curation: Lessons learned from the KDD challenge cup. *Bioinformatics*, 19(1):i331–i339.