

**GOAnnotator: linking protein GO
annotations to evidence text**

Francisco M. Couto

Mário J. Silva

Vivian Lee

Emily Dimmer

Evelyn Camon

Rolf Apweiler

Harald Kirsch

Dietrich Rebholz-Schuhmann

DI-FCUL

TR-05-25

December 2005

Departamento de Informática
Faculdade de Ciências da Universidade de Lisboa
Campo Grande, 1749-016 Lisboa
Portugal

Technical reports are available at <http://www.di.fc.ul.pt/tech-reports>. The files are stored in PDF, with the report number as filename. Alternatively, reports are available by post from the above address.

GOAnnotator: linking protein GO annotations to evidence text

Francisco M. Couto ^{*†} Mário J. Silva [†] Vivian Lee [‡]
Emily Dimmer [‡] Evelyn Camon [‡] Rolf Apweiler [‡]
Harald Kirsch [‡]
Dietrich Rebholz-Schuhmann [†]

*Contact author: fcouto@di.fc.ul.pt

[†]Departamento de Informática, Faculdade de Ciências,
Universidade de Lisboa, Portugal

[‡] European Bioinformatics Institute, Hinxton, Cambridge, UK

December 2005

Abstract

Annotation of proteins with gene ontology (GO) terms is ongoing work and a complex task. Manual GO annotation is precise and precious, but it is time-consuming. Therefore, instead of curated annotations most of the proteins come with uncurated annotations, which have been generated automatically.

Text-mining systems that use literature for automatic annotation have been proposed but they do not satisfy the high quality expectations of curators. In this paper we describe an approach that links uncurated annotations to text extracted from literature. The selection of the text is based on the similarity of the text to the term from the uncurated annotation. Besides substantiating the uncurated annotations, the extracted texts also lead to novel annotations. In addition, the approach uses the GO hierarchy to achieve high precision.

Our approach is integrated into GOAnnotator, a tool that assists the curation process for GO annotation of UniProt proteins. The GO curators assessed GOAnnotator with a set of 66 distinct UniProt/SwissProt proteins with uncurated annotations. GOAnnotator provided correct evidence text at 93% precision. This high precision results from using the GO hierarchy to only select GO terms similar to GO terms from uncurated annotations in GOA. Our approach is the first one, which achieved high precision, which is crucial for the efficient support of GO curators.

GOAnnotator is available at: <http://xldb.fc.ul.pt/rebil/tools/goa/>

Keywords: Bioinformatics (genome or protein) databases, Text mining

1 Introduction

The core objective of GOA (GO Annotation) is to provide high-quality GO (Gene Ontology) annotations to proteins within the UniProt Knowledgebase [2, 6, 1]. Manual GO annotation produces high-quality and granular GO term assignments, but tends to be slow and therefore covers less than 3% of UniProt. For better coverage, the GOA team integrates uncurated GO annotations deduced from automatic mappings between UniProt and other manually curated databases (e.g. Enzyme Commission numbers or InterPro domains). Although these assignments have high accuracy, the GOA team still has to verify them by extracting experimental results from peer-reviewed papers.

Reading these papers takes time, which motivates the research of text-mining methods. Very early on Andrade et al. proposed the text-mining system AbXtract, which identifies keywords from MEDLINE abstracts and scores their relevance for a protein family. Other systems have been developed in recent years to identify GO terms from the text: MeKE by Chiang et al. identified potential GO terms based on sequence alignment and Kim et al. created BioIE which uses syntactic dependencies to select GO terms from the text [3, 8]. Furthermore, Perez et al., Müller et al. (Textpresso) and Koike et al. suggested IT solutions where GO terminology is applied as a dictionary [9, 13, 12]. However, none of these systems have been integrated into the GOA curation process. Moreover, only Perez makes use of the topology of the hierarchical structure of GO to measure the distance between two terms based on the number of edges that separate them. Neglecting the semantic of the hierarchical structure of GO causes incorrect annotations by over-predicting too deep-level GO terms, or useless annotations by predicting too general GO terms.

The selection of pieces of text that mention a GO term was assessed as part of the BioCreAtIvE competition [7]. 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 of 41). Without improvements to the precision, such automatic extractions are unhelpful to curators. This reflects the importance of integrating domain knowledge when designing tools to aid in the curation effort.

When manually annotating, GOA curators use pre-existing uncurated annotations as a guide, which can also be used to direct text-mining tools. Since GOA curators primarily require high precision in a text-mining solution, we expect that the information from the uncurated annotations will support this goal without the complex issues of creating rules and patterns encompassing all possible cases, and creating training sets that are too specific to be extended to new domains.

Section 2 describes the GOAnnotator, and Section 3 describes the methods incorporated. The assessment of GOAnnotator is presented in Section 4. In Section 5, we discuss the results. Finally, Section 6 expresses our main conclusions.

2 GOAnnotator

GOAnnotator is a tool for assisting the GO annotation of UniProt entries by linking the GO terms present in the uncurated annotations with evidence text





















| PubMedId | Title | MostSimilar Term Extracted | Scope | Authors | Year | Extract | Add Text |
|------------------------------------|--|--|--------------------|---------|------|---|---|
| 11594756(FullText) | Distinct phosphoinositide binding specificity of the GAF1 family proteins; characterization of the pleckstrin homology domains of MRASAL and KIAA0538. | 100% GTPase activator activity (f) | GeneRIF | 3 | 2001 |   |   |
| 11448776(FullText) | CAPRI regulates Ca(2+)-dependent inactivation of the Ras-MAPK pathway. | 100% GTPase activator activity (f) | SEQUENCE FROM N.A. | 3 | 2001 |   |   |
| 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 |   |   |
| 14702038(FullText) | Complete sequencing and characterization of 21,243 full-length human cDNAs. | - | GeneRIF | 154 | 2004 |   |   |
| 12853948(FullText) | The DNA sequence of human chromosome 7. | - | SEQUENCE FROM N.A. | 107 | 2003 |   |   |

Figure 1: 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) [-] | 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) [-] | 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: <input type="text"/> | New Terms: <input type="text"/> Evidence: [-] <input type="button" value="Add"/> |

Figure 2: 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.

automatically extracted from the documents linked to UniProt entries. Initially, the curator provides a UniProt accession number to GOAnnotator. GOAnnotator follows the bibliographic links found in the UniProt database and retrieves the documents. Additional documents are retrieved from the GeneRIF database or curators can add any other text [11]. 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 1). Any extracted GO term is an indication for the topic of the document, which is also taken from the UniProt 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-AtIvE competition (see Section 3) [5]. GO terms are similar if they are in the same lineage or if they share a common parent in the GO hierarchy. A semantic similarity measure is used to determine the degree of similarity between two GO terms (see Section 3).

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 2). 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.

3 Methods

The two main components of GOAnnotator comprise the method to extract GO terms from text, and the similarity measure between GO terms.

3.1 FiGO

FiGO receives a piece of text and returns the GO terms that were detected in the given text. To each 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 (LEC) 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.

FiGO starts by analyzing all entries of the ontology and calculating what we call the *evidence content* (EC) of those words that are part of at least one entry. This evidence content is inversely correlated to the frequency of a given word in the ontology and it is supposed to measure its amount of evidence needed to detect an entry in the text. For instance, consider the word "binding" that is used in many GO terms. If this word is encountered in the text, there is a low probability that the text mentions to the GO term "pant binding". Instead, if the word 'pant' is also encountered, then we have strong evidence that the GO term is mentioned in the text, since 'pant' is not part of any other GO term.

FiGO receives i) an ontology *Ont* and ii) a piece of text, *Txt*, as input. FiGO returns the entries in *Ont* that were detected in the given text. These

entries are ranked according to their likelihood of being mentioned in the text. For example, *Ont* can be the GO with each entry of *Ont* representing one GO term, and *Txt* can be a sentence taken from a document.

The Words

FiGO derives a map between the entries and their names:

$$Names(e) = \{n_0, \dots, n_k\},$$

where $e \in Ont$ and n_0, \dots, n_k are its name and synonyms in the ontology. If e does not have synonyms, then $k = 0$ and $Names(e) = \{n_0\}$. The set of words that compose a name n is given by:

$$Words(n) = \{w_0, \dots, w_l\}.$$

In addition, we define the set of words contained in an entry e as:

$$Words(e) = \{w \in Words(n) | n \in Names(e)\}$$

Furthermore, the words of the ontology are

$$Words(Ont) = \{w \in Words(e) | e \in Ont\}$$

Evidence Content

The evidence content of each word decreases with its frequency. The frequency of a word w is the number of entries that contain the word:

$$Freq(w) = \#\{e \in Ont | w \in Words(e)\}.$$

A word present in only one name has high evidence content. On the other hand, the word with the maximum frequency has no evidence content. The maximum frequency is defined using the following equation:

$$MaxFreq = \max\{Freq(w) | w \in Words(Ont)\}.$$

Thus, $WordEC(w)$, the evidence content of a word w , is defined using the following equation:

$$WordEC(w) = -\log_2\left(\frac{Freq(w)}{MaxFreq}\right).$$

Since each name is composed of a set of words, we can define the evidence content of a name n as the sum of the evidence content of its words:

$$NameEC(n) = \sum_{w \in Words(n)} WordEC(w)$$

The evidence content of an entry e is defined as the highest evidence content of all its names:

$$EC(e) = \max\{NameEC(n) | n \in Names(e)\}.$$

Local Evidence Content

The input text is transformed into a set of words:

$$Txt = \{w_0, \dots, w_l\}.$$

The local evidence content (LEC) is used to measure the likelihood that a given name n is mentioned in the text Txt . LEC is the sum of the evidence content of those words, which are present in the text as well as in the name:

$$NameLEC(n, Txt) = \sum_{w \in (Txt \cap Words(n))} WordEC(w).$$

The LEC is also used to measure the likelihood that a given entry e is mentioned in the text Txt :

$$LEC(e, Txt) = \max_{n \in Names(e)} \{NameLEC(n, Txt)\}.$$

The LEC divided by the EC is a confidence level for the entry e occurring in the Txt :

$$Conf(e, Txt) = \frac{LEC(e, Txt)}{EC(e)}.$$

$Conf(e, Txt) \in [0, 1]$, since LEC is smaller than EC by definition.

If the confidence level is larger than a given threshold $\alpha \in [0, 1]$, then e is considered to occur in Txt :

$$Conf(e, Txt) \geq \alpha.$$

If $\alpha = 1$, the complete name has to appear in the text to be selected. Thus, the α parameter is used to tune recall and precision of FiGO. An increase in α increases precision, a decrease in α increases recall. $Conf(e, Txt)$ is used to rank the returned entries, and represents the likelihood of each entry of the ontology occurring in text.

3.2 Similarity Measure

To calculate the similarity between two GO terms, we decided to implement the Lin's semantic similarity measure [10]. This measure combines the structure and content of an ontology with statistical information from corpora, and it is based on the notion of commonality and differences between the information content of two concepts. The information content of a concept c is defined as the negative logarithm of its probability:

$$IC(c) = -\log_2(Prob(c)).$$

The probability of a GO term can be calculated as the number of annotations containing the term over the total number of annotations. Given two concepts, c_1 and c_2 , their similarity is the information content of their most specific common ancestor a over their information content:

$$Sim_{Lin}(c_1, c_2) = \frac{2 \times IC(a)}{IC(c_1) + IC(c_2)}.$$

The most specific common ancestor of two GO terms is their common ancestor that has the largest information content.

| 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.

4 Assessment

From the set of UniProt/SwissProt proteins with uncurated annotations and without manual annotations, we selected 66 proteins for which GOAnnotator identified evidence texts with more than 40% similarity and 50% confidence. 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.

In most cases, where the evidence text was correct, the GO term present in the extracted annotation was the same as the GO term present in the uncurated annotation (65 cases, see Table 3). Although the evidence text being correct, most of the times it did not exactly contain any of the known representations of the extracted GO term (see Section 3.1). 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 obvious as well that GOAnnotator can propose new GO terms.

4.1 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

| 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.

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).

5 Discussion

Researchers need more than facts, they need the source from which the facts derive [14]. 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, of which in 78% of the cases the GO term present in the uncurated annotation was confirmed. This performance meets the expectations of the curation process. However, 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 applied 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 solve this problem by retrieving full text documents from the Internet [4]. In addition, the list of documents cited in the UniProt 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. This 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 [3]. Koike et al. distinguished between gene and family names to deal with general terms [9]. 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 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 [8]. 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.

6 Conclusions

We presented GOAnnotator, a system that automatically identifies evidence text in literature for GO annotation of Uniprot/SwissProt proteins. GOAnnotator provided evidence text at high precision (93%, 66 sample proteins) taking advantage of existing uncurated annotations and the GO hierarchy. GOAnno-

tator incorporates text-mining methods to extract GO terms from text, and a similarity measure to select GO terms similar to GO terms from uncurated annotations.

GOAnnotator assists the curation process by allowing fast verification of uncurated annotations from evidence texts, which can also be the source for novel annotations. GOAnnotator is available through a Web interface, which enables the verification of uncurated annotations of any UniProt entry with evidence extracted from literature.

References

- [1] 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.
- [2] 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*, 32:262–166, 2004.
- [3] J. Chiang and H. Yu. MeKE: discovering the functions of gene products from biomedical literature via sentence alignment. *Bioinformatics*, 19(11):1417–1422, 2003.
- [4] D. Corney, B. Buxton, W. Langdon, and D. Jones. BioRAT: Extracting biological information from full-length papers. *Bioinformatics*, 20(17):3206–3213, 2004.
- [5] F. Couto, M. Silva, and P. Coutinho. Finding genomic ontology terms in text using evidence content. *BMC Bioinformatics*, 6(S1):S21, 2005.
- [6] GO-Consortium. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research*, 32(Database issue):D258–D261, 2004.
- [7] L. Hirschman, A. Yeh, C. Blaschke, and A. Valencia. Overview of BioCre-AtIvE: critical assessment of information extraction for biology. *BMC Bioinformatics*, 6(Suppl 1):S1, 2005.
- [8] J. Kim and J. Park. BioIE: retargetable information extraction and ontological annotation of biological interactions from literature. *Journal of Bioinformatics and Computational Biology*, 2(3):551–568, 2004.
- [9] A. Koike, Y. Niwa, and T. Takagi. Automatic extraction of gene/protein biological functions from biomedical text. *Bioinformatics*, 21(7):1227–1236, 2005.
- [10] D. Lin. An information-theoretic definition of similarity. In *Proc. of the 15th International Conference on Machine Learning*, 1998.

- [11] J. Mitchell, A. Aronson, J. Mork, L. Folk, S. Humphrey, and J. Ward. Gene indexing: characterization and analysis of NLM's GeneRIFs. In *Proc. of the AMIA 2003 Annual Symposium*, 2003.
- [12] H. Müller, E. Kenny, and P. Sternberg. Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLOS Biology*, 2(11):E309, 2004.
- [13] A. Pérez, C. Perez-Iratxeta, P. Bork, G. Thode, and M. Andrade. Gene annotation from scientific literature using mappings between keyword systems. *Bioinformatics*, 20(13):2084–2091, 2004.
- [14] D. Rebolz-Schuhmann, H. Kirsch, and F. Couto. Facts from text - is text mining ready to deliver? *PLoS Biology*, 3(2):e65, 2005.