Prediction of GO annotation by combining entity specific sentence sliding window profiles

Martin KRALLINGER
Protein Design Group,
CNB-CSIC,
Calle Darwin,
Cantoblanco,
E-28049 Madrid,
martink@cnb.uam.es

María Magdalena PADRÓN
Protein Design Group,
CNB-CISC,
Calle Darwin,
Cantoblanco
E-28049 Madrid,
mpadron@cnb.uam.es

1 Overview

Information extraction aims to derive from free text significant information related to a given query (Hobbs, 2002). The "Critical assessment of text mining methods in molecular biology (BioCreative)" is a community wide text mining and information extraction contest. It is divided into two major tasks, each composed by further sub-tasks. While task 1 refers mainly to named entity extraction of gene and protein names, task 2 is concerned with functional annotation of gene products from free text and is organized into four distinct sub-tasks. The BioCreative sub-task 2.1, aims to extract protein annotations using full length biomedical articles and taking Gene Ontology (GO) as reference. This means that for a given protein and GO term, a text passage should be returned where a traceable association between them is provided. In order to participate in sub-task 2.1 we developed a method which returns text passages associating a given protein to a GO-term. It is based on a strong name identification system, a catalog of GO related terms and patterns, and statistics derived from a collection of sentences derived from information contained in the Gene Ontology Annotation (GOA) database. The system generates distinct *sub-tag* sets or word lists for the query entities (protein and GO term). Each *sub-tag* set has associated a characteristic *sub-tag score*. To extract the sentences participating in the annotation event we used a *sentence sliding window* approach applying previously calculated *sub-tag scores* for the protein and GO term *sub-tag* matches.

2 Methods

GOA dataset

The Gene Ontology Annotation database (GOA) (Camon et al., 2004), contains a list of associations were proteins are linked to GO terms through scientific articles. Those articles contain text passages which describe the annotated protein in context with GO terms.

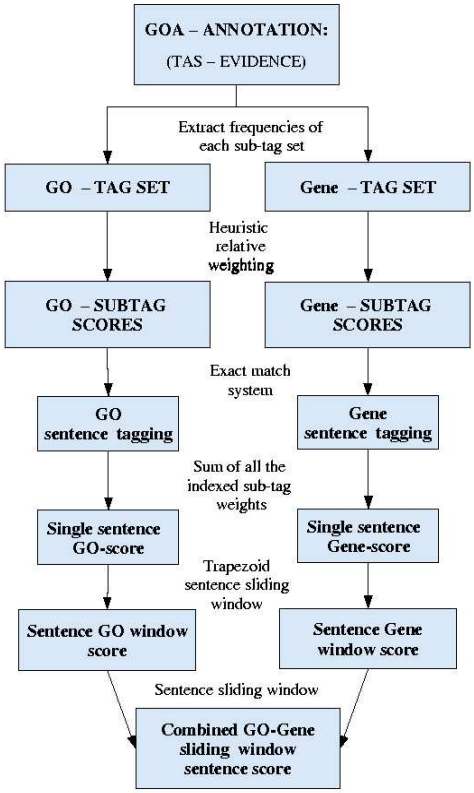

Figure 1: Flow chart of the construction of the combined gene/protein and GO-term sentences sliding scores.

We derived a "training" dataset of 560 Medline abstracts were traceable author statements (TAS evidence code) of protein annotations are provided by the GOA database. Those abstracts were used for further analysis for each of the entities involved in the annotation, namely the protein and the GO term. It is important to take into account sources of noise due to different annotation standards depending on accuracy of the manual annotation and

GO sub-tag set	Gene sub-tag set
GO term (original)	Gene name / symbol
NL-GO term	Variants of Gene name
Externally linked terms	Externally linked names
GO word tokens	Gene name word tokens
GO definition tokens	GOBO mutation term
GO co-occurence tokens	GOBO sequence term

Table 1: The basic *sub-tag* sets for the GO entity class and protein/gene entity class. For each *sub-tag* set a stemmed and lower case converted version was also derived.

the changes in the curation standards over time. As curator annotation was performed using full length articles, an additional problem concerning the abstracts is that they might not contain the passage of text relevant for annotation extraction.

Protein/gene tag set

Protein or gene names form one of the *entity classes* involved in the GO-annotation event. In order to allow tagging of this entity at different levels we constructed a manually derived *sub-tag* system (see table 1). Each *sub-tag* set contains a list of names, symbols or word types associated with the specific query protein. Among the *sub-tag* sets used was the original gene/protein name, symbol and identifier. Within textual sources the protein symbol is often expressed in form of typographical variants (Yeh et al., 2003). Thus an other *sub-tag* set contained a list of protein typographical variants obtained through a manual rule based pipeline. Moreover using cross references provided by other database sources (HUGO, OMIM, SwissProt, UniGene, LocusLink) allows extraction of protein name synonyms, that were organized in a sub-tag set containing a list of synonyms provided by externally linked databases.

In order to take into account pragmatic context information, the meronymic relations between the protein and the terms contained in Global Open Biology Ontologies (GOBO) dataset, e.g. the terms referring to mutation events and sequence ontology were incorporated as a separate *sub-tag* set.

Gene Ontology term tag set

The lexical properties of GO (McCray et al., 2002) were previously analyzed in order to establish whether they are suitable for Natural Language Processing (NLP) approaches. In analogy to the gene/protein *entity class* also a set of manually derived *sub-tags* for the GO-term entity was constructed (see table 1). Tagging of gene ontology terms is even more cumbersome, as the way terms are expressed in GO often does not correspond to the way they appear in free text. This is especially the case for certain terms within the categories molecular function and cellular component, which do not correspond to natural language expressions. A significant difference between the protein symbols/names and the GO terms is that gene names (symbols) correspond to proper nouns while GO terms are adverbial nouns, which are more difficult to identify in free text as they often lack morphological characteristics present in proper nouns. To modify GO terms in order to reach a higher degree of resemblance to the way they are used in free text, a rule based system which converts an input GO term in its corresponding "NL-variants" was implemented. This system carries out, besides minor morphological changes, acronym substitution, word token synonym substitution, collocation shuffling and preposition insertions. A sample NL-variant for the original GO term (GO-id 0000780) "*condensed nuclear chromosome/ pericentric region*" would be "*pericentric region of condensed nuclear chromosome*".

The Gene Ontology consortium also provides synonyms and external links to terms and keywords derived from other annotation databases which were included as a separate *sub-tag* set.

Analysis of sub-tag sets using GOA abstracts

Tagging of each of the entries belonging to the *sub-tag* sets for the protein and the GO entities was conducted for the GOA abstracts and their average sentence occurrences was calculated (see figure 2).

As the number of used GOA abstracts was small and contained considerable noise, it constitutes a not very representative text corpus. Therefore we used it only as a rough guide to derive the domain heuristic weighting scheme. Nevertheless, the average occurrence of each *sub-tag* reflects somehow its specificity. More specific sub-tags were given a higher weighting score then more general ones. The stemmed versions of each sub-tag class have lower scores compared to the original word/s.

The heuristic sub-tag score h_i for each *sub-tag*

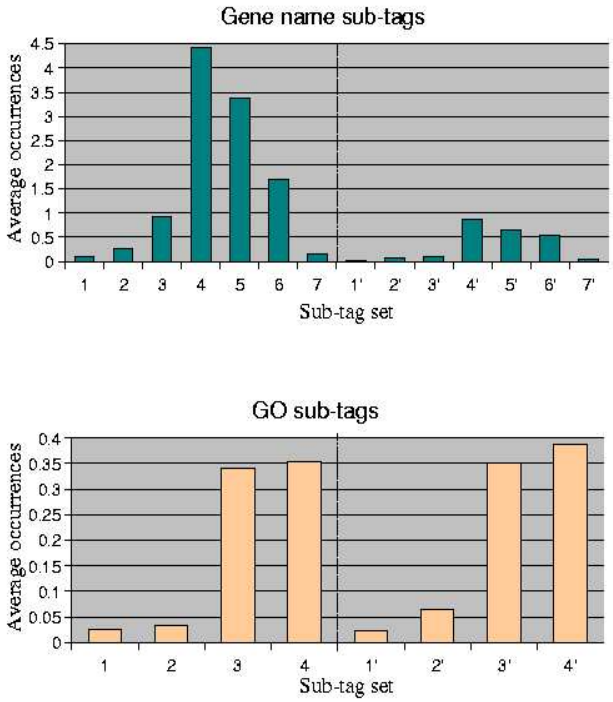

Figure 2: Average occurences of members of each sub-tag set within GOA abstract sentences.
Gene name sub-tags, 1: original gene name provided by GOA, 2: heuristic typographical variants of the gene name, 3: variants extracted from links to external databases, 4: word types which build up the gene names, 5: word types which build up the external linked gene names, 6 and 7: GOBO sequence ontology and mutation event terms respectively.
GO-sub-tags, 1: original GO term, 2: NL-variant of GO-term, 3: word types which build up the GO term, 4: word types which build up the GO-term definitions. Note that not all the categories are displayed in the bar diagram, co-occuring word types for GO-terms which were extracted from PubMed sentences have an avergae occurence in GOA abstract sentences of 11.3337254243. Primed numbers correspond to lower case and stemmed versions.

i was constructed as follows:

$$h_i = \bar{o}_i * e_i \quad (1)$$

where $\bar{o}_i$ corresponds to the average number of occurences of elements of *sub-tag* i in GOA sentences and e_i is the relative heuristic weight used for *sub-tag* i based on domain knowledge.

Trapezoid sentence sliding window
Sliding window models have been widely used in signal processing for analysis of frequent items (Datar et al., 2002), in many bioinformatics applications related with sequence analysis, e.g. (Sipos and vonHeijne, 1993) and in statistical natural language processing for collocation identification. We used an "averaging sliding window" approach to extract relevant information for intervals (windows) of sentences units.

To calculate the average *sentence score* for each of the entities over a fragment of text, a trapezoid *sentence sliding window* was used. The sentence position weight w_i within the trapezoid sliding window with length L (total number of sentences forming the window) was determined by

$$w_i = \begin{cases} 1 & \text{if } 1 < i < L, \\ 0.5 & \text{otherwise} \end{cases} \quad (2)$$

Hence the flanking sentences comprising the sliding window have a lower weight compared to the central window sentences.

The average *sentence score* for each entity, $\bar{H}$ can thus be calculated by

$$\bar{H} = \frac{\sum_{i=1}^{L} h_i w_i}{L} \quad (3)$$

where h_i is the sum of the scores of the matched *sub-tags* of the given sentence, w_i is its corresponding sentence position weight and L is the sentence window size, in the case of the entity profiles L=5 sentences.

Document Entity profiles
The entity *sentence sliding windows* result in a protein and a GO-term *document entity profile* respectively, when the sentence number is plotted versus the average sentence score. Each sentence has thus an entity score reflecting the average sub-tag score for this sliding window position. Those sentence scores can be used to determine relevant text passages for each entity. In general, the higher the sentence score for a certain sentence, the more high scoring sub-tags are matched on average for this entity within the sentence window.

Document annotation profile
In order to extract the relation between the entities, their profiles or sentence scores must be combined into an unique profile, a combined document annotation score reflecting the relation between both entities involved in the annotation. This is accomplished through a combining sliding window, which in principle is similar to the entity sliding windows. The combining window size was reduced to L=4

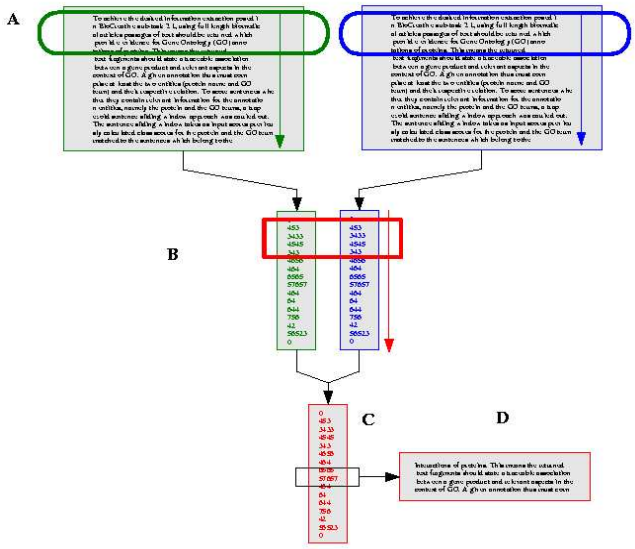

Figure 3: Flow chart ilustrating the combination of the distinct entity sentence scores. A: Trapezoid sentence sliding window generates the average sentence scores for each entity using the sub-tag scores of the matched words (document entity profiles), B: The average sentence scores of each entity are used to generate the combined average annotation score using a second step sliding window (document annotation profile), C: Selection of the highest combined average annotation score, D: Return sentences corresponding to the sentence window with the highest combined average annotation score.

sentences and the flanking regions of the combining sliding window have the same position weight w_i as the central sentence. We considered that semantic information expressing the relation of two entities should be restrained to a distance expressed in sentences.

The average annotation sentence score $\bar{A}$ of this combined window is calculated as the sum of the product of the entity sentence scores ($\bar{H}$) namely $\bar{x_i}$, for GO and $\bar{y_i}$ for the protein, divided by the window length L. If the average annotation sentence score is plotted versus the sentence number a *document annotation profile* was obtained. For a sample output of a *document annotation profile*, refer to figure 3.

The combined average annotation score $\bar{A}$ for a given window position is thus given by

$$\bar{A} = \frac{\sum_{i=1}^{L} \bar{x_i}\bar{y_i}}{L} \qquad (4)$$

The sentences comprising the highest scoring window, namely the highest average annotation sentence score, $\max \bar{A}$ are returned as annotation evidence text. Notice that the number of returned sentences for annotation evidence is dependent on the combining window size.

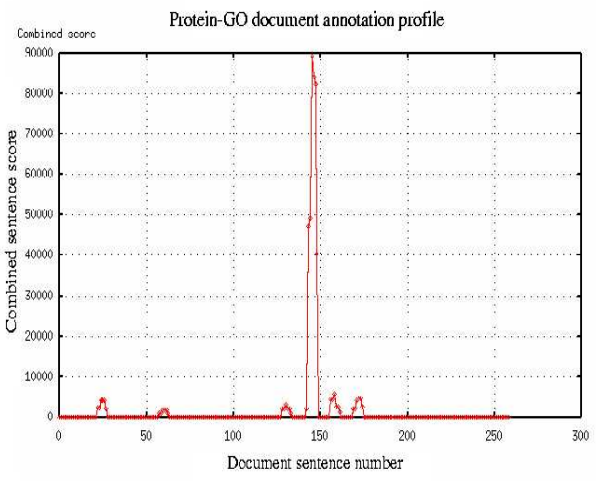

Figure 4: Sample output of an *annotation document profile* generated combining the entity document profiles: document sentence number vs combining annotation score.

3 Results

The test set provided for task 2.1. consisted in full text articles of the Journal of Biological Chemistry. The aim was to extract relevant passages of text for a given GO identifier, a SwissProt accession number and the article identifier. After low-level processing, junk formating and parsing, a sentence splitting algorithm was applied, indexing of *sub-tags* for each entity was performed and then the annotation profiles and highest scoring sentence window was calculated as described in the methods section.

Protein/GO Matches	Prediction Category				
	Low	General	High	None	**Total**
High	21.05	6.57	**28.85**	0	**56.47**
General	4.48	2.28	10.67	0	17.43
Low	12.10	4.10	8.19	0	24.39
None	0.10	0	0	1.61	1.71
Total	37.73	12.95	**47.71**	1.61	100

Table 2: Result summary for task 2.1. The table shows the percentages of evaluated evidences organized by precision categories for proteins (rows) versus precision categories of GO terms (columns). The label corresponds to, high: correct prediction, general: not totally wrong prediction but too general to be really useful for protein annotation (for GO-terms) and that the specific protein is not there but a homologue from another organism or a reference to the protein family is contained (for Protein), low: means basically wrong. Total refers to the entity extraction (protein or GO term) and None are not evaluated cases.

A total of 1076 text fragments were extracted

as candidate text annotation passages and submitted as evidences for GO annotation of proteins. Out of those submissions 1050 were evaluated by expert curators (see table 2).

Within the assessment of the submissions, also the extraction of the involved entities themselves was evaluated separately. About 56% (594 evidences) of our predictions were assessed as corresponding to high precision (correct) protein extraction and about 48% (501 evidences) were evaluated as high precision GO-term extraction.

GO-category Matches	Prediction Category			
	Low	Generally	High	None
Function	30.29	11.76	52.35	5.58
Component	28.10	14.59	56.21	1.08
Process	47.71	13.11	39.70	3.46
All	37.71	12.95	47.71	1.61

Table 3: Percentage of evidences per precision categories for the GO term entity extraction organized by its corresponding GO-category.

A closer analysis of the evaluated GO terms, revealed the existence of differences in prediction performance depending on the associated GO-category (see table 3). Within each GO-category, the highest percentage of correct predictions corresponded to the category cellular component, followed by the molecular function category and in the biological process category. The cellular component document profiles often show high average sentence scores, due to matching of high scoring *sub-tags* (e.g. the original GO term or its NL-variant). Scoring of GO-terms belonging to the category biological process was more difficult, this agrees to previous attempts to identify process terms in biomedical abstracts (Marquet et al., 2003). A plausible reason might be the broad diversity of describing biological processes, often colloquially expressed. In the case of annotation predictions, the highest number of accurately predicted associations corresponded to the GO category molecular function. Cellular component terms often are formed by word tokens which can be used in other contexts this leads to an increased number of false predictions.
We obtained an overall result of 28.8% (303 textual evidences) of high precision, predictions of protein GO-annotations (see table 2). An example of high precision annotation extraction is displayed in figure 5, were the evidence text relevant for the annotation was correctly identified.

```
<protein>
<namefile>JBC_2001-2\bc4501042445.gml<\namefile>
<idTask>2.1<\idTask>
<participant>user14<\participant>
<nameProtein><\nameProtein>
<dbId>O15023<\dbId>
<sourceDb>Swiss-Prot<\sourceDb>
<goCode>
<name>phosphatidylinositol binding<\name>
<code>0005545<\code>
<evidenceText>In addition, a single point mutation in the FYVE finger motif at cysteine residue 753 (C753S) is sufficient to abolish its endosomal association. Its endosomal localization is also sensitive to the phosphatidylinositol 3-kinase inhibitor, wortmannin. Using in vitro liposome binding assays, we demonstrate that Myc-tagged endofin associates preferentially with phosphatidylinositol 3-phosphate, whereas the C753S point mutant was unable to do so. We also show that endofin co-localizes with SARA but that they are not associated in a common complex because they failed to co-immunoprecipitate in co-expressing cells.<\evidenceText>
<\goCode>
<\protein>
```

Figure 5: Example output of a GO protein annotation predicted with high precision.

Furthermore, cases where the protein entity was evaluated as 'high' and the GO term was evaluated as 'generally' constituted 6.57% of out submissions.

4 Conclusions

It is generally believed that information extraction of relationships is more challenging when compared to single entity extraction. This is also reflected in the overall results of the BioCreative task 2.1 result, and was previously pointed out by Xie et al. (Xie et al., 2002). Attempts have been made to use automatic extraction of sentence patterns which could aid in annotation beyond the involved entity detections (Chiang and Yu, 2003). In practice, evidence relevant for protein annotations can often not be detected by mere co-occurrence of the protein entity and the annotation term within a single sentence. We thus developed a method which analyze text fragments in form of sliding sentence windows, which allows to score whether they contain relevant information for a given entity. The obtained results of out method are promising, but still some improvements could be carried out, in particular related to the subtag entity recognition. Also the window size and entity sentence score combination could be optimized.

5 Acknowledgements

This work was sponsored by DOC, doctoral scholarship programme of the Austrian Academy of Sciences and the ORIEL (IST-2001-32688) and TEMBLOR (QLRT-2001-00015) projects.

References

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

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

M. Datar, A. Gionis, P. Indyk, and R. Motwani. 2002. Maintaining stream statistics over sliding windows. *SODA.*, pages 635–644.

J.R. Hobbs. 2002. Information extraction from biomedical text. *J Biomed Inform.*, 35:260–264.

G. Marquet, A. Burgun, F. Moussouni, E. Guerin, F. LeDuff, and O. Loreal. 2003. BioMeKe: an ontology-based biomedical knowledge extraction system devoted to transcriptome analysis. *Stud Health Technol Inform.*, 95:80–85.

A.T. McCray, A.C. Browne, and O. Bodenreider. 2002. The lexical properties of the gene ontology. *Proc AMIA Symp*, pages 504–508.

L. Sipos and G. vonHeijne. 1993. Predicting the topology of eukaryotic membrane proteins. *Eur J Biochem.*, 213:1333–1340.

H. Xie, A. Wasserman, Z. Levine, A. Novik, V. Grebinskiy, and A. Shoshan. 2002. Large-scale protein annotation through gene ontology. *Genome Res.*, 12:785–794.

A.S. Yeh, L. Hirschman, and A.A. Morgan. 2003. Evaluation of text data mining for database curation: lessons learned from the KDD Challenge Cup. *Bioinformatics.*, 19:331–339.