

BioCreative Task 2.1: The Edinburgh-Stanford system

Yuval Krymolowski and Beatrice Alex and Jochen L. Leidner

Institute for Communicating and Collaborative Systems

University of Edinburgh, United Kingdom

{ykrymolo, vlbalex, s0239229}@inf.ed.ac.uk

Abstract

We describe a system for BioCreative Task 2.1: finding evidence that supports a GO term annotation for a given protein in a given biomedical paper. We approach the problem as a question answering task, where the query is constructed from a protein name, a GO term and its definition.

1 Introduction

The GO ontology is a hierarchical collection of biological processes, cellular components, and molecular functions. The terms in the ontology tell i) what are the processes in which the protein takes part (“biological process” terms), ii) where in the cell they occur (“cellular component” terms), and iii) how the protein operates (“molecular function” terms). The goal of BioCreative Task 2.1 was to extract pieces of evidence for GO codes from a given set of articles published between the years 1998-2002 in the Journal of Biological Chemistry (JBC).

In this paper, we describe a question-answering approach to this task. Our approach starts with word-level normalisations of the term definitions, along with a heuristic search in the paper body for acronyms explicitly referring to the protein. We then construct a set of queries from the words comprising the protein name, explicit protein references, and the GO term definition. These queries are further expanded using the GO ontology. Our query approach is similar to the one taken by Osborne et al. (2003, Section 3) for finding evidence for GeneRIF codes in paper abstracts.

2 Input and Pre-processing

For each JBC article, the accession number of a protein as well as a GO code were specified (usually several GO codes were given per article). Using the ac-

cession number, we acquired the full name of the protein from the Human Genome Organisation (HUGO¹) database². In addition, we retrieved the name of the GO term and its definition from the GO database.

The protein name, GO term and term definition served as a basis for generating a query for potential evidence text. In this section we describe and motivate the pre-processing steps that were carried out in order to generate a query. The following steps were used:

- Stemming: normalising nouns, verbs, and adjectives derived from a single root into one lemma.
- Acronym lookup: finding acronyms in the paper that refer to the protein in question.
- Query expansion: adding query terms based on the GO ontology.
- Markup: marking certain words with labels that would be retrieved by queries.

2.1 Stemming

The GO definition often contains nominalisations or adjectives that describe certain functions, whereas the body of the paper may contain verbs for describing the same functions, for example:

PMID: 10026212

go code: GO:0004337

go name: geranyltranstransferase activity

go def: Catalysis of the reaction: geranyl diphosphate + isopentenyl diphosphate = diphosphate + trans,trans-farnesyl diphosphate.

evidence: Geranylgeranyl diphosphate (GGPP) synthase (GGPPSase) catalyzes the synthesis of GGPP,

...

¹<http://www.hugo-international.org/hugo>

²<http://www.gene.ucl.ac.uk/public-files/nomen/nomeids.txt>

Similarly, the body of the paper may contain an adjective derived from the same lemma as a nominalisation that appears in the GO information, as this example shows:

PMID: 10037736

prot name: NF-E2-related factor 3

go code: GO:0003700

go name: transcription factor activity

evidence: ... indicating that Nrf3 is a transcriptional activator.

As these examples demonstrate, such nominalisations and adjectives need to be reduced to a common lemma form in order for sentences such as the above to be retrieved. We therefore made the following conversions:

- Adjectives ending with “lytic” or “otic” were converted to nominalisations, e.g., “catalytic” → “catalysis”, “hidrotic” → “hydrosis”,
- Words ending with “ional” were converted to nominalisations by removing the final “al”, e.g., “transcriptional” → “transcription”, and finally
- Nominalisations were converted to the corresponding verb lemma, e.g. “catalysis” → “catalyse” using a list from the Unified Medical Language System[®] (UMLS³).

2.2 Acronym lookup

In general, paper authors tend to use an acronym when referring to the protein, either one or several aliases or a novel acronym, for example:

1. **PMID:** 10026212

prot name: Geranylgeranyl pyrophosphate synthetase (GGPP synthetase) (GGPPSASE)
(Geranylgeranyl diphosphate synthase)

reference in paper: GGPPSASE

2. **PMID:** 10037736

prot name: NF-E2-related factor 3

reference in paper: Nrf3

3. **PMID:** 10066790

prot name: Adapter-related protein complex 4
 μ 1 subunit

reference in paper: AP4, AP-4

Since existing acronym lists may be incomplete or not entirely up to date, we used the heuristic shown in Figure 1 for finding the acronym used in an article.

Find protein reference:

1. Test the 10 most frequent words or bigrams in descending order of frequency. For each word (or bigram) :
2. Test whether it:
 - contains two upper-case letters, or
 - contains alphabetic as well as digits or greek letters.(ignoring “DNA” and “RNA”)
3. If the word satisfies these conditions, identify it as the acronym.
4. Otherwise test whether the word/bigram is part of the protein name (ignoring the words “gene” and “protein”) and accept it as a reference to the protein if this is the case

Figure 1: Heuristic for finding a synonym that refers to the protein which is the subject of a biomedical paper.

Where the algorithm does not find an acronym, it proposes a sub-string of the protein name where possible.

The heuristic is based on the hypothesis that this acronym would be among the most common acronyms in the article, because it refers to the protein which is being discussed.

For the purpose of acronym search, words were standardized by removing dashes. In addition, since sometimes a protein name may start with a lower-case letter indicating the organism, we removed lower-case letters from the beginning of the name where followed by an upper case letter. This standardisation resulted, for example, in the strings “hGCAP-3” and “GCAP3” being counted as the same symbol.

While in general the acronym is derived from the protein name, in some articles the most frequent acronym denoted a homologue of the protein, or was generated by a different naming scheme:

1. **PMID:** 10066793

prot name: LAK-1

reference in paper: TRF4

2. **PMID:** 10066796

prot name: Ankyrin-like protein

reference in paper: p120

3. **PMID:** 10075682

prot name: Potassium channel subfamily K member 6

³<http://www.nlm.nih.gov/research/umls>

reference in paper: TWIK2, TWIK-2

Although we found this heuristic generally helpful in recovering acronyms, it failed under certain circumstances:

1. **PMID:** 10075657

prot name: Neurogranin

acronym found: CaM (abbreviation for Calmodulin)

reason: The paper discusses interactions between Neurogranin and Calmodulin, hence the high frequency of “CaM”.

2. **PMID:** 10318868

prot name: Dynactin 6

acronym found: OXPHOS

reason: The paper discusses the effect of lack of the ANT1 isoform of the adenine nucleotide translocator on up-regulation of nuclear and mitochondrial genes in mice. Lack of this protein is related to the human oxidative phosphorylation (OXPHOS) disease, hence the acronym denoting this process is very frequent here.

These problems could be amended by further linguistic processing, but restricted the analysis to the word level. In addition, we ignored the experimental section of the biomedical articles because this section is unlikely to contain evidence texts for GO terms.

2.3 Query expansion

The evidence text may contain words that do not appear in the GO definition or term, but are related to it. For example, evidence for “signal transduction” (GO:0007165) may stem from sentences containing:

- verbs like “inhibit”, “stimulate”, “activate”,
- references to cellular components like the Golgi,
- descriptive words like “intracellular” or “receptor”.

Looking for evidence for a GO term, we used words from a list compiled by human experts for query expansion. In addition, we used words from the titles of more specific GO terms.

2.4 Markup

In the last preprocessing step we lemmatised nouns and verbs that appear in the protein name, the GO term name and the GO term definition for each document. This processing relied on certain nominalisations and

adjectives being reduced to verb lemmas in the stemming step, and its output was a list of noun and verb lemmas from the protein name and the GO information.

We then marked up nouns and verbs that have these lemmas in the article with labels denoting their part-of-speech, and the information field where each lemma appears. For example, the noun “GGPP-SASE” that appeared in the protein name of the article with the PMID 10037736 was annotated with the tag “PROTNAME-NN”. In case a verb or noun appeared in several fields, e.g. in the protein name and the GO definition, it was annotated with all corresponding tags. In addition, acronyms and words that can be used for query expansion were marked up with special tags.

2.5 Summary

At the end of the preprocessing phase we had article words annotated with tags referring to potential evidence they provide. These tags were then used as input into *qtile*, the query-based passage ranking tool.

3 Question Answering System

We re-cast Task 2.1 as a *query-based passage segmentation and ranking* problem: given a query and a document, find the top-N sentences in the document whose content is most closely related to the query.

This view made the problem directly susceptible to re-use of an existing query-based passage ranking tool: *qtile* (Leidner et al., 2003) was originally developed to reduce the amount of text to be parsed in QED, the Edinburgh question answering system for TREC-12.

This “tiler” extracts from the set of documents a set of segments (“tiles”) based on the occurrence of relevant words in a query, which comprises the words of the question. A sliding window is shifted sentence by sentence over the text stream, retaining only the window tiles that contain: i) *at least one* of the words in the query and ii) *all upper-case query words*.

The scoring function for each tile considers

- the number of non-stopword query word tokens occurring in the tile;
- a bonus for corresponding capitalisation of each term in query and tile, respectively
- term bigrams and trigrams that occur in both question and tile

The score for every tile is multiplied with a triangular window function to weight sentences closer to the middle of a window higher.

	total annotations	GO term missed		evidence loosely related		evidence too general		evidence accurate for GO term	
biological process	517	110	21%	232	44%	76	14%	99	19%
cellular component	181	37	20%	81	44%	18	9%	45	24%
molecular function	319	85	26%	102	31%	39	12%	93	29%

Table 1: Evaluation of the quality of GO term evidence text by run 1.

4 Application

For the application in Task 2, the query consisted of the different information tags generated by the pre-processing phase. In order to avoid over-scoring by a chance alignment of tagged words, we inserted a dummy word between query tags to ensure that no bigrams or trigrams from the query appear in the processed document.

In order to get more likely sentences prior to less likely ones, we applied filters to the query output in the following order:

1. Sentences that contain a word from the query expansion, (derived from the GO term)
2. a verb from the GO definition,
3. a verb from the GO name,
4. a noun from the GO name, or
5. no restrictions

The filters were applied as a series of queries with different constraints. We use a window of one sentence on each side for scoring and extract the highest ranking tile satisfying each query. The top three matching sentences were selected as outputs of three runs.

5 Results

The success rates in retrieving evidence for both a GO term and the protein were approximately 15% for all three runs. Table 1 shows the results of run1 when evaluated according to the success in retrieving evidence for a GO term regardless of the protein.

A possible explanation for the significant difference in performance between the results with and without taking the protein into account, is that the acronym and the protein name served as additional query items but were not imposed by any of the filters. In addition, of the 108 articles in the test data, 20 articles had annotations for two or more proteins. In these cases, the acronym would be misleading for at least one of the proteins.

Analysis of the system output reveals a few reasons for failing to provide the most precise evidence

text. In some cases, certain words in the GO term are more important than others. For example, the word “vesicle” in “vesicle organization and biogenesis” (GO:0016050) is more important than the other words in this term. Our system extracted the evidence sentence

“Members of the FYVE domain family of proteins have been implicated in protein trafficking and signal transduction.”

for this term from the article in file JBC_2001-2/bc4501042445, while better evidence is provided by the sentence:

“The vesicles appeared clustered or fused together into larger structures, an effect that was most pronounced when only the 100-amino acid FYVE domain region was expressed.”

It may be possible to single out an important word from the term name by comparing it with the term directly above in the GO hierarchy, in this case “organelle organization and biogenesis” (GO:0006996). A word that does not appear in the more general term is likely to be more related to the specific meaning of the term for which we seek evidence.

Other retrieval errors resulted from different wording used in the article for describing a certain function, or sentences with high concentration of words that appeared in the term definition but were not indicative enough.

6 Related Work

The secondary task of the TREC-2003 Genomics Track (Hersh and Bhupatiraju, 2003) was similar to the present task. This task concerned with recovering GeneRIF annotations from abstracts. As preliminary analysis had revealed, 77% of the GeneRIF snippets contained either complete or partial chunks of text from the title or the abstract. The current task, of retrieving evidence from the whole paper, and when there is no reason to assume that there will be common chunks of text between the GO term data and the article text, is therefore harder.

Chiang and Yu (2003) describe a work that is more similar to the BioCreative task 2.1, recovering GO evidence from paper abstracts. Having observed that there are a number of commonly used patterns for describing functions of proteins, they trained a sentence-alignment system for recognizing these patterns. For the alignment process, they marked up gene products and function names in each sentence. The patterns identified in this way were then used for constructing a naive Bayes model for sentence classification.

The overall success rate in recovering product and GO evidence pairs using the sentence classification algorithm was 14.6%. The authors also note the difficulty of finding the protein name in the abstract, reporting a success rate of 75% .

While abstracts are clearly a good source for functional annotation, an article may contain GO evidence in the introduction or conclusion as well. Such evidence may be expressed in a way which is less dense than evidence from an abstract.

In future work, we will aspire to improve the quality of word-level modelling by taking more advantage of the structure of the GO ontology.

Acknowledgements

We thank Kirsty Newman and Frances Turner for their help in reasoning about the relevance of sentences for providing evidence for GO terms.

This work was supported by a Scottish Enterprise Edinburgh-Stanford Link Grant (R36759).

References

- Jung-Hsien Chiang and Hsu-Chun Yu. 2003. MeKE: discovering the functions of gene products from biomedical literature via sentence alignment. *Bioinformatics*, 19(11):1417–1422.
- W. Hersh and R.T. Bhupatiraju. 2003. The TREC 2003 genomics track overview. In *Proceedings of the Twelfth Text Retrieval Conference (TREC 2003)*, Gaithersburg, MD.
- Jochen L. Leidner, Johan Bos, Tiphaine Dalmás, James R. Curran, Stephen Clark, Colin J. Bannard, Mark Steedman, and Bonnie Webber. 2003. The QED open-domain answer retrieval system for TREC 2003. In *Proceedings of the Twelfth Text Retrieval Conference (TREC 2003)*, pages 595–599, Gaithersburg, MD.
- Miles Osborne, Jeffrey Chang, Mark Cumiskey, Nipun Mehra, Gail Sinclair Veronica Rotemberg, Matthew Smillie, Russ B. Altman, and Bonnie Webber. 2003. Edinburgh-stanford TREC 2003 genomics track. In *Proceedings of the Twelfth Text Retrieval Conference (TREC 2003)*, Gaithersburg, MD.