

The Edinburgh-Stanford system for BioCreative Task 2.1

Yuval Krymolowski, Beatrice Alex, Jochen L. Leidner

School of Informatics
University of Edinburgh



Approach to the Task

- We approach the problem as a querying task
- The query is constructed from:
 - The protein name,
 - the GO term name, and
 - the GO term definition.
- Evidence text is retrieved by extracting sentences that match the query.

System outline:

Given an article, a GO code, and a protein:

- Pre-process the article text, as well as the GO term name and definition, and the protein name,
- Construct a query and retrieve the ranked list of sentences from the article,
- Suggest the best-matching sentence as evidence text.

Pre-processing Steps:

- **Stemming:**

Normalising nouns, verbs, and adjectives derived from a single root into one lemma.

- **Acronym lookup:**

Finding acronyms in the article 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 the query.

Pre-processing Steps:

⇒ **Stemming:**

Normalising nouns, verbs, and adjectives derived from a single root into one lemma.

- **Acronym lookup:**

Finding acronyms in the article 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 the query.

Observation: word variation

- Nominalisation in the GO definition
⇒ verb in text:

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, ...

Observation: word variation

- Nominalisation in the GO term name
⇒ adjective in text:

GO code: GO:0003700

GO name: transcription factor activity

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

Pre-Processing: Stemming

- Converting adjectives ending with “lytic” or “otic” to nominalisations:
 - “catalytic” → “catalysis”, “hidrotic” → “hydrosis”
- Converting words ending with “ional” to nominalisations by removing the final “al”:
 - “transcriptional” → “transcription”
- Converting nominalisations to the corresponding verb lemma: (using a list from the UMLS)
 - “catalysis” → “catalyse”
- Lemmatising verbs.
 - “transcribing” → “transcribe”

Pre-processing Steps:

- **Stemming:**

Normalising nouns, verbs, and adjectives derived from a single root into one lemma.

⇒ **Acronym lookup:**

Finding acronyms in the article 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 the query.

Observation: Use of Acronyms

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

- **prot name:** Geranylgeranyl prophosphate synthetase
reference in paper: GGPPSASE
- **prot name:** Aapter-related protein complex 4 μ 1
subunit
reference in paper: AP4, AP-4
- **prot name:** LAK-1
reference in paper: TRF4 (a homologue)

Pre-Processing: Acronym lookup

We used a heuristic in order to find the reference to the protein discussed:

- Test the 10 most frequent words in descending order of frequency. For each word, test whether it:
 - contains two upper-case letters, or
 - contains alphabetic as well as digits or greek letters.
(ignoring “DNA” and “RNA”)
- If it does, identify the word as the acronym.
- Otherwise:
 - If the word is part of the protein name
then accept it as a reference to the protein.

Pre-processing Steps:

- **Stemming:**

Normalising nouns, verbs, and adjectives derived from a single root into one lemma.

- **Acronym lookup:**

Finding acronyms in the article 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 the query.

Observation: related words

- The evidence text may contain words that do not appear in the GO definition or name, but are related to it.
- Example: evidence for “signal transduction” may contain:
 - verbs like “inhibit”, “stimulate”, “activate”,
 - references to cellular components like the Golgi,
 - descriptive words like “intracellular” or “receptor” .

Pre-Processing: Query expansion

We expanded the query using words:

- From a list compiled by an expert, indicating which words can be used in order to find evidence for GO terms based on the term name.
- From more specific GO terms, direct descendants in the GO hierarchy. For example:
 - signal transduction
 - cell surface receptor linked signal transduction
 - interpretation of external signals that regulate cell growth
 - intracellular signaling cascade
 - regulation of signal transduction
 - two-component signal transduction system

Pre-processing Steps:

- **Stemming:**

Normalising nouns, verbs, and adjectives derived from a single root into one lemma.

- **Acronym lookup:**

Finding acronyms in the article 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 the query.

Pre-Processing: Markup

After the words in the input were stemmed, they were replaced with the tags: (possibly multiple tags)

GODEF_NN, GODEF_VB:

A noun or verb in the GO term definition,

GONAME_NN, GONAME_VB:

A noun or verb in the GO term name,

PROTNAME_NN, PROTNAME_VB:

A noun or verb in the protein name,

PROTFOUND:

An acronym that refers to the protein, and

GOCOND:

The word appears in the query expansion

System outline:

Given an article, a GO code, and a protein:

- Pre-process the article text, as well as the GO term name and definition, and the protein name,
- ⇒ Construct a query and retrieve the ranked list of sentences from the article,
- Suggest the best-matching sentence as evidence text.

Querying the article text

- The query consisted of the set of tags that actually appeared in the text.
- We used a passage ranking system (Leidner *et al.*) that ranks sentences according to the fraction of query words that they contain.
- We applied filters to the query output in the following order:
 1. Sentences that contain a **GOCOND** word
 2. " a **GODEF_VB** word,
 3. " a **GONAME_VB** word,
 4. " a **GONAME_NN** word,
 5. no restriction

Results

The success in retrieving both GO code and protein evidence is 15% .

Performance in retrieving GO term evidence:
(regardless of protein)

GO term type	total annotations	evidence accurate for GO term	
biological process	517	99	19%
cellular component	181	45	24%
molecular function	319	93	29%

Note: the protein name is not included in any of the query filters, as the system is more focused on the GO terms.

Discussion

Possible reasons for failing to provide the most precise evidence text:

- Some words in the GO term name are more important than others:
 - The word “vesicle” in “vesicle organization and biogenesis”

Future work:

- ignore very frequent words in the term name,
- compare with the names of less specific (parent) GO terms.

Discussion

- Certain parts of the paper are more likely to contain evidence:
 - The introduction and the concluding paragraphs are more likely to contain statements about the nature of the protein being studied.

Future work:

- assign more weight to evidence from these parts.
- weights may differ according to the GO term type, annotated training data would be of help.