**INTRODUCTION**

Gene Ontology (GO) database plays an important role in modern bioinformatics [1]. GO annotation, the fundamentals of GO database, is a human intensive work, which heavily relies on the expertise of the curators. To approach the ultimate goal of automated annotation, we need a GO concept recognition tool at the very beginning.

GO concept recognition is difficult because of the following reasons:

1. A GO concept is usually not presented in a sentence as its formal name nor its documented synonym.
2. It is not a name-entity recognition (NER) task. There is something more than a name-entity in a GO term, and there is no boundary of GO concept in a sentence.
3. More the tokens a GO term has, more difficult to recognize it in the sentence.

Our system called GREPC is capable of recognizing the GO concepts in the whole GO spectrum, including the complicated concepts deep down in the GO hierarchy.

**OBJECTIVES**

The most important question in building a GO recognition tool is “What is a GO concept?”

1. Dissect the GO terms to define the divisions of a GO concept.
2. Mining the GO database to fulfill the information of the divisions for each GO concept.
3. Build a function to recognize GO concepts from a given sentence from literature.
4. Calculated the precision, recall, and F1-measure of our system.

**METHODS**

For a GO concept, we divided its name, and synonyms, by regular-expression rules into the main part and the constrain part. The keywords used here are “involved in”, “during”, “via”, “using”, “by”, and “acting on”.

In the main part of the concept, we identified the pattern used in the concept by regular-expression based on table 1. At the same time, the entity was also identified by removing the heading or tailing tokens that were identified as patterns. In the constrain part, we did the same process as in the main part, discarded the patterns in the constrain part, and kept the entity in this part as the constrain. Thereafter, we built two inverted indexes from entity to GOIDs and from pattern to GOIDs for searching in the next step.

**RESULTS**

GREPC has remarkable results of recall rates and stable scores in both developing and testing datasets, shown in table 2.

**CONCLUSION**

By dividing a GO concept into divisions of entity, pattern, and constrain, we can apply strict, lenient, and strict matching respectively, to recognize a GO concept.

We only used the information from GO database and not yet used the data from BioCreative. We believe that GREPC will be an ultimate GO concept recognition engine after integration of these data. To our knowledge, this is the first GO concept recognizing system, which promising full GO spectrum search, deep GO search, and is open-sourced.

**REFERENCES**


**CONTACT**

Chia-Jung Yang
jeroyang@gmail.com
GREPC demo here!