Protein Subcellular Localization Prediction: A Natural Language Processing Approach

Yifeng Liu*, Alona Fyshe and Duane Szafron

yifeng@cs.ualberta.ca*

Abstract

- Swiss-Prot 51.3 protein entries
  - only 54% have subcell annotation
  - ~90% reference to PubMed abstracts
- BIG IDEA:
  - Build classifier to predict subcell location based on these abstracts!
- Challenge:
  - Abstracts are written in natural language, not database entries.
- Our approach:
  - Baseline approach: map abstract to a set of text features.
  - Improvements: include synonyms and generalized GO terms.

Baseline Approach

(Figure 1: Flowchart)

- How to make a text classifier
  - retrieve associated PubMed abstracts for protein p
  - tokenize, stem, form feature vector with stemmed words
  - assign weight to each feature according to its inverse document frequency statistics (TFIDF)
  - train binary support vector machine (SVM) classifier on a labeled set of feature vectors
  - classify protein p by classifying associated abstracts

Improvements

- Two major improvements:
  - synonym resolution
  - term generalization
  - ~ both rely on the GO hierarchy
- What is GO?
  - controlled vocabulary of biology
  - concept hierarchy organized as direct acyclic graph (DAG)
- We use GO as:
  - thesaurus for words with synonyms
  - source of generalized features

Figure 1: Flowchart for processing biological abstracts and text classification.

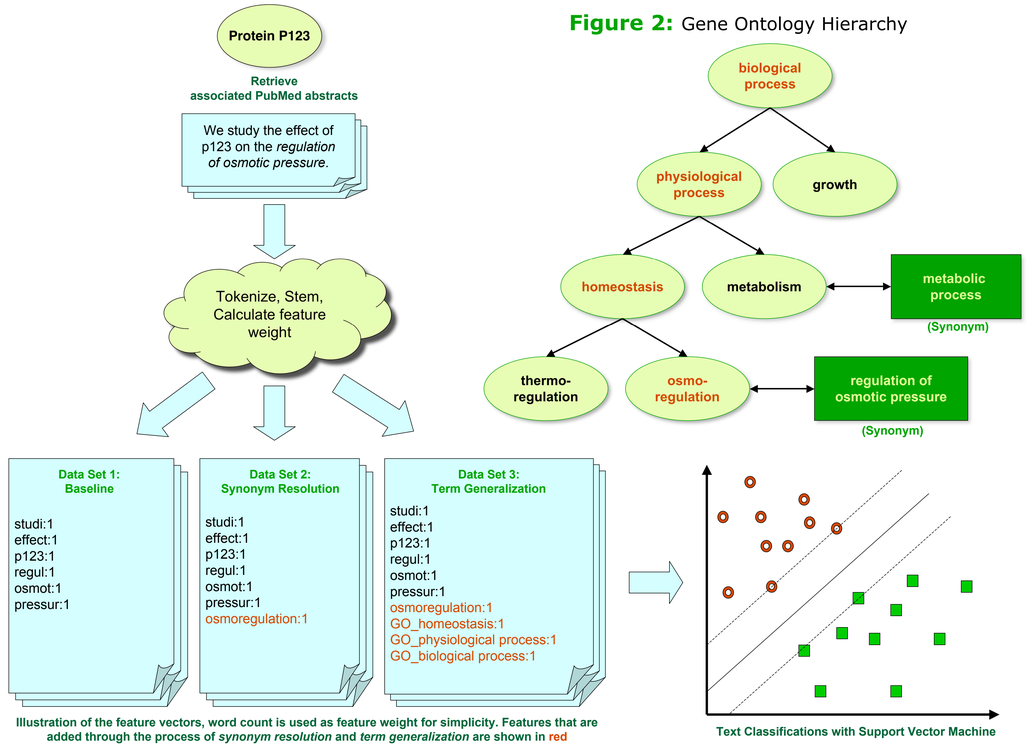

Illustration of the feature vectors, word count is used as feature weight for simplicity. Features that are added through the process of *synonym resolution* and *term generalization* are shown in red

Figure 2: Gene Ontology Hierarchy

Table 1: F-measures for stratified 10-fold cross validation on the Proteome Analyst animal data set. Results deemed significantly improved over the baseline (p=0.05) appear in **bold**. Stand deviation is shown in parentheses.

Class	Data Set 1 Baseline	Data Set 2 Synonym Resolution		Data Set 3 Term Generalization	
	F-measure	F-measure	Δ	F-measure	Δ
cytpopasm	0.740 (±0.049)	**0.758** (±0.042)	+0.017	**0.761** (±0.042)	+0.021
endoplasmic reticulum	0.760 (±0.055)	**0.779** (±0.068)	+0.019	**0.786** (±0.072)	+0.026
extracellular	0.931 (±0.009)	**0.935** (±0.009)	+0.004	**0.935** (±0.010)	+0.004
lysosome	0.746 (±0.107)	**0.787** (±0.100)	+0.041	**0.820** (±0.089)	+0.074
mitochondrion	0.840 (±0.041)	**0.848** (±0.038)	+0.008	0.852 (±0.039)	+0.012
nucleus	0.885 (±0.014)	0.885 (±0.016)	+0.001	0.887 (±0.019)	+0.003
peroxisome	0.790 (±0.054)	**0.823** (±0.042)	+0.033	**0.868** (±0.046)	+0.078
Average	0.815 (±0.016)	**0.832** (±0.012)	+0.017	**0.845** (±0.009)	+0.030

Reference: A. Fyshe and D. Szafron, Term Generalization and Synonym Resolution for Biological Abstracts: Using the Gene Ontology for Subcellular Localization Prediction, BioNLP Workshop at the Human Language Technology Conference – North American chapter of Association for Computational Linguistics (HLT-NAACL), June 2006, 17-24.

Improvements (cont.)

(Figure 2: GO hierarchy)

- **Synonym Resolution**
  - search GO for synonyms and include them in feature vector.
  - why: combine the weight of several synonyms, allowing SVM to more accurately model the author's intent.
- **Term Generalization**
  - search for GO node name (or synonym) in an abstract
  - include all names of the ancestors (red) along the GO hierarchy in the feature vector
  - why: allow the SVM algorithm to learn the correlations between general terms and subcell localizations.

Evaluation & Performance

(Table 1)

- **Evaluation**
  - Using the Proteome Analyst animal data set (***www.cs.ualberta.ca/~bioinfo/PA***)
  - F-measures for the 10-fold cross-validation are reported in *Table 1*.
  - *F-measure*
    - *= 2*precision*recall / (precision + recall)*
- **Performance**
  - Baseline F-measure ~ [0.740, 0.931]
  - With the two major improvements
    - Synonym resolution ~ [0.758, 0.935]
    - Term generalization ~ [0.761, 0.935]
  - Are these improvements significant?
    - perform paired t-tests with p=0.05 between three data sets
    - classifiers with significantly better performance over the baseline appear in **bold**.

Conclusion

- External information is beneficial in processing biological abstracts.
- GO can be used as a reference for both synonym resolution and term generalization and doing so significantly improve the performance of our text classifier.