A Machine Learning Model and Evaluation of Text Mining for Protein Function Prediction

Jari Björne*, Tapio Salakoski
Turku Centre for Computer Science TUCS
University of Turku, Department of IT, FI-20014, Turku, Finland
*To whom correspondence should be addressed: jari.bjorne@utu.fi

1. INTRODUCTION

We present a machine learning model which uses multiple datasets for improving GO term prediction, and evaluate the suitability of text mining as an additional dataset. To the best of our knowledge this is the first evaluation of *event*-type text mining data as a resource for automated function prediction.

The 2011 CAFA Shared Task concerns the prediction of GO terms from the Molecular Function (MFO) and Biological Process (BPO) Ontologies (1). We approach this problem as a number of independent classification tasks, predicting for each protein whether one of the 385 most common GO terms from these ontologies applies for it. We participate in the primary, eukaryotic track of the CAFA task.

2. DATA PREPARATION

We use support vector machines for GO term prediction, so we need to build a dataset for training and testing the system. This dataset is constructed from the manually annotated Swiss-Prot Knowledgebase and contains all 164985 eukaryotic proteins. From this set, we leave aside the 35261 CAFA target proteins as a final test set. The remaining proteins are divided, for each of the predicted terms, into a training (50%), parameter optimization (25%) and test (25%) set, with the ratio of positive examples consistent across all sets.

A protein is considered positive for a GO term if that term has been annotated for that protein with an *experimental*, *traceable author statement* or *inferred by curator* evidence code. Since the absence of an evidence code may mean that that function has simply not yet been found for that protein, we include as negatives only proteins with at least one GO term with one of the mentioned evidence codes. These proteins have already been the subject of at least some research and should therefore be less likely to be unknown positives.

3. CLASSIFICATION MODEL

We develop a GO term classification system that combines several biological data sources and then evaluate the impact of text mining as additional data. We predict each term independently using an SVM with a fast linear kernel (2).

Blast2GO is a widely used functional annotation tool that can predict GO terms (3). We use the output of this rule-based system as data for training our classifier, in a combined meta-system approach. We use the precalculated Blast2GO annotations provided by SIMAP (Similarity Matrix of Proteins) (4). The Blast2GO predictions also form a baseline against which we compare our methods.

For additional features, we use Uniprot information on protein structures (domains, repeats, zinq fingers) and families (5). All tissues where the protein is known to be expressed are used as features, based on the UniGene database (6). If a protein is from one of the seven CAFA target species, we mark this as a feature.

4. TEXT MINING

Event extraction is a biomedical text mining approach designed to extract detailed information about protein interactions. It was popularized by the BioNLP'09 Shared Task on Event Extraction, where our text mining system had the best performance. Since then we have applied that system for extracting events from all the publicly available PubMed abstracts, creating a dataset of 19 million statements about protein and gene relations (7).

For GO term prediction, we convert, for each protein, all extracted statements describing it into features and evaluate their impact both alone and with the classification model described in Section 3.

5. RESULTS

To determine the overall performance of the different methods, we use F-score microaveraged over all predicted terms. This allows us to establish the relative performance of the methods when tested on the same datasets. As performance baselines we use the all-positive baseline, i.e. consider all proteins positive for all terms, and Blast2GO predictions. Especially on smaller classes lack of training data can cause machine learning to reduce performance, but since this can be detected during parameter optimization, we choose in these cases to fall back on the baseline Blast2GO prediction, resulting in a greater overall prediction improvement, shown in the *mixed* performance metric (Table 1).

	New Model	**Text Events**	**New Model + Text**
All-positive baseline	0.7%	0.7%	0.7%
Blast2GO baseline	47.7%	47.7%	47.7%
Classification	48.3%	9.4%	40.2%
Classification (mixed)	52.9%	47.7%	50.9%

Table 1: F-scores, microaveraged over the 385 predicted terms. *New Model* is described in Section 3.

The results indicate that our classification model was able to build on Blast2GO predictions and improve overall GO term prediction by around 10%. Text mining alone provided an F-score around 9.4%, clearly above the ~0% all-positive baseline, indicating that the extracted statements contain information usable for GO term prediction. However, combined with the full model, text mining reduced performance, perhaps indicating the presence of a lot of noise. The impact of such misleading signals could also have been amplified by the relatively low average number of positive examples in the training data for several classes. Our classification model improved performance most over the Blast2GO baseline (33–42 pp) on terms *growth*, *pathogenesis* and *translational elongation* from the BPO ontology.

6. CONCLUSIONS

We developed an SVM based classification model that using varied biological datasets can improve GO term prediction done with the popular Blast2GO software. We tested features based on text mining, showing that they can be used for GO term prediction. While event-type text mining is a potential new data source for GO term prediction, more work is needed to integrate it with other approaches. We will publish our software free for download and use under an open source license.

7. REFERENCES

1. http://biofunctionprediction.org/

2. Tsochantaridis I., Joachims T., Hofmann T. and Altun Y. 2005. Large margin methods for structured and interdependent output variables. Journal of Machine Learning Research (JMLR), 6(Sep):1453–1484.

3. Conesa A., Götz S., García-Gómez J. M., Terol J., Talón M. and Robles M. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 2005 21: 3674-3676

4. http://boincsimap.org/boincsimap/

5. Uniprot protein structure and family annotation – http://www.uniprot.org/docs/similar

6. UniGene – http://www.ncbi.nlm.nih.gov/unigene

7. Björne J., Ginter F., Pyysalo S., Tsujii J. and Salakoski T. (2010) Complex event extraction at PubMed scale. Bioinformatics. 2010 Jun 15;26(12):i382-90.