Predicting 400 GO Functions of Proteins

Part of the Proteome Analyst Suite of Tools

Roman Eisner*, Alona Fyshe, Russell Greiner, Paul Lu, Brandon Pearcy, Brett Poulin, Duane Szafron, David Wishart

*eisner@cs.ualberta.ca

www.cs.ualberta.ca/~bioinfo/PA

Introduction

- We predict over 400 molecular function categories from Gene Ontology (www.geneontology.org)
- We predict functions of proteins from sequence information
- Our technique is evaluated against experimentally verified data
- Contributions:

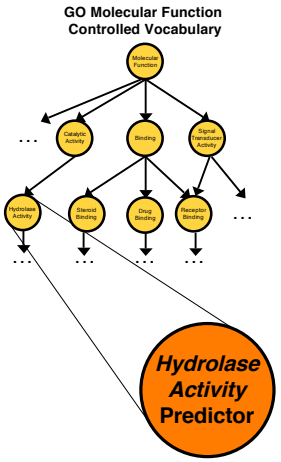
1. More **accurate** than BLAST, especially on remotely related proteins
2. Coverage of a **large ontology** with accurate predictions
3. Exploiting the **hierarchy** to increase accuracy and minimize computational complexity

Experimentally Consistent

- GO Term Predictors are trained, and evaluated against experimentally annotated proteins. No electronic annotations are used for evaluating our predictions.
- All accuracy statistics are obtained using 5-fold cross-validation.

References

1) The Gene Ontology Consortium [http://www.geneontology.org/]
2) ProtFun 2.2 Server [http://www.cbs.dtu.dk/services/ProtFun/]
3) Slim GO @ EBI [http://www.ebi.ac.uk/GOA/]
4) Proteome Analyst Server [http://www.cs.ualberta.ca/~bioinfo/PA]
5) GO Slim @ MGI [http://www.spatial.maine.edu/~mdolan/MGI_GO_Slim.html]



The Hydrolase Activity Predictor predicts whether a given protein is a hydrolase enzyme or not.

Figure 1

☆ *Creating Predictors for each GO Term*

- Machine-learned
- Binary (predict yes/no for each function term)
- Created for every GO function with at least 20 experimentally verified proteins (Total of 406)
- A Weighted Combination of:
  - Probabilistic Suffix Trees (PSTs)
  - PFAM with Support Vector Machines
  - Proteome Analyst (See Proteome Analyst Poster)

1 Predicting the Function of Proteins

- Goal: To find all functions of a query protein. Input is a protein sequence (Fasta format)
- We increase the predictive accuracy on those proteins that are similar to experimentally verified ones (Table 1). Here, similar means BLAST E-value ≤ 0.001
- Our predictors work significantly better for query proteins which do not have a good BLAST result against the set of experimentally verified proteins (Table 2)

Proteins with a good BLAST hit (≤ 1e-3)		
	Overall Precision	**Overall Recall**
BLAST (E-value ≤ 0.001)	77%	78%
Our Method	**78%**	**80%**
How often does this occur?		
60% of *D. melanogaster* proteins		
62% of *S. cerevisiae proteins*		

Table 1

Proteins with no good BLAST hit		
	Overall Precision	**Overall Recall**
BLAST (any hit accepted)	19%	20%
Our Method	**54%**	**31%**
How often does this occur?		
40% of *D. melanogaster* proteins		
38% of *S. cerevisiae proteins*		

Table 2

2 Large Ontology

- 406 categories of Molecular Function
- Allows for very specific and general predictions of function

GO Predictor	Size of Ontology
Our Ontology	406
ProtFun[2]	14
Slim-GO @ EBI[3]	30
Proteome Analyst[4]	12
GO Slim @ MGI[5]	13

3 Exploiting the Hierarchy

- To find the functions of a query protein, first BLAST against experimental data. When BLAST provides a good match (Table 1), we use the hit's annotations as a guide to which term predictors should be computed. This reduces runtime, without penalty to accuracy.
- Currently working on reducing the computational runtime of predicting function for those proteins which do not return a good BLAST result.
- Predictions of predictors are propagated upwards in the hierarchy to maintain consistency.
- To create classifiers (Figure 1), positive and negative training examples for each term predictor are selected to maintain consistency with the hierarchy. This increases the accuracy of each term predictor.