PA-GOSUB

Gene Ontology Molecular Function and Subcellular Localization Predictions for Model Organisms

Alona Fyshe*, Paul Lu, Duane Szafron, Russell Greiner, David S. Wishart, Brandon Pearcy, Brett Poulin, Roman Eisner, Danny Ngo, Nicholas Lamb
***alona@cs.ualberta.ca**

http://www.cs.ualberta.ca/~bioinfo/PA/GOSUB

What is PA-GOSUB?

- PA-GOSUB is an online database with
  - Gene Ontology molecular function (**GO**) predictions
  - Subcellular localization (**SUB**) predictions

  for proteomes of genetic models.
- PA-GOSUB has predictions for
  - > 24 organisms (see handout)
  - > 200,000 proteins
- GO and SUB predictions are provided by Proteome Analyst's (PA) Classifiers [2].
- PA's classifiers have high accuracy
  - Sub 81%-96%
  - GO 98%

How can I use PA-GOSUB?

- **SEARCH** - (Figure 2) Find proteins of interest using searches by
  - Organism
  - Keywords
  - Subcellular localization
  - Molecular function

  or by using
  - sequence similarity (BLAST).
- **DOWNLOAD** - Users can download entire model organism proteomes complete with GO and SUB predictions.

Figure 1: PA-GOSUB was created using the proteomes of model organisms and Proteome Analyst's GO and SUB classifiers.

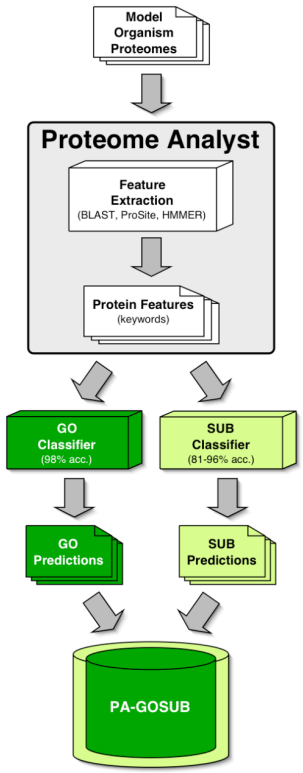
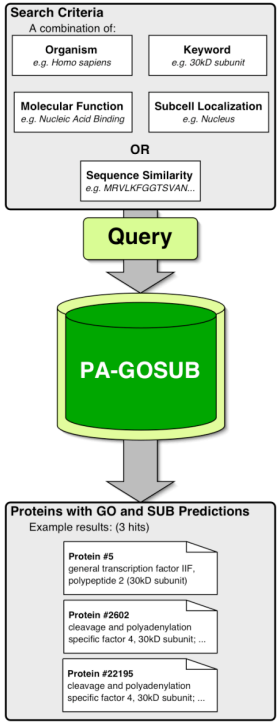


Figure 2: Users can query the PA-GOSUB database using several different criteria (ex. organism type & PA's classifier predictions or sequence similarity)



How was PA-GOSUB made?

(Figure 1)

- Proteomes are submitted to PA
- Sequences are BLASTed against SwissProt
- Keywords are extracted from the top BLAST hits to create features
- Features are input to PA's classifiers.
- Predicted molecular function and subcellular localization are made available in PA-GOSUB

PA-GOSUB Results

PA contributes to the annotation coverage of model organisms

Model Organism	Number of Proteins	GO MF		subcellular localization	
		GOA	PA-G	SP 42.7	PA-G
M. thermoauto-trophicum	1,868	497	1,250	157	1,100
B. subtilis	4,105	1,534	3,187	862	2,999
E. coli	4,353	3,524	3,772	2,167	3,627
P. falciparum	5,257	78	4,309	85	4,275
S. cerevisiae	6,195	3,017	5,049	2,024	4,978
D. melanogaster	16,371	1,535	12,924	1,246	12,869
C. elegans	21,821	1,459	14,379	933	14,297
A. thaliana	26,173	1,891	18,338	1,528	18,130
M. musculus	26,556	5,520	22,512	4,912	22,431
H. sapiens	27,954	8,230	23,064	7,136	22,978
Total	140,653	27,285	108,784	21,050	107,684

Table 1. Model Organisms and Annotation Coverage in PA-GOSUB

References

1. **Lu et al. 2005. PA-GOSUB: A Searchable Database of Model Organism Protein Sequences with Their Predicted GO Molecular Function and Subcellular Localization**
2. **Szafron et al. 2004. Proteome Analyst: Custom Predictions with Explanations in a Web-Based tool for High-Throughput Proteome Annotations. NAR 32: w365 - W371**