

PA-GOSUB:

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

Gene Ontology Molecular Function (GO) and Subcellular Localization (SUB) Predictions for Model Organisms

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



What is PA-GOSUB?

- PA-GOSUB¹ is a new online database of model organism proteomes containing predictions of:
 - GO molecular function (GO)
 - subcellular localization (SUB)
- 24 complete model organisms.
- Over 200 000 proteins.
- GO and SUB predictions are made by Proteome Analyst² (see Figure 2).
- Accuracy of predictors:
 - GO: 98%
 - SUB: 81-96%

How can I use PA-GOSUB?

You can use the high quality GO and SUB predictions to streamline annotation, as well as to enhance wetlab research. You can get the data in two different ways:

1. **Search.** You can find proteins of interest using PA-GOSUB's online search capability (see Figure 1). You can search by:
 - organism
 - keyword(s) (such as IDs)
 - subcellular localization(s)
 - molecular function(s)
 - sequence similarity
2. **Download.** You can download entire proteomes complete with GO and SUB predictions for any organism in the system.

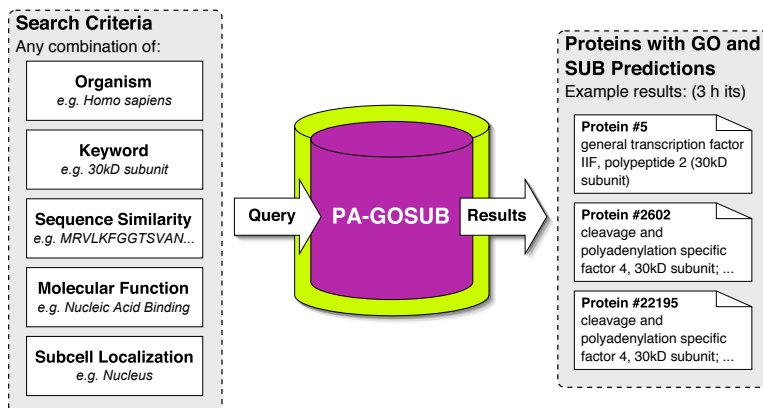


Figure 1. Users can query the PA-GOSUB database of model organism proteomes using a variety of criteria to retrieve the proteins of interest. Given a combination of search criteria, PA-GOSUB returns all proteins from an organism that match, including predicted molecular function and subcellular localization.

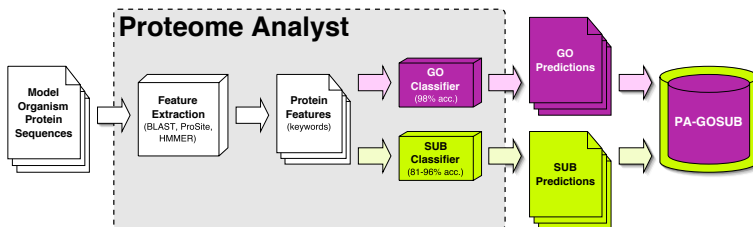


Figure 2. PA-GOSUB was created by processing model organism proteomes with Proteome Analyst². A variety of tools are used to extract features for each protein. The extracted features are inputted into PA's GO and SUB classifiers. The resulting predictions are compiled together to form PA-GOSUB.

Organisms in PA-GOSUB

- | | | |
|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Animal (5) <ul style="list-style-type: none">· Caenorhabditis elegans· Drosophila melanogaster· Homo sapiens· Mus musculus· Plasmodium falciparum | Gram-negative (10) <ul style="list-style-type: none">· Agrobacterium tumefaciens· Bartonella henselae· Bdellovibrio bacteriovorus· Chlamydia trachomatis· Coxiella burnetii· Escherichia coli· Hamophilus ducreyi· Helicobacter pylori· Thermus thermophilus· Vibrio vulnificus | Gram-positive (5) <ul style="list-style-type: none">· Bacillus subtilis· Deinococcus radiodurans· Mycobacterium tuberculosis· Mycoplasma genitalium· Mycoplasma pneumoniae |
| Archaea (2) <ul style="list-style-type: none">· M.thermoautotrophicum· Methanococcus jannaschii | Plant (1) <ul style="list-style-type: none">· Arabidopsis thaliana | |
| Fungi (1) <ul style="list-style-type: none">· Saccharomyces cerevisiae | | |

What's in PA-GOSUB?

- PA-GOSUB has analyzed proteomes covering a wide spectrum of life (see Organisms in PA-GOSUB below).
- We are continually expanding our database of organisms. If you would like to make any requests, please let us know.

How was PA-GOSUB made?

1. We submit model organism proteomes to Proteome Analyst² (see Figure 2).
2. PA extracts features (keywords) for each protein, such as:
 - SwissProt keywords from the most similar sequences (BLAST)
 - Pfam domains (HMMER)
 - sequence patterns (ProSite)
3. PA uses the extracted features as the inputs to the GO and SUB classifiers. PA's high quality classifiers output predictions for each protein.
4. The GO and SUB predictions are added to each protein's annotation, and are published 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. NAR 33: D147 - D153.
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.

