# Proteome Analyst: An Overview

Alona Fyshe*, Roman Eisner, Russell Greiner, Paul Lu, David Meeuwis,
Brett Poulin, Duane Szafron, David Wishart, Chris Upton
*alona@cs.ualberta.ca

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

## Introduction: Bridging the Gap

Current methods of genomic sequencing allow huge amounts of data to be produced quickly, while laboratory techniques for determining protein function and localization can take many times longer to complete. There is a real need to bridge this time gap; Proteome Analyst (PA) attempts to do that.

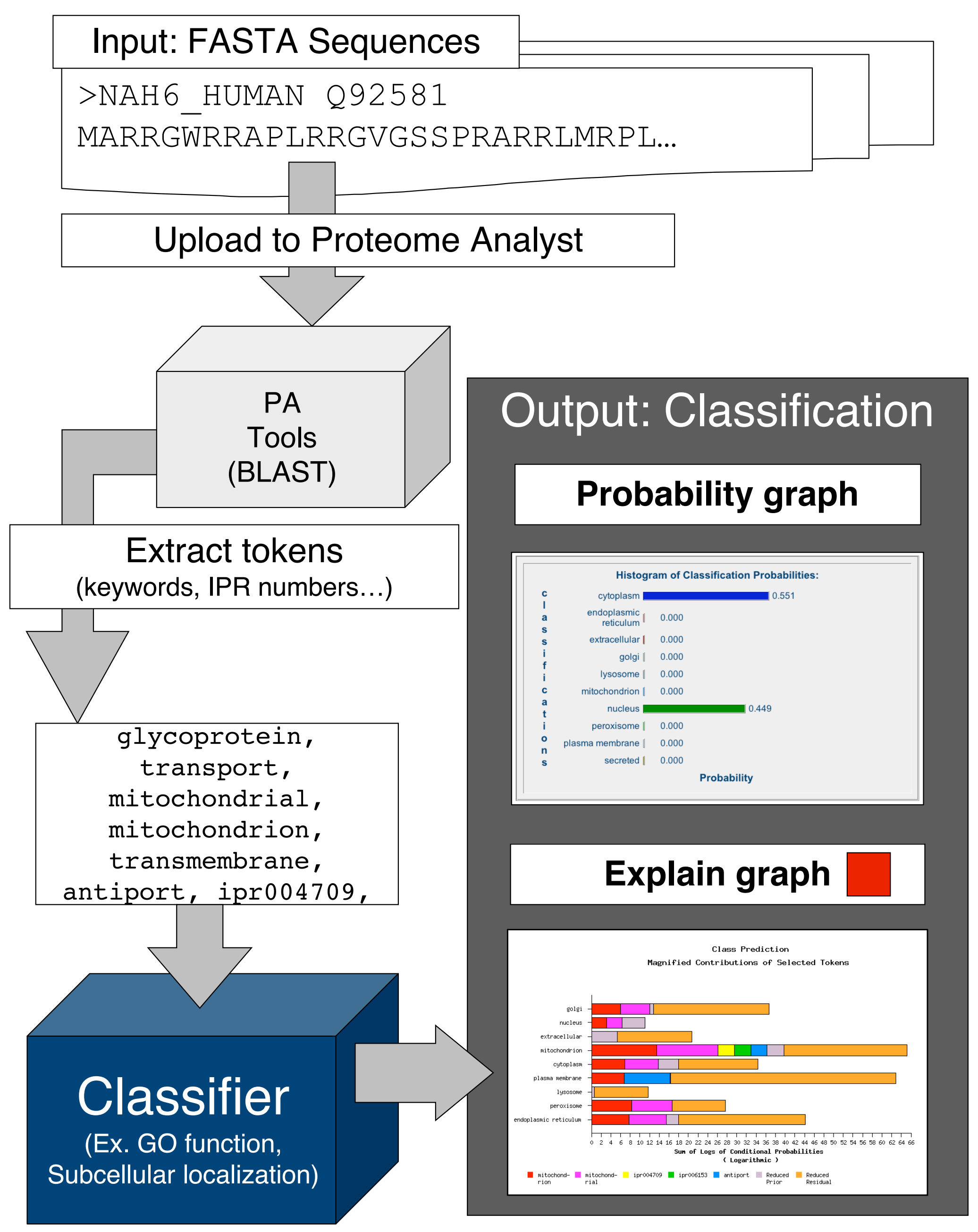
PA is a free web-based service for one-stop automatic high-throughput analysis that includes the most accurate subcellular predictors across the widest set of organisms ever published[1].

## What does PA do?

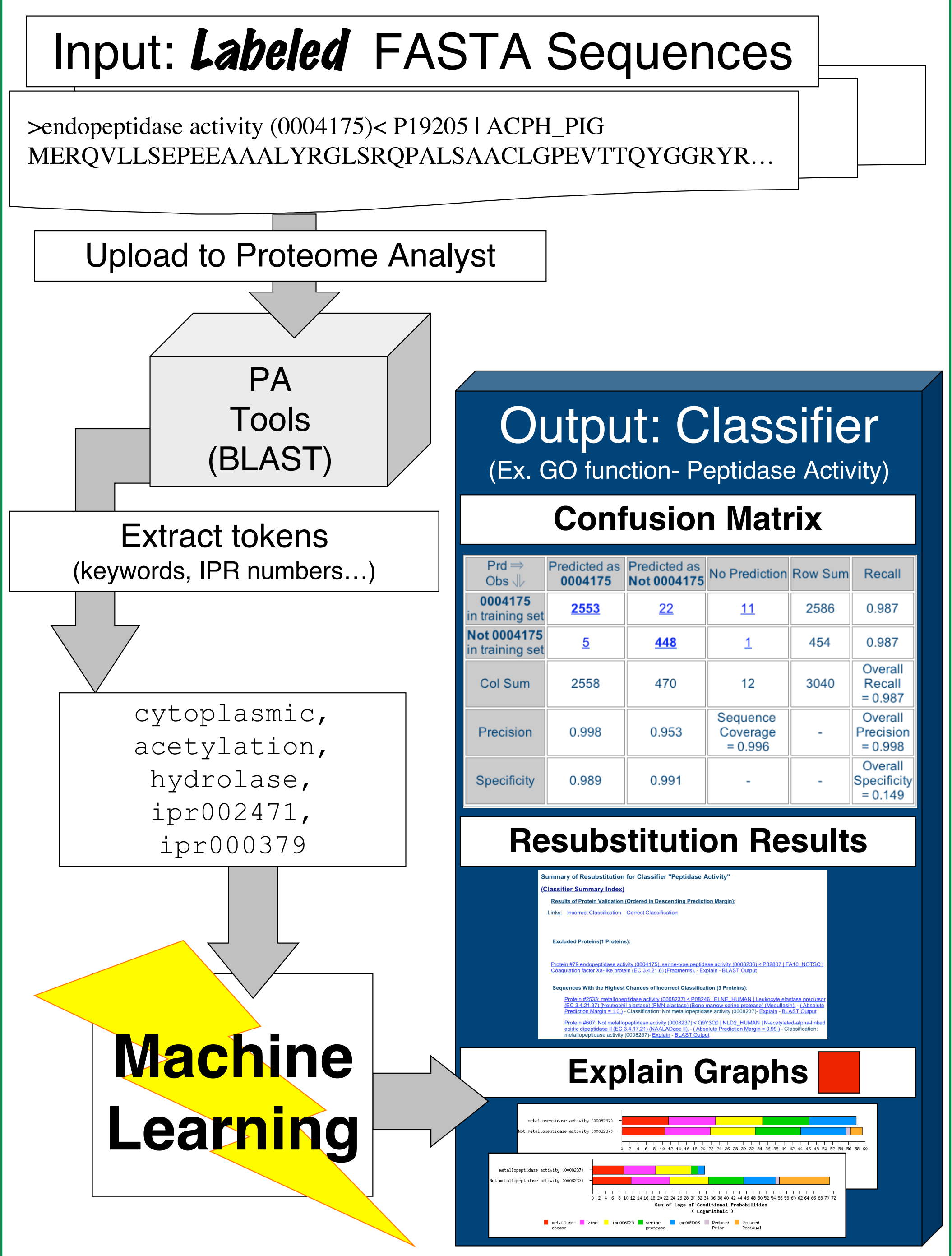
PA provides 2 main services:

- **Analysis ( ◆ )**
  - Upload sequences in fastA format
  - Process the sequences with tools (runs BLAST, Prosite)
  - Parse tokens from the tool's output
  - Use tokens to predict the class of the protein (Ex. Hydrolase Activity, Cytoplasm) using Machine Learning.
  - Provide an Explanation ( 🟥 ) for the prediction
- **Custom Classifer Creation ( ● )**
  - Upload ***Labeled*** sequences in fastA format
  - Process the sequences with tools (runs BLAST, Prosite)
  - Parse tokens from the tool's output and use them to detect similarities within classes using Machine Learning.
  - Use detected similarities to classify new proteins with unknown properties.

## ◆ Analysis



## ● Custom Classifier Creation


Input: Labeled FASTA Sequences
>endopeptidase activity (0004175)< P19205 | ACPH_PIG
MERQVLLSEPEEAAALYRGLSRQPALSAACLGPEVTTQYGGRYR…
Upload to Proteome Analyst
PA
Tools
(BLAST)
Extract tokens
(keywords, IPR numbers…)
cytoplasmic,
acetylation,
hydrolase,
ipr002471,
ipr000379
Machine
Learning
Output: Classifier
(Ex. GO function- Peptidase Activity)
Confusion Matrix
Resubstitution Results
Explain Graphs
metallopeptidase activity (0008237)
Not metallopeptidase activity (0008237)
Sum of Logs of Conditional Probabilities
( Logarithmic )
metalloprotease
zinc
ipr006025
serine protease
ipr009003
Reduced Prior
Reduced Residual

| Prd ⇒<br>Obs ⇓ | Predicted as 0004175 | Predicted as Not 0004175 | No Prediction | Row Sum | Recall |
|---|---|---|---|---|---|
| 0004175 in training set | 2553 | 22 | 11 | 2586 | 0.987 |
| Not 0004175 in training set | 5 | 448 | 1 | 454 | 0.987 |
| Col Sum | 2558 | 470 | 12 | 3040 | Overall Recall = 0.987 |
| Precision | 0.998 | 0.953 | Sequence Coverage = 0.996 | - | Overall Precision = 0.998 |
| Specificity | 0.989 | 0.991 | - | - | Overall Specificity = 0.149 |

Summary of Resubstitution for Classifier "Peptidase Activity"

(Classifier Summary Index)

Results of Protein Validation (Ordered in Descending Prediction Margin):

Links: Incorrect Classification Correct Classification

Excluded Proteins(1 Proteins):

Protein #79 endopeptidase activity (0004175), serine-type peptidase activity (0008236) < P82807 | FA10_NOTSC | Coagulation factor Xa-like protein (EC 3.4.21.6) (Fragments). - Explain - BLAST Output

Sequences With the Highest Chances of Incorrect Classification (3 Proteins):

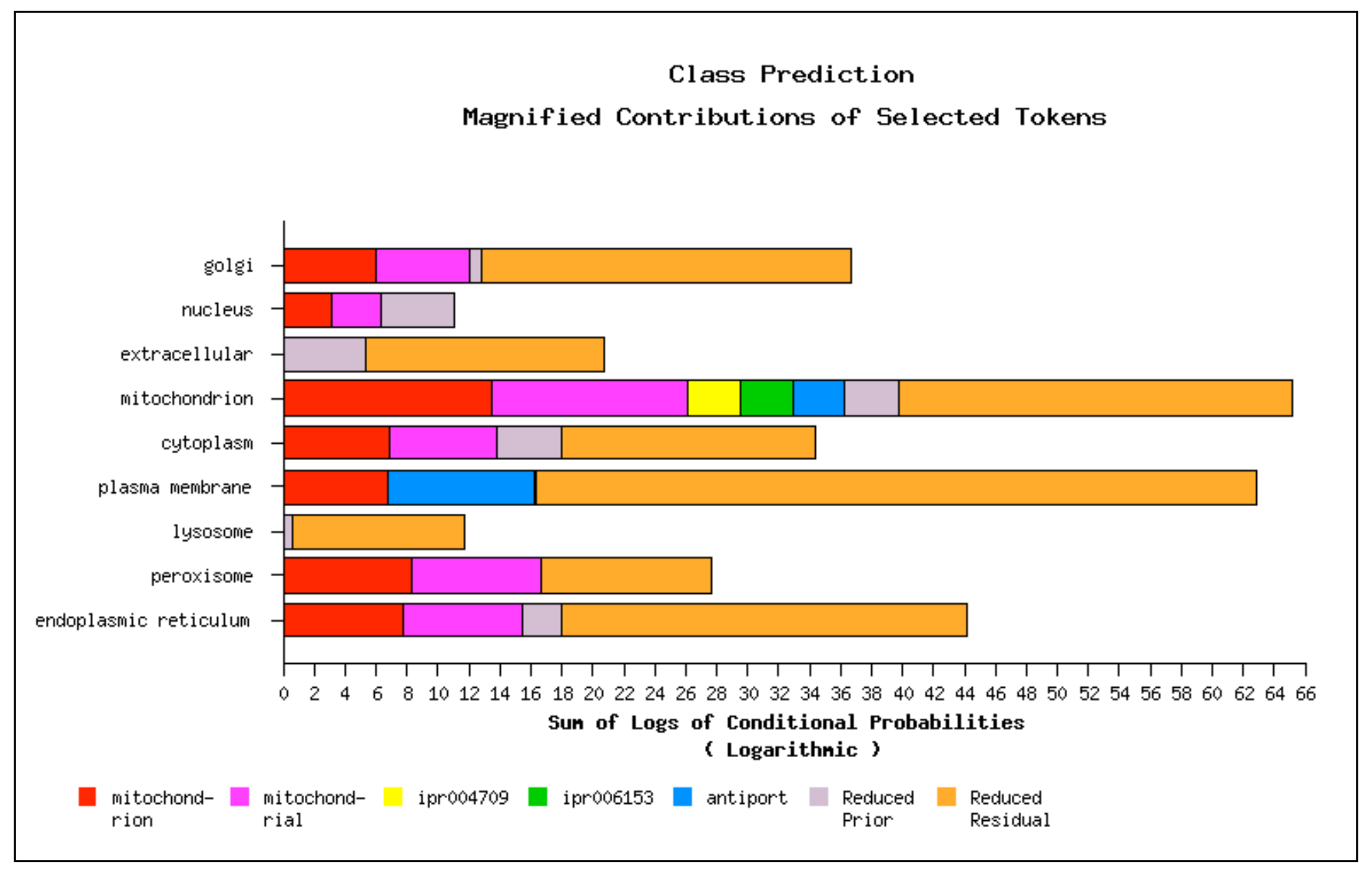
Protein #2533: metallopeptidase activity (0008237) < P08246 | ELNE_HUMAN | Leukocyte elastase precursor (EC 3.4.21.37) (Neutrophil elastase) (PMN elastase) (Bone marrow serine protease) (Medullasin). - ( Absolute Prediction Margin = 1.0 ) - Classification: Not metallopeptidase activity (0008237)- Explain - BLAST Output

Protein #607: Not metallopeptidase activity (0008237) < Q9Y3Q0 | NLD2_HUMAN | N-acetylated-alpha-linked acidic dipeptidase II (EC 3.4.17.21) (NAALADase II). - ( Absolute Prediction Margin = 0.99 ) - Classification: metallopeptidase activity (0008237)- Explain - BLAST Output

## 🟥 "Explain"

PA delivers *transparent* predictions; a mechanism, called "Explain", is provided that helps users understand why a classifier made a particular classification.

Example: NAH6_HUMAN classified by PA's Animal Subcellular Classifier



Note:

- IPR004709: Sodium/hydrogen exchanger subfamily
- IPR006153 Sodium/hydrogen exchanger

- This graph is on a logarithmic scale
- The color of a bar indicates the token, and it's length indicates that token's contribution to the class's probability.
- In this graph the tokens "mitochondrion" and "mitochondrial" contribute greatly to the class Mitochondrion.
- The token "antiport" actually contributes more to the class Plasma Membrane than Mitochondrion, but because only Mitochondrion has the tokens "IPR004709" and "IPR006153" it has a longer probability bar, thus Mitochondrion it is the predicted class for this protein.

## Explaining "Explain"

Explain will feature a textual analysis of the classification:

```
Your protein has a 83.4% chance of
belonging to the class "Mitochondrion".
The 3 features associated with your
protein that were the most important in
making this classification are
"mitochondrion", "mitochondrial" and
"ipr004709". Together these 3 tokens
account for 45.2% of "Mitochondrion"'s
total probability.
The second most likely class is "Plasma
Membrane" with a probability of 16.6%.
The same tokens "mitochondrion",
"mitochondrial" and "ipr004709" account
for 10.7% of "Plasma Membrane"'s total
probability.
```

## Gene Ontology - Function

PA recently finished training a new Gene Ontology (GO) Function classifier.

- 12 Classes
- 102,225 sequence training set
- Built using EBI's GO mapping & the SwissProt database

- Precision: 93%
- Recall: 97%

**Also see Proteome Analyst's Gene Ontology Poster**

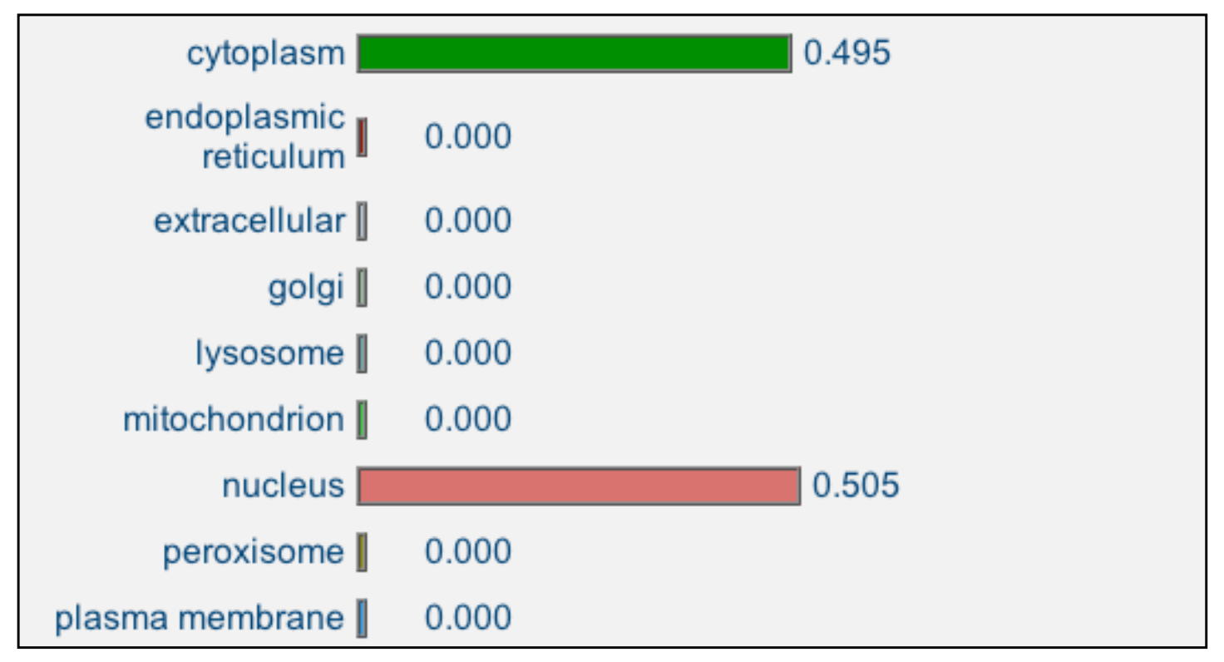
## What if…?

Proteome Analyst has the ability to change the tokens associated with a protein.

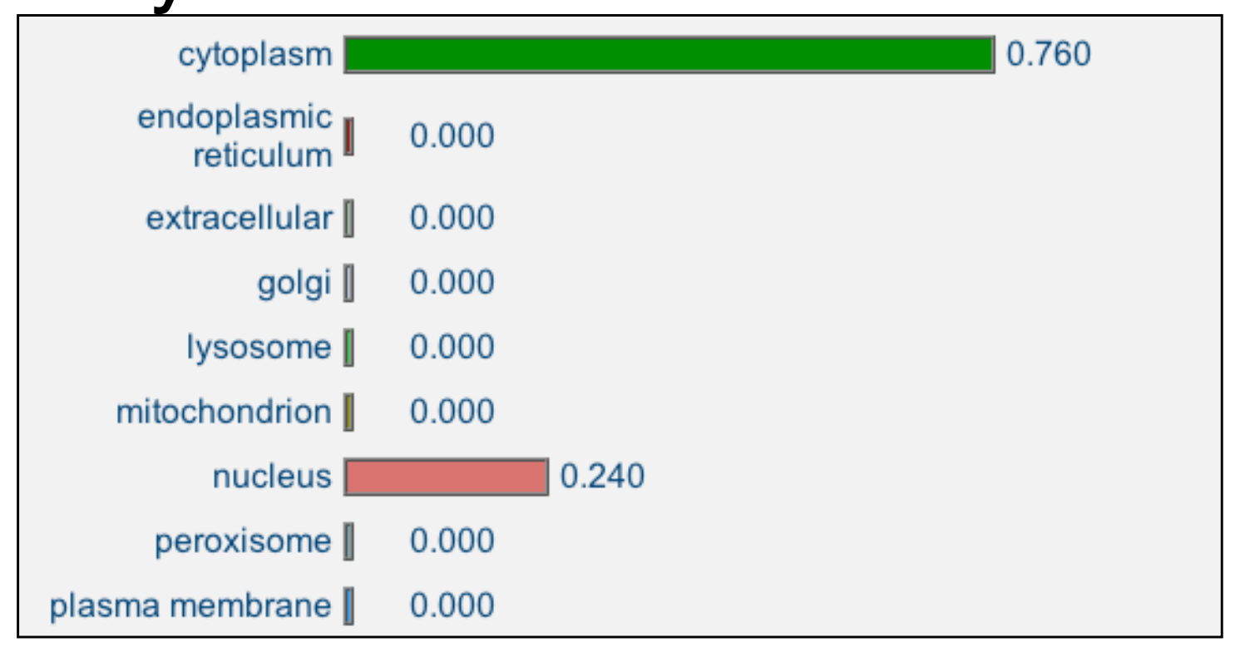
- Useful when features are determined experimentally, but are not in the tokens produced by PA's tools.

Example: Protein classified "nucleus" (50.5%), "cytoplasm" (49.5%)

Associated Tokens: repeat, nuclear, antigen, cytoplasmic, lectin, galectin, nuclear protein, ipr008985 ,ipr001079 , polymorphism



It is deduced experimentally that this protein is involved in apoptosis. A biologist can now add the token "apoptosis" to the protein's token list and re-classify.



This new token changes the classification to "cytoplasm" with probability 76.0%

References

1) Z. Lu, et al. Predicting Subcellular Localization of Proteins using Machine-Learned Classifiers, Bioinformatics 2004 20(4):547-556
2) D. Szafron, P. Lu, R. Greiner, D. Wishart, B. Poulin, R. Eisner, Z. Lu, J. Anvik, C. Macdonell, A. Fyshe, D. Meeuwis. Proteome Analyst: Custom Predictions with Explanations in a Web-based Tool for High-Throughput Proteome Annotations. Nucleic Acids Research, in press, July 2004.