Proteome Analyst: Machine learning for protein annotation

Introduction

PA annotates proteins with
- GO (Gene Ontology) molecular function
- Subcellular localization

Proteome Analyst (PA) is
- Publicly available
- High-throughput
- Web-based

and uses
- Sequence alignment
- Machine learning

PA has
- High accuracy
- Broad sequence coverage

What does PA do?

1) Analysis (Figure 1)
- Upload protein sequences
- Process the sequences with PA's tools (BLAST, Prosite)
- Use keywords from tool's output along with machine learning to predict properties
  - GO - ex. "Hydrolase Activity"
  - Subcellular localization - ex. "Cytoplasm"
- Explain reason for the classifier's prediction

2) Custom Classifier Creation (Figure 2)
- Upload **labeled** sequences
- Process the sequences with tools (BLAST, Prosite)
- Parse keywords from the tool's output and use them to detect similarities within protein classes using Machine Learning.
- Use similarities to classify new unannotated proteins

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

PENCE NSERC CRSNG

ALBERTA INGENUITY CENTRE FOR MACHINE LEARNING

CORE Sun microsystems

Figure 1: Analysis of a proteome using a classifier

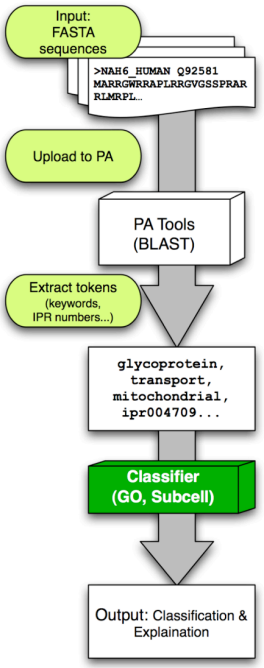


Figure 2: The creation of a new custom classifier

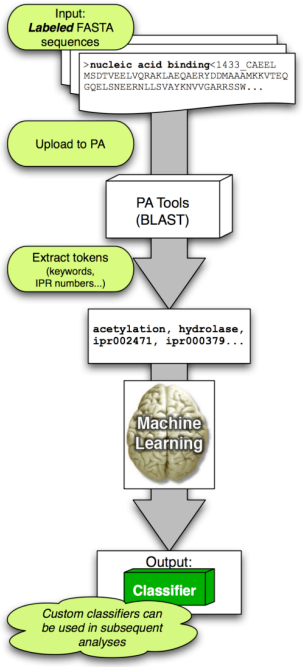
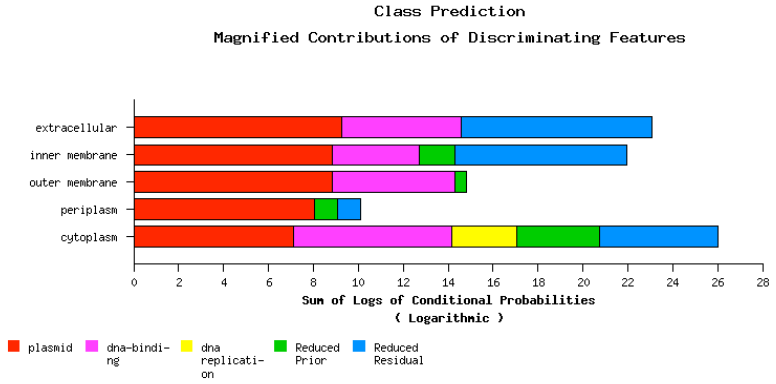


Figure 3: An explain graph, provided to help users understand why a classifier gave a particular classification.



PA's performance

- PA has 2 kinds of built in classifiers:
  - Gene Ontology (GO) Function
    - 98% accurate
    - >100,000 protein training set
  - Subcellular Localization
    - 5 organism types (animal, plant, fungi, gram negative, gram positive)
    - All 5 >90% accurate, except fungi which is 81% accurate

PA's Explain Feature

(Figure 3)
- Explain helps users understand how each keyword contributes to the final prediction of the classifier
- Each row of bars corresponds to the probability of a localization
- Logarithmic Scale
- *Cytoplasm* bar (Figure 3) is the longest because it represents the most probable class
- Each colored sub-bar corresponds to a keyword in the legend
- Each sub-bar's length indicates the corresponding token's contribution to the class's probability.
- In this graph the keyword "*DNA binding*" contributes the greatest probability to the class *cytoplasm*, and "*DNA replication*" contributes only to *cytoplasm* (*E. coli* example - no nucleus)