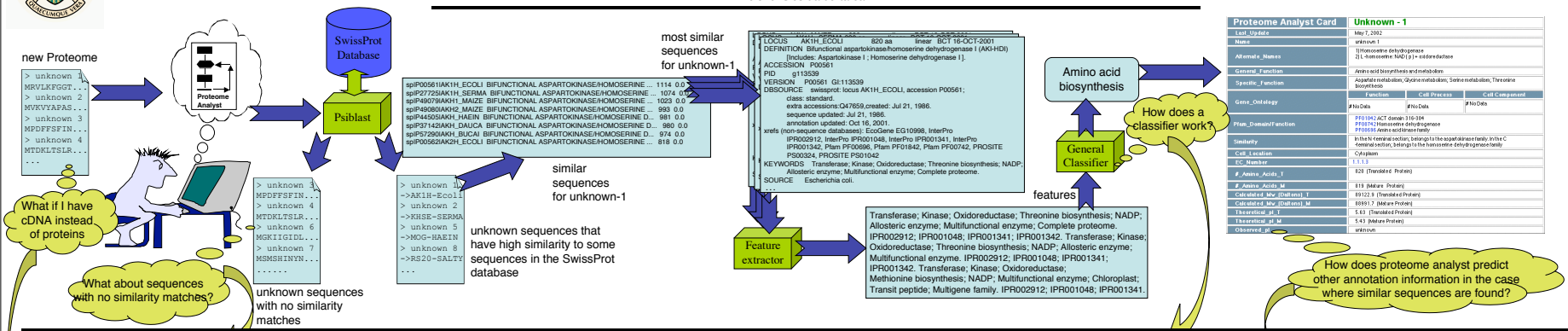
Proteome Analyst

PENCE Bioinformatics-1

Roman Eisner*, Brett Poulin, Duane Szafron, Paul Lu, Russell Greiner, David Wishart and Bahram Habibi-Nazhad
University of Alberta
*eisner@cs.ualberta.ca



Low Similarity Sequences! What do we do?

Analysis techniques are being developed based on motifs

Motif finder

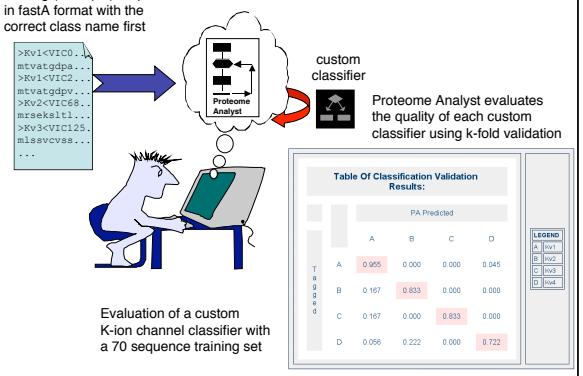
Secondary Structure P.

cDNA?

No problem.
Proteome Analyst will pre-translate to protein.

Creating a Custom Classifier

training (example) sequences in fastA format with the correct class name first



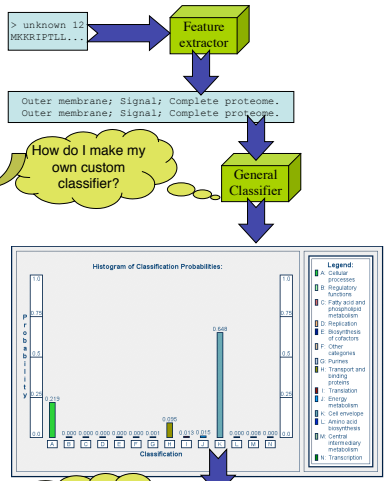
Proteome Analyst evaluates the quality of each custom classifier using k-fold validation

Classifiers

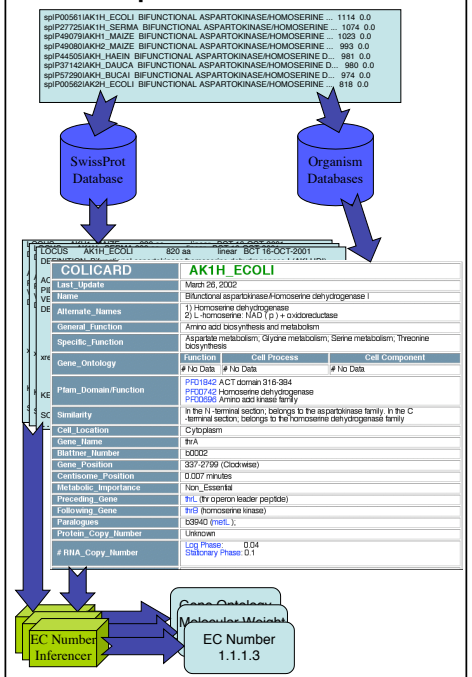
Classifier Algorithms (Machine Learning)

- Naive Bayes
- Tree-Augmented Naive Bayes (TAN)
- Neural Networks
- Decision Trees
- Support Vector Machine (SVM)
- Feature Selection - Wrapping

Classifier Decisions

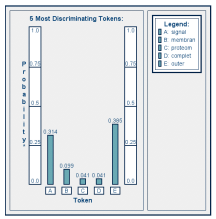


Predicting other annotation information when similar sequences are found



a set of prediction algorithms, one for each field -
some simple and some as complex as general function prediction

Classifier Decisions Explained



The features (tokens) in the sequences that were similar to the unknown sequence, that were most important in making the classification decision

Acknowledgements

Thanks to the 2001 summer students who worked on Proteome Analyst:
Cynthia Luk
Samer Nassar
Kevin McKee
Shannon Brooks

Thanks to the micro-biologists who used Proteome Analyst or provided other feedback or input:
Warren Gallin
Kathy Magor
Chris Upton

Proteome Analyst has received a generous equipment donation from Sun Microsystems.