The Proteome Analyst Suite of Automated Function Prediction Tools

Poulin B., Szafron D., Lu P., Greiner R., Wishart D.S., Eisner R., Fyshe A., Pearcy B., and Pireddu L.

Department of Computing Science, University of Alberta
221 Athabasca Hall, University of Alberta
Edmonton, Alberta, Canada T6G 2E8
{poulin,szafron,paullu,greiner,dwishart,eisner,alona,bpearcy,luca}@cs.ualberta.ca

Abstract

Proteome Analyst (PA) is a publicly available, high-throughput, web-based system for automatically predicting the function and properties of proteins. Biologists can use PA to predict, for example, the Gene Ontology (GO) molecular function and subcellular localization of a protein based on sequence information. Using sequence analysis tools and machine learning, PA gives high accuracy and broad coverage for both molecular function and subcellular localization predictions.

Introduction

Proteome Analyst (PA) is a publicly available, high-throughput, web-based system (http://www.cs.ualberta.ca/~bioinfo/PA/) for automatically predicting the function and properties of proteins. Biologists who have sequences for uncharacterized proteins can use PA (Szafron *et al.*, 2003; Szafron *et al.* 2004) to predict, for example, the Gene Ontology (GO) molecular function and subcellular localization of a protein based on sequence information. Using sequence analysis tools and machine learning, PA gives high accuracy and broad coverage for both molecular function and subcellular localization predictions.

PA predictions are made using a combination of sequence alignment and motif-based techniques. The information from these sources is combined using machine-learning classifiers. First, PA takes advantage of the availability of high-quality protein annotations by aligning each protein against the SwissProt database with BLAST. Annotated features (usually text keywords) of the nearest-neighbors of each protein (the highest scoring BLAST hits) are used to create a feature set for the unannotated protein. Motif-based techniques using pattern databases such as Pfam and PROSITE are then used to supplement this feature set. Based on the set of features, a machine-learned classifier predicts the classification – be it function, subcellular localization, etc. – of the protein. As the classifier has been trained on many examples of proteins that have been annotated with the function (or localization) of interest, it is able to make a prediction for the unannotated protein with high accuracy.

The PA prediction method offers both high accuracy and broad coverage across many protein classes. For Gene Ontology (GO) molecular function, PA achieves an overall accuracy of 98% across 12 high level GO classes on a cross-validation test set of over 100 000 proteins. For subcellular localization, PA is able to predict with more than 90% accuracy on cross-validation protein test sets from animals, plants, Gram-negative bacteria and Gram-positive bacteria (fungi results are slightly lower).

Although the predictions themselves are the key result of PA processing, an important contribution of Proteome Analyst system is a sophisticated explanation feature that shows why one prediction is chosen over another. The current PA classifier of choice is a naïve Bayes classifier, which is amenable to a graphical and interactive approach for explanations of its predictions. The transparency of these predictions increases the user's confidence in, and understanding of, PA. It also makes PA much more useful for identifying potential mistakes in predictions or previous annotations. This explanation facility is also extendable to other classifier technologies such as support vector machines (Poulin *et al.*, 2005).

In addition to the pre-trained GO and subcellular localization classifiers, users can create a custom PA classifier to predict a new property without any programming. The user simply provides a set of labeled training data (i.e. a set of examples, each with the correct classification label) to the system that PA uses to create the custom classifier. PA has been used, for example, to create custom classifiers for potassium-ion channel proteins and other general function ontologies.

The proteome analyst suite includes a variety of services.

- Proteome Analyst 2.0 – the main PA server (http://www.cs.ualberta.ca/~bioinfo/PA/). PA (Szafron *et al.*, 2003; Szafron *et al.* 2004) predicts high-level GO function, subcellular location, and any custom predictions (with explanations) for high-throughput proteome annotations.

- PA-SUB – the subcellular localization server (http://www.cs.ualberta.ca/~bioinfo/PA/Sub). PA-SUB (Lu *et al.*, 2004) predicts subcellular localization of proteins across a broad range of organisms with specialized classifiers for animals, plants, fungi, Gram-negative bacteria and Gram-positive bacteria. These classifiers are 81% accurate for fungi and 92–94% accurate for the other four categories.
- PA-GOSUB – Proteome Analyst: Gene Ontology Molecular Function and Subcellular Localization (http://www.cs.ualberta.ca/~bioinfo/PA/GOSUB). PA-GOSUB (Lu *et al.*, 2005) is a publicly available, web-based, searchable and downloadable database that contains the sequences, predicted GO molecular functions and predicted subcellular localizations of more than 200 000 proteins from 24 model organisms (and growing), covering the major kingdoms and phyla for which annotated proteomes exist. The PA-GOSUB database effectively expands the coverage of subcellular localization and GO function annotations that are available for the model organisms. More model organisms are being added to PA-GOSUB as their sequenced proteomes become available. PA-GOSUB can be used in three main ways. First, a researcher can browse the pre-computed PA-GOSUB annotations on a per-organism and per-protein basis using annotation-based and text-based filters. Second, a user can perform BLAST searches against the PA-GOSUB database and use the annotations from the homologs as simple predictors for the new sequences. Third, the whole of PA-GOSUB can be downloaded in either FASTA or comma-separated values (CSV) formats.

The PA tools and services are being constantly updated with more data, more predictions, better explanation capabilities and more efficient and robust service. A summary of some current work follows.

- Pathway Analyst – a pathway-centric prediction tool: Pathway Analyst will provide prediction facilities similar to Proteome Analyst specifically for the prediction of metabolic and signaling pathways. The organization of the tool will allow researchers who are interested in specific pathways to easily visualize the data and predictions.
- Specific Gene Ontology Predictions: Our current GO predictions focus on high-level (general) GO classes. This work will take advantage of the GO hierarchy and predict more specific function.
- Probabilistic Suffix Trees: Efficient Markov chains have shown promise in classifier many classes of protein sequences. We are incorporating these tools into our set of sequence-based features.
- Improvement of Prediction Accuracy: Prediction accuracy may be improved through a variety of methods. We are constantly improving pre-processing to give more and cleaner training data. We are also incorporating other state-of-the-art classification techniques such as support vector machines into the prediction and explanation facilities of PA.

The PA suite of tools provides accurate and high-throughput automated protein function prediction for biological research.

References

Lu P., Szafron D., Greiner R., Wishart D.S., Fyshe A., Pearcy B., Poulin B., Eisner R., Ngo D., and Lamb N. 2005. PA-GOSUB: A Searchable Database of Model Organism Protein Sequences With Their Predicted GO Molecular Function and Subcellular Localization. *Nucleic Acids Research* 33:D147-D152.

Lu Z., Szafron D., Greiner R., Lu P., Wishart D.S., Poulin B., Anvik J., Macdonell C., and Eisner R. 2004. Predicting Subcellular Localization of Proteins using Machine-Learned Classifiers. *Bioinformatics* 20(4):547 - 556.

Poulin B., Eisner R., Szafron D., Lu P., Greiner R., and Wishart D. 2005. Explaining Classification with Additive Evidence. Submitted for publication.

Szafron D., Lu P., Greiner R., Wishart D.S., Poulin B., Eisner R., Lu Z., Anvik J., Macdonell C., Fyshe A., and Meeuwis, D. 2004. Proteome Analyst: Custom Predictions with Explanations in a Web-based Tool for High-throughput Proteome Annotations. *Nucleic Acids Research* 32:W365-W371.

Szafron D., Lu P., Greiner R., Wishart D., Lu Z., Poulin B., Eisner R., Anvik J. and Macdonell C. Proteome Analyst - Transparent High-throughput Protein Annotation: Function, Localization and Custom Predictors. *International Conference on Machine Learning Workshop on Machine Learning in Bioinformatics (ICML Workshop - Bioinformatics)*. August 2003, Washington, U.S.A., pp. 2-10.