PROJECT AIMS

MS-based proteomics typically yields thousands of protein identifications. We have developed a software tool, STRAP (Software Tool for Researching Annotations of Proteins), which automates interpretation of protein lists generated from high-throughput proteomics experiments.

BACKGROUND

After proteins in a proteomics experiment are identified by mass spectrometry (or any other method), information about these proteins must be collected, organized, and interpreted to reach a conclusion. Collecting this data often requires many hours of searching online protein databases. The subsequent organization and interpretation of these protein lists is challenging due to a lack of a way to conceptualize and visualize the extensive protein information. Software must be used to automate this annotation process in order to accelerate extensive protein information.

Where Do Protein Lists Come From?

There are several types of tools for large-scale proteomics data analysis:

- Lab Information Management System (LIMS)-type software, such as Genelincs Proteus, compile data from heterogeneous sources, and perform analysis.
- Server-based Software, such as the Institute for Systems Biology’s Trans-Proteomic Pipeline (TPP) (http://systemsbiology.net/), allow its users to process data, but require setting up a web server.
- PC-based packages, such as ProteinCenter ($$$), Scaffold ($$$), and web applications, such as PPI from the Institute for Systems Biology, allow for analysis without a complicated setup.

PROBLEMS WITH TODAY’S TOOLS

Today’s tools help with data interpretations, but are limited because:

- It is difficult to interpret large data sets from a holistic perspective due to limited protein annotation.
- Commercial packages for protein annotation (e.g., ProteinCenter and Scaffold) are costly, use proprietary formats and do not necessarily present manageable annotation data tables.
- Few free programs can annotate protein lists or show trends within and between protein lists.

METHODS

STRAP is written in C# and was developed in the Microsoft Visual C# 2008 environment. Additionally, it uses the open-source ZeigGraph (http://zegigraph.org) and 3D Pie Chart libraries to create charts (http://www.codeproject.com/KB/graphical/piepiechart.aspx). STRAP was designed to be intuitive and easy to implement in a common PC-based laboratory. Users who are familiar with Microsoft Windows will be comfortable using STRAP. In order to use STRAP data with Gaggle, FireCore (the Gaggle Toolkit for Firefox) must be installed. The data set displayed in the screenshots here is from a recent lymphoma biomarker study (Romesser, et al., 2009).

STRAP OVERVIEW AND SCREENSHOTS

Figure 1. Protein Lists are generally produced by submitting raw mass spectrometric data to search engines; the search results are then ready for further analysis by, e.g., the TPP.

Figure 2. Schematic representation of STRAP functionality, including data input and output. STRAP reads protein lists in UniProt format obtained from several file types, and then gathers protein GO-term annotation data from public databases. These annotations can be edited, saved to disk to disk, or exported to Gaggle. Additionally, STRAP can visualize these GO annotations in tabular, pie chart, or bar graph formats to aid in interpretation and differential comparison of multiple data sets.

Figure 3. The STRAP protein annotation table, complete with gene ontologies. Multiple datasets can be opened at once in a multi-threaded fashion. Additionally, gene ontology annotations can be manually edited.

Figure 4. This pie chart was generated from the gene ontology terms of a set of 22 proteins. Each slice represents a high-level category of gene ontology (GO) terms. In order from left to right, the pie charts correspond to biological process, cellular component, and molecular function GO terms.

Figure 5. This bar graph compares the amount of biological process GO term annotations between proteins in three different samples.

Figure 6. The Gene Ontology Term Browser. The browser presents all GO terms associated with a particular protein entity, as well as each GO term’s complete lineage.

STRAP’S FEATURES

STRAP is an easy-to-use open-source C# application with the following features:

- Automates the protein annotation and GO-term visualization process that is otherwise extremely laborious when done manually.
- Provides pie and bar charts of gene ontology terms to aid annotation interpretation.
- Takes input from several popular formats (Mascot, TPP)
- Can communicate with Gaggle’s geese for further interpretation by submitting data sets for, e.g., a KEGG database search

CURRENT AND FUTURE WORK

- Add the capability to search databases other than the UniProtKB (e.g., the NCBIhm and PI databases).
- Account for post-translational modifications
- Extrapolate gene ontology annotations for proteins with incomplete annotations

ACKNOWLEDGMENTS

- NIH/NHLBI contract N01 HV28178
- NIH/NCCR grant P41 RR10888
- Special thanks to the members of the CPC and MSR