Software Tool for Researching Annotations of Proteins (STRAP): Open-Source Software for Protein Annotation and Data Visualization

Vivek Bhatia, David H. Perlman, Catherine E. Costello, Mark E. McComb

Cardiovascular Proteomics Center, Center for Biomedical Mass Spectrometry,
Boston University School of Medicine, Boston, MA 02118.

Introduction: MS-based proteomics may yield thousands of protein identifications per experiment. Post-identification, information about these proteins must be collected, organized, and interpreted to fully understand the results. Amassing this data requires a laborious search of multiple online databases (e.g. protein, genomic, relational). Even after one compiles this information, organizing and interpreting the results of this search is challenging due to a lack of convenient means to visualize the vast amount of protein information at hand. Bioinformatics approaches must be used to automate annotation and interpretation of results in order to accelerate the pace of research. Here we present an automated way of performing such meta-analyses using our open-source software program, the Software Tool for Rapid Annotation of Proteins (STRAP).

Methods: STRAP is an application with a user-friendly graphical user interface (GUI). STRAP is written in C# and runs on Microsoft Windows PCs with Windows XP or higher with version 3.5 of the Microsoft .NET Framework. C# was chosen due to its clean syntax, object-oriented nature, and productive programming environment. The code is written such that any parsed information is imported into an object data model. This data can be queried for alternate applications; this feature makes STRAP easily extensible. Gene ontology visualization is available through custom bar and pie charts created with the ZedGraph and PieChart3D libraries. The program currently pulls information from a web-hosted UniProt Knowledgebase database as its main source of information. Additional databases may be specified.

Results: STRAP is an easy-to-use, open-source application that helps automate the protein annotation process. STRAP allows collection and annotation of information about the proteins in a data set. First, it imports protein lists from several file formats. It currently supports import from Mascot DAT, protXML (Trans-Proteomics Pipeline, TPP, ISB) and ASCII plain text files consisting of protein lists. STRAP then downloads information about each protein from several online databases, focusing on information from the UniProt Knowledgebase database. STRAP then compiles all of the protein annotation information and displays it in a table. This information includes name, ID, protein length and gene ontology (GO) terms.

From an annotated list of proteins, STRAP renders various gene ontology charts to aid data interpretation and thus dramatically reduces the effort required to interpret cumbersome annotation tables. STRAP can generate a pie chart to represent a data set's composition of GO terms, which is useful for easily seeing a global overview of all the proteins' characteristics. Additionally, STRAP can create bar charts that allow comparison of multiple GO annotations from several samples. The tables and the charts can be exported into other software for presentation and saved for later use. Output also includes creation of searchable FASTA format databases for iterative search schemas. Additionally, in support of the TPP, output tables may be exported to Gaggle (ISB) for use with other proteomics tools on the web. Charts and tables derived by STRAP from complex data sets generated during completed and current proteomics experiments will be used to demonstrate the features of STRAP.

Conclusions: STRAP is user-friendly, open-source software that automates the protein annotation and GO-term visualization process that is otherwise extremely laborious when done manually. It can read protein lists from a variety of formats, including Mascot and TPP search results, and then annotate these lists using the online UniProtKB database. From an annotated list of proteins, it can generate various GO term graphs and charts to aid data interpretation and thus expedite proteomic data analysis. This PC-based software allows researchers to rapidly parse and annotate large sets of proteins. STRAP and its C# source code are available under the lesser GNU public license (LGPL) at http://sourceforge.net/projects/cpctools/.

Acknowledgements: This research was funded by NIH-NHLBI contract N01 HV28178 and NIH-NCRR grant P41 RR10888.

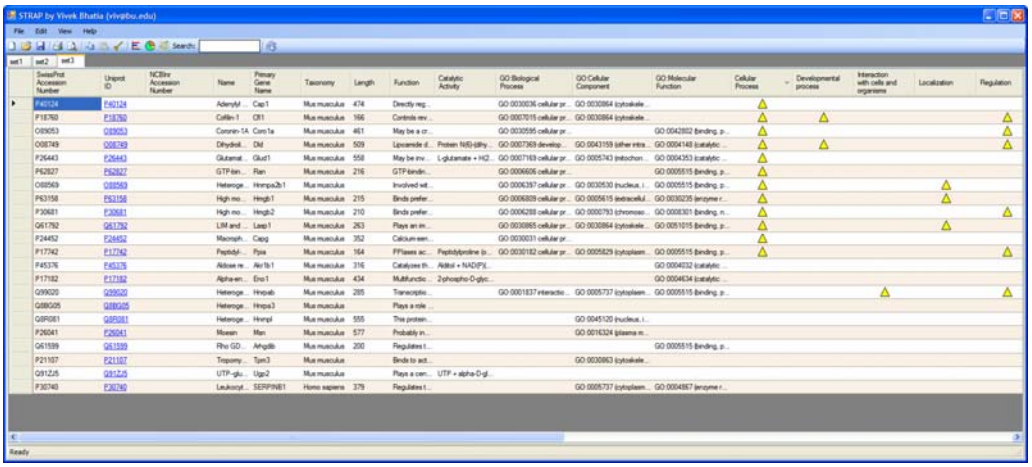

Figure 1. The STRAP protein annotation table, complete with gene ontologies. Multiple datasets can be opened at once in a multi-threaded fashion. Comparison is afforded accross multiple proteins and GO-functions. Gene ontology annotations can also be manually edited.

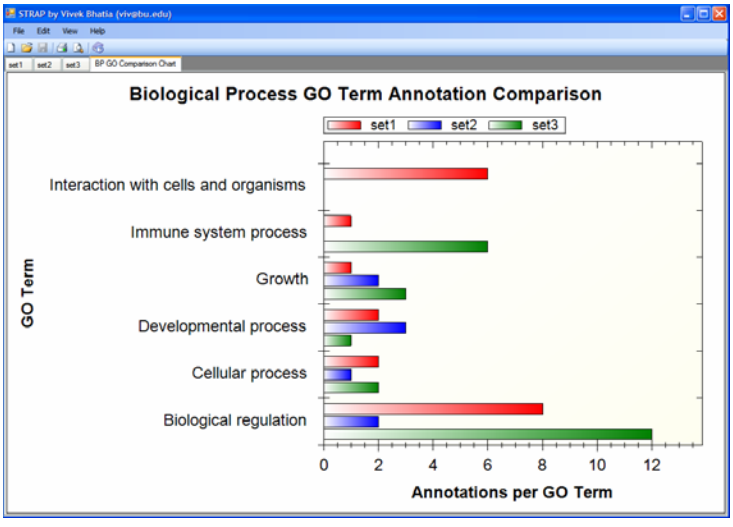

Figure 2. Comparison between multiple samples: bar graphical representation illustrating the different biological process gene ontology (GO) term annotations between proteins in three different samples.