

## Technical Notes

# Software Tool for Researching Annotations of Proteins: Open-Source Protein Annotation Software with Data Visualization

Vivek N. Bhatia, David H. Perlman, Catherine E. Costello, and Mark E. McComb\*

Cardiovascular Proteomics Center, Boston University School of Medicine, 670 Albany Street, Room 504, Boston, Massachusetts 02118

In order that biological meaning may be derived and testable hypotheses may be built from proteomics experiments, assignments of proteins identified by mass spectrometry or other techniques must be supplemented with additional notation, such as information on known protein functions, protein–protein interactions, or biological pathway associations. Collecting, organizing, and interpreting this data often requires the input of experts in the biological field of study, in addition to the time-consuming search for and compilation of information from online protein databases. Furthermore, visualizing this bulk of information can be challenging due to the limited availability of easy-to-use and freely available tools for this process. In response to these constraints, we have undertaken the design of software to automate annotation and visualization of proteomics data in order to accelerate the pace of research. Here we present the Software Tool for Researching Annotations of Proteins (STRAP), a user-friendly, open-source C# application. STRAP automatically obtains gene ontology (GO) terms associated with proteins in a proteomics results ID list using the freely accessible UniProtKB and EBI GOA databases. Summarized in an easy-to-navigate tabular format, STRAP results include meta-information on the protein in addition to complementary GO terminology. Additionally, this information can be edited by the user so that in-house expertise on particular proteins may be integrated into the larger data set. STRAP provides a sortable tabular view for all terms, as well as graphical representations of GO-term association data in pie charts (biological process, cellular component, and molecular function) and bar charts (cross comparison of sample sets) to aid in the interpretation of large data sets and differential analyses experiments. Furthermore, proteins of interest may be exported as a unique FASTA-formatted file to allow for customizable re-searching of mass spectrometry data, and gene names corresponding to the proteins in the lists may be encoded in the Gaggie microformat for further characterization, including pathway analysis. STRAP, a

tutorial, and the C# source code are freely available from <http://cpctools.sourceforge.net>.

Most mass spectrometry based proteomics experiments are designed to address the question, “What are the differences between the observed sets of expressed proteins in normal versus diseased or modified states?” Typically, proteins are isolated from cells, tissues, or biological fluids obtained from host organisms experiencing a challenged state relative to a control. These proteins are then identified and characterized by analytical techniques such as mass spectrometry. The results often yield lengthy lists of differentially expressed proteins, with tens to thousands of entries. Deriving meaning from these lists of proteins is quite challenging, but is necessary, if biological conclusions are to be drawn from the experiment. Interpretation of this data begins with annotation of the protein lists with information such as known protein function, followed by analyses of whether these attributes display meaningful trends. In order to obtain these protein annotations, knowledge must be gleaned from experts in the field, through an exhaustive search of the literature, or through the consultation of online protein databases. Accomplishing this process manually is extremely laborious and scales in time and effort linearly with the size of the protein lists such that it is feasible with only the smallest of data sets. Clearly, there is a widespread and growing need for software to facilitate this process.

To address these requirements, a number of genomics and proteomics software applications have been proposed with varying degrees of utility and success. While not all-inclusive, examples of gene-based and GO-related software may be readily found within the tools section of the Gene Ontology Web site ([www.geneontology.org](http://www.geneontology.org)),<sup>1</sup> and mass spectrometry and proteomics based tools may be found at ProteomeCommons ([proteomecommons.org/tools-browse.jsp](http://proteomecommons.org/tools-browse.jsp)).<sup>2</sup> One goal of such software is to retrieve

\* To whom correspondence should be addressed. Phone: (617) 638-4280. E-mail: [mccomb@bu.edu](mailto:mccomb@bu.edu).

(1) Ashburner, M.; Ball, C. A.; Blake, J. A.; Botstein, D.; Butler, H.; Cherry, J. M.; Davis, A. P.; Dolinski, K.; Dwight, S. S.; Eppig, J. T.; Harris, M. A.; Hill, D. P.; Issel-Tarver, L.; Kasarskis, A.; Lewis, S.; Matese, J. C.; Richardson, J. E.; Ringwald, M.; Rubin, G. M.; Sherlock, G. *Nat. Genet.* 2000, 25, 25–29.

(2) Falkner, J. A.; Falkner, J. W.; Andrews, P. C. *Bioinformatics* 2006, 22, 632–633.

and collate information from online databases and to provide a straightforward way to interpret proteomics data. For functional annotation, the most mature and widely used online repository is the Gene Ontology database, the product of a project to establish a standardized language and hierarchical framework for gene and gene-product annotation. In the field of proteomics, many protein information repositories have utilized the GO annotation nomenclature, including the highly annotated and curated UniProtKB protein knowledgebase ([www.uniprot.org](http://www.uniprot.org)),<sup>3</sup> an online repository of protein sequence, annotation, and functional information.

Specific software tools have been created for the purpose of functionally annotating gene and gene product lists. Among the commercial packages, the most popular for proteomics annotation are ProteinCenter, available from Proxeon ([www.proxeon.com](http://www.proxeon.com)), and Scaffold, from Proteome Software ([www.proteomesoftware.com](http://www.proteomesoftware.com)). These commercial packages are, unfortunately, quite costly and this property limits their adoption, particularly by academic laboratories. Meanwhile, there are examples of free tools which are capable of browsing/parsing or annotating entire protein lists at once, such as g:Profiler,<sup>4</sup> PANDORA,<sup>5</sup> and DAS.<sup>6</sup> Their annotation lists, however, are general and lack the granularity required to show detailed trends within or between protein lists. Recently released by the Institute for Systems Biology, PIPE: Protein Information Property Explorer,<sup>7</sup> and another recently developed server based application, PIKE: Protein Information and Knowledge Extractor ([prote.cnb.uam.es:8080/pike](http://prote.cnb.uam.es:8080/pike)),<sup>8</sup> have more extensive and comprehensive protein list annotation capabilities. Nevertheless, the complicated organization of GO terms in Scaffold, ProteinCenter, PIPE, and PIKE makes it difficult to quickly interpret the biological meaning of a protein set or a comparative proteomics data set. Additionally, none of these programs allows the user to manually edit GO annotations within a data set. This is critical as the extent and validity of GO annotations are currently limited. Furthermore, a main focus of proteomics experiments is to investigate the differential expression of proteins between pairs or among groups of samples. For this reason, it is important for any protein annotation software to have capabilities to compare protein sets. Scaffold is restricted to tabular protein set comparison, while PEAK and PIPE lack such abilities altogether. Although ProteinCenter does have extensive data set comparison capabilities, it requires the user to upload complete data sets to the ProteinCenter server as part of its web-based subscription package. This loss of control over valuable proprietary data may discourage some users.

While there is no question as to the utility of existing tools, there is no single tool that is open-source and easy to use, yet capable of automatically annotating large protein lists and visually rendering the annotations such that large scale proteomics data

sets can be compared. STRAP was designed to meet these specific needs while being a freely available tool that runs locally on the user's computer. It uses dedicated columns for GO terms, but the global presentation of the category names has been streamlined for ease of interpretation and manual GO term annotation.

## EXPERIMENTAL SECTION

The Software Tool for Researching Annotations of Proteins (STRAP) was written in C# and was developed in the Visual C# 2008 environment. Additionally, it uses the open-source ZedGraph ([zedgraph.org](http://zedgraph.org)) and 3D Pie Chart ([www.codeproject.com/KB/graphics/julianpiechart.aspx](http://www.codeproject.com/KB/graphics/julianpiechart.aspx)) libraries to create charts. STRAP was designed to be intuitive and easy to implement in a common PC-based laboratory such that users would be comfortable using STRAP. Although Windows-based machines are ubiquitous in the proteomics world, STRAP may also be used on Apple computers with the use of Boot Camp (included with Mac OS) or virtualization software such as Sun VirtualBox (free) or Parallels Desktop (commercial). STRAP data can be exported to the Gagggle<sup>9</sup> network for analysis with other tools provided that FireGoose (the Gagggle Toolbar for Firefox)<sup>10</sup> is installed on the host PC. The STRAP installer, a tutorial and C# source code can be obtained directly from <http://cpctools.sourceforge.net>.

In order to demonstrate STRAP's capabilities, we used data from a recent lymphoma biomarker study.<sup>11</sup> This example data set, which is included with the STRAP installer, is composed of three protein sets reflecting the differentially expressed proteomes of three B-cell proliferative states.

## RESULTS

**Overview of STRAP and Accepted File Formats.** STRAP provides proteomics researchers with an easy way to obtain, group, and compare protein annotations for lists of proteins. STRAP begins with the input of protein lists from a proteomics experiment. For example, such an experiment may begin with liquid chromatography–tandem mass spectrometry (LC–MS/MS) and end with a protein ID list following a standard database search using a search engine such as Mascot ([www.matrix-science.com](http://www.matrix-science.com)).<sup>12</sup> For initial dissemination, we chose several input and output formats for STRAP which are summarized schematically in Figure 1. Input files for STRAP must contain protein entry or accession numbers in the UniProt Knowledgebase ([www.uniprot.org](http://www.uniprot.org)) format (e.g., HBB\_HUMAN or P002833). These lists may be in any of three currently supported formats: plain text lists of proteins (ASCII format), Mascot database search results (.dat format), or protXML results (XML format) obtained from Protein Prophet<sup>13</sup> using the Trans Proteomic Pipeline (TPP)<sup>14,15</sup> ([tools.proteomecenter.org/software.php](http://tools.proteomecenter.org/software.php)). Mascot and TPP file formats were chosen because of their popularity in proteomics, but the text file format gives STRAP the flexibility to be integrated into

(3) Apweiler, R.; Bairoch, A.; Wu, C. H.; Barker, W. C.; Boeckmann, B.; Ferro, S.; Gasteiger, E.; Huang, H.; Lopez, R.; Magrane, M.; Martin, M. J.; Natale, D. A.; O'Donovan, C.; Redaschi, N.; Yeh, L.-S. L. *Nucleic Acids Res.* **2004**, *32*, D115–119.

(4) Reimand, J.; Kull, M.; Peterson, H.; Hansen, J.; Vilo, J. *Nucleic Acids Res.* **2007**, *35*, W193–200.

(5) Kaplan, N.; Vaaknin, A.; Liniel, M. *Nucleic Acids Res.* **2003**, *31*, 5617–5626.

(6) Olason, P. I. *Nucleic Acids Res.* **2005**, *33*, W468–470.

(7) Ramos, H.; Shannon, P.; Aebersold, R. *Bioinformatics* **2008**, *24*, 2110–2111.

(8) Rodriguez-Perez, M. A.; Medina-Aunon, A.; Encarnacion-Guevara, S. M.; Bernal-Silvia, S.; Barrera-Saldana, H.; Albar-Ramirez, J. P. *Clin. Transl. Oncol.* **2008**, *10*, 604–617.

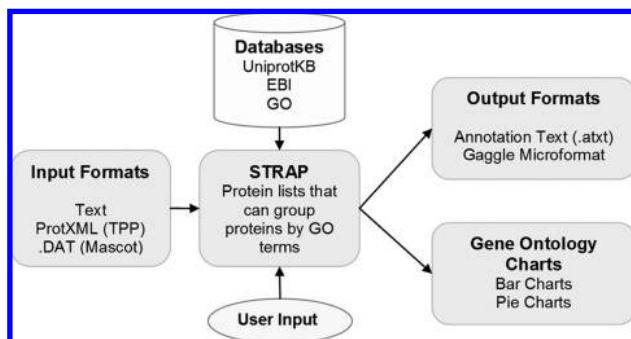
(9) Shannon, P. T.; Reiss, D. J.; Bonneau, R.; Baliga, N. S. *BMC Bioinf.* **2006**, *7*, 176.

(10) Bare, J. C.; Shannon, P. T.; Schmid, A. K.; Baliga, N. S. *BMC Bioinf.* **2007**, *8*, 456.

(11) Romesser, P. B.; Perlman, D. H.; McComb, M. E.; Faller, D. V.; Costello, C. E.; Denis, G. V. *Am. J. Pathol.* **2009**, in press.

(12) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. *Electrophoresis* **1999**, *20*, 3551–3567.

(13) Nesvizhskii, A. I.; Keller, A.; Kolker, E.; Aebersold, R. *Anal. Chem.* **2003**, *75*, 4646–4658.



**Figure 1.** Schematic representation of STRAP functionality, including data input and output. STRAP can read protein lists in UniProt entry or accession number format obtained from plain text files, as well as from Mascot and TPP ProteinProphet results files. STRAP then gathers protein GO-term annotation data from the public UniProtKB and the EBI GOA databases and allows editing of this data, providing the capacity to integrate in-house expertise on the proteins of study. STRAP can save the annotations to disk or export them to the Gaggle framework via Firegoose.

alternative workflows. The Mascot .dat search results format contains identifiers through which protein accession numbers are retrieved by STRAP. ProtXML results obtained using the TPP tools are in an XML format defined by the Institute for Systems Biology (ISB) that is mined by STRAP for accession numbers. STRAP focuses on data obtained from the TPP because of its open-source, community-driven nature. Each protein hit in the TPP's protXML file also has an associated probability which reflects the likelihood of valid protein assignment to the raw data. A STRAP user may choose to import only proteins above a given probability threshold. This probability cutoff acts as a filter to reduce the presence of low-scoring proteins and allows for more accurate and manageable data sets.

Following the import of protein identifications and subsequent annotations, the results may be saved as an annotated text file (.atxt) that contains all pertinent information regarding the data set or a FASTA formatted protein data set containing the names and sequences of the identified proteins. The .atxt files may be reloaded into STRAP at any later time to reconstitute the project for further analysis without the need to download all of the protein annotations again while the FASTA data set export may be used for subsequent searching via a database search algorithm when incorporating iterative search strategies. Additionally, gene names corresponding to the proteins in the lists may be encoded in the Gaggle microformat ([gaggle.systemsbio.org/docs](http://gaggle.systemsbio.org/docs)) for use with more diverse data analysis and interpretation options that are part of the Gaggle.<sup>9</sup> The Gaggle microformat file may be submitted for analyses by other Gaggled programs (e.g., KEGG Pathway Search) via FireGoose.<sup>10</sup> We chose to integrate STRAP with Gaggled programs to increase the overall utility of STRAP and allow for incorporation within open-source proteomics workflows which make use of the TPP.

**Annotations.** STRAP parses protein annotation information from XML data in the online UniProtKB database. Because downloading annotations can take several minutes, depending on the size of a protein list, annotation is carried out in a multi-threaded fashion to allow a user to simultaneously continue using STRAP, and other programs, for other tasks. At the time of downloading a new protein list, the user may select the option to also import GO annotations from the EBI GOA database (<http://www.ebi.ac.uk/GOA/>).<sup>16</sup> Downloaded annotation information is categorized into the following results: protein name, primary gene name, taxonomy, length, function, catalytic activity, and several gene ontology (GO) terms. A screenshot of the tabular results

SwissProt Accession Number	Uniprot ID	Name	Primary Gene Name	Taxonomy	Length	Function	Catalytic Activity	GO:Biological Process	GO:Cellular Component	GO:Molecular Function	Cellular process	Develop process	Interaction with cells and organisms	Nucleus
P18760	P18760	Cofilin-1	Cfl1	Mus musculus	166	Controls rev...		GO:0007015 cellular pr...	GO:0030864 cytoske...		▲	▲		
O08749	O08749	Dihydrofol...	Dld	Mus musculus	509	Lipoamide d...	Protein N(6){dihy...	GO:0007369 develop...	GO:0043159 (other in...	GO:0004148 [catalytic ...	▲	▲		
P40124	P40124	Adenylyl ...	Cap1	Mus musculus	474	Directly reg...		GO:0030036 cellular pr...	GO:0030864 cytoske...		▲			
P45376	P45376	Aldose re...	Aku1b1	Mus musculus	316	Catalyzes th...	Alditol + NAD(P)[...]			GO:0004032 [catalytic ...				
P17182	P17182	Alpha-en...	Eno1	Mus musculus	434	Multifunctio...	2-phospho-D-glyc...			GO:0004634 [catalytic ...				
O89053	O89053	Coronin-1A	Coro1a	Mus musculus	461	May be a cr...		GO:0030595 cellular pr...		GO:0042802 [binding, p...	▲			
P26443	P26443	Glutamat...	Glud1	Mus musculus	558	May be inv...	L-glutamate + H <sub>2</sub> O	GO:0007169 cellular pr...	GO:0005743 (mitocho...	GO:0004353 [catalytic ...	▲			
P62827	P62827	GTP-bin...	Ran	Mus musculus	216	GTP-bindin...		GO:0006606 cellular pr...		GO:0005515 [binding, p...	▲			
Q99020	Q99020	Heteroge...	Hnripab	Mus musculus	285	Transcriptio...		GO:0001837 interactio...	GO:0005737 (cytopla...	GO:0005515 [binding, p...		▲		●
Q88G05	Q88G...	Heteroge...	Hnripa3	Mus musculus		Plays a role ...								●
Q8R081	Q8R081	Heteroge...	Hnrip1	Mus musculus	555	This protein...			GO:0045120 (nucleus...					●
O88569	O88569	Heteroge...	Hnripa2b1	Mus musculus		Involved wit...		GO:0006397 cellular pr...	GO:0030530 (nucleus...	GO:0005515 [binding, p...	▲			●
P63158	P63158	High mo...	Hmgb1	Mus musculus	215	Binds prefer...		GO:0006809 cellular pr...	GO:0005515 [extracel...	GO:0030235 [enzyme r...	▲			●
P30681	P30681	High mo...	Hmgb2	Mus musculus	210	Binds prefer...		GO:0006288 cellular pr...	GO:0000793 (chromo...	GO:0008301 [binding, n...	▲			●
Q61792	Q61792	LIM and ...	Lasp1	Mus musculus	263	Plays an im...		GO:0030865 cellular pr...	GO:0030864 cytoske...	GO:0051015 [binding, p...	▲			
P24452	P24452	Macroph...	Capp	Mus musculus	352	Calcium-sen...		GO:0030031 cellular pr...			▲			
P26041	P26041	Moesin	Msx	Mus musculus	577	Probably in...			GO:0016324 (plasma ...					
P17742	P17742	Peptidyl...	Ppia	Mus musculus	164	PFases ac...	Peptidylproline [o...	GO:0030182 cellular pr...	GO:0005829 (cytopla...	GO:0005515 [binding, p...	▲			
Q61599	Q61599	Rho GD...	Arhgdb	Mus musculus	200	Regulates t...				GO:0005515 [binding, p...				
P21107	P21107	Tropomy...	Tpm3	Mus musculus		Binds to act...			GO:0030863 cytoske...					
Q91ZJ5	Q91ZJ5	UTP-glu...	Ugp2	Mus musculus		Plays a cen...	UTP + alpha-D-gl...							

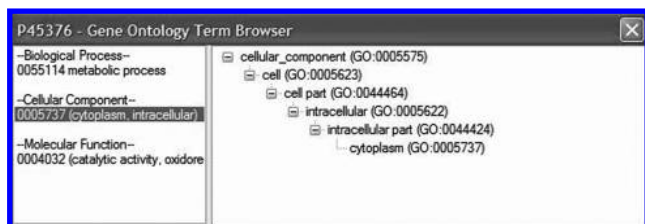
**Figure 2.** GUI interface of STRAP showing the main protein annotation view. All columns in the main annotation table, including the GO category column, allow sorting. This allows users to group proteins by any gene ontology. Annotation attributes can be edited to include or eliminate GO-term associations with particular protein entries according to the users' expertise.



**Table 1. STRAP Categorization of Gene Ontology (GO) Terms<sup>a</sup>**

major GO category	associated GO term	no. of subcategories
biological process	GO:0008150	10
cellular component	GO:0005575	17
molecular function	GO:0003674	12

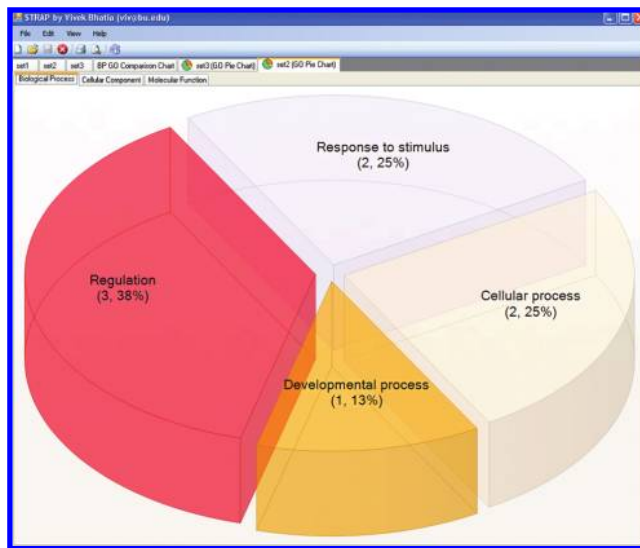
<sup>a</sup> The three major categories of GO terms are incorporated within strap as shown here. For the complete list of categories and subcategories, see Supplemental Table 1 in the Supporting Information.



**Figure 3.** STRAP's built in Gene Ontology Term Browser. The browser presents all GO terms associated with a particular protein entry, as well as each GO term's complete lineage.

shown in STRAP is presented in Figure 2. Because GO terms are organized within nested hierarchies below the three major categories of Biological Process, Cellular Component, and Molecular Function, STRAP displays the GO term annotations on its main annotation table in columns corresponding to these three categories, each with a different representative symbol. Further organization of the GO terms into subcategories is based on a simplified version of the categories originally published by the Gene Ontology Consortium, with major categories being taken from [geneontology.org](http://geneontology.org).<sup>17</sup>

The main GO term columns used are as shown in Table 1. A complete description of GO terms subcategory may be found in the Supplemental Table 1A–C in the Supporting Information under the main category to which they belong. For simplicity, any GO column is made visible only if there is at least one protein in the list that has the corresponding GO term annotation. STRAP displays a simplified GO category list to improve viewing capabilities; however, the user may still view the original GO lineage of a protein's GO terms via STRAP's built-in Gene Ontology Term Browser (Figure 3). In addition to the full annotation display of GO term attributes obtained from the online database, users can manually edit these terms within STRAP as necessary if a user has additional expert information about the proteins of interest. Both inclusion into, and exclusion from, previously associated GO term categories can be specified for each protein entry. These results then may be stored in the STRAP atxt file.

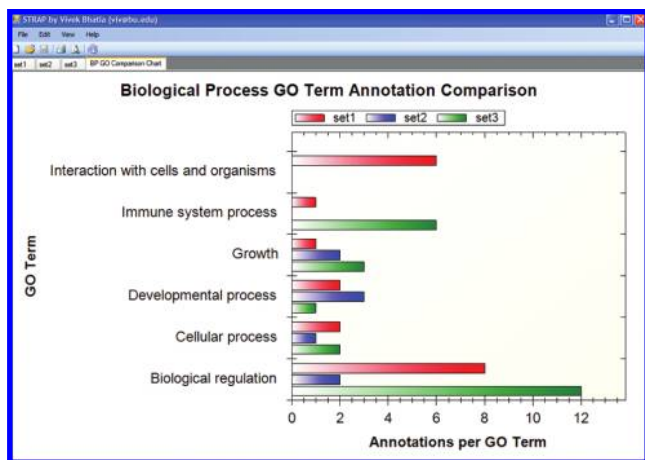


**Figure 4.** STRAP pie chart rendering. To graphically display the GO-term subcategories for each data set, STRAP can generate pie charts for each of the three main GO categories, wherein each slice represents a subcategory. Each pie slice is labeled with the GO subcategory name, the number of GO annotations within the category, and the percentage of annotations associated with that particular GO term. Shown is a Biological Process pie chart generated from the GO terms associated with a set of 10 proteins, using the example data set as described within the text. Note that each unit of the pie represents one GO term rather than one protein, as one protein can be assigned multiple GO terms.

**GO Annotation Visualization.** There are currently two ways to visualize protein annotations in STRAP: pie charts and bar graphs. Selecting pie-chart rendering allows for display of each major category of GO term (biological process, cellular component, and molecular function) as shown in Figure 4. Each slice of the pie chart is labeled with the appropriate subcategory title and the number of GO annotations that fall under this category. In addition to showing the number of GO annotations, a percentage value is listed to indicate the fraction of GO terms that is represented out of the complete list of GO annotations for the entire set of proteins. For example, from the particular data set shown in Figure 4, it appears that "Regulation" is a frequent annotation in common to many proteins in the current protein list. On the basis of the biological context of the data, this might be an interesting subset of proteins for further study.<sup>11</sup>

Whereas pie charts provide visualization of single data sets that are easy to interpret on a per-sample or per-project basis, bar graphs may be rendered for ease of comparative analysis of multiple data sets. STRAP can render bar graphs on both single results sets and multiple sets of proteins and thus be used to qualitatively and semiquantitatively interpret the biological significance of proteomics data. Comparison may be performed on any of the protein data sets' GO terms, including the number of annotations for subcategories within each top-level GO category. In the example shown here in Figure 5, a comparison is made across three complementary proteomics data sets. This easily allows the visualization of global changes in protein functions across the data sets and provides information for global data analysis and comparison and affords planning of future experiments. In this example, among the three data sets, "growth" is an annotation that is found more frequently in set 3 than the other

- Pedrioli, P. G.; Eng, J. K.; Hubley, R.; Vogelzang, M.; Deutsch, E. W.; Raught, B.; Pratt, B.; Nilsson, E.; Angeletti, R. H.; Apweiler, R.; Cheung, K.; Costello, C. E.; Hermjakob, H.; Huang, S.; Julian, R. K.; Kapp, E.; McComb, M. E.; Oliver, S. G.; Omenn, G.; Paton, N. W.; Simpson, R.; Smith, R.; Taylor, C. F.; Zhu, W.; Aebersold, R. *Nat. Biotechnol.* **2004**, *22*, 1459–1466.
- Keller, A.; Eng, J.; Zhang, N.; Li, X. J.; Aebersold, R. *Mol. Syst. Biol.* **2005**, *1*, 2005 0017.
- Barrell, D.; Dimmer, E.; Huntley, R. P.; Binns, D.; O'Donovan, C.; Apweiler, R. *Nucleic Acids Res.* **2009**, *37*, D396–D403.
- Beisvag, V.; Junge, F. K.; Bergum, H.; Jolsum, L.; Lydersen, S.; Gunther, C. C.; Ramampiaro, H.; Langaas, M.; Sandvik, A. K.; Laegreid, A. *BMC Bioinf.* **2006**, *7*, 470.



**Figure 5.** STRAP bar chart rendering for the comparison of multiple data sets. STRAP can generate bar charts allowing for comparison and visualization of large data sets based upon GO terms. This bar graph compares the amount of Biological Process GO term annotations between three sets of proteins as described within the text.

sets, which may imply that growth-related proteins are up-regulated in set 3.<sup>11</sup> When there are many data sets (e.g., more than 5) being compared, the user can elect to generate a comparison table in addition to a bar chart.

## 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 and the EBI GOA 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. Furthermore, these graphs provide qualitative information that is instrumental to planning future experiments. This easy to use PC-based software allows researchers to rapidly parse and an-

notate large sets of proteins from a variety of sources. Additionally, it can export data to FASTA databases or to the large pool of Gaggles framework software for further data interrogation including pathway analysis.

STRAP currently mines the UniProtKB and EBI GOA databases for annotation information, but these databases do not contain an exhaustive list of proteins, and proteins that are in these databases are not always completely annotated. Fortunately, both databases are rapidly growing to encompass more complete annotations and additional protein entries. While we anticipate that future releases of STRAP will have the capability to automatically search databases other than the UniProtKB and EBI GOA, users can use the ID Mapping tool provided by UniProt ([www.uniprot.org/?tab=mapping](http://www.uniprot.org/?tab=mapping)) for preanalysis file conversion should they be interested in characterization of proteins obtained from other database driven proteomics experiments.

Because gene ontology annotation is still incomplete, it is useful to access gene ontology annotations from various sources. There are several applications in the public domain which focus on predicting gene ontology annotations: GoFigure,<sup>18</sup> Blast2GO,<sup>19</sup> GOanna,<sup>20</sup> and GeneTools<sup>17</sup> predict GO terms with a homology search, and EBIMed<sup>21</sup> predicts them using a literature-based search. We plan to integrate such GO term prediction as well as GO term enrichment and depletion calculation capabilities in a future release of STRAP. Furthermore, to meet the increasing focus on post-translational modifications (PTMs) in proteomics, future releases of STRAP will include the ability to visualize and compare the character and extent of PTMs across data sets in addition to simple protein comparisons. Finally, as the source code is readily available to the community, others may add their own functionality to STRAP as they see fit.

## ACKNOWLEDGMENT

We thank all the members of the Center for Biomedical Mass Spectrometry for evaluating STRAP and providing feedback during the course of program development. This research was funded by NIH-NHLBI Contract N01 HV28178 and NIH-NCRR Grant P41 RR10888.

## SUPPORTING INFORMATION AVAILABLE

Additional information as noted in text. This material is available free of charge via the Internet at <http://pubs.acs.org>.

Received for review June 19, 2009. Accepted October 5, 2009.

AC901335X

- (18) Khan, S.; Situ, G.; Decker, K.; Schmidt, C. J. *Bioinformatics* **2003**, *19*, 2484–2485.  
 (19) Conesa, A.; Gotz, S.; Garcia-Gomez, J. M.; Terol, J.; Talon, M.; Robles, M. *Bioinformatics* **2005**, *21*, 3674–3676.  
 (20) McCarthy, F. M.; Wang, N.; Magee, G. B.; Nanduri, B.; Lawrence, M. L.; Camon, E. B.; Barrell, D. G.; Hill, D. P.; Dolan, M. E.; Williams, W. P.; Luthe, D. S.; Bridges, S. M.; Burgess, S. C. *BMC Genomics* **2006**, *7*, 229.  
 (21) Rebholz-Schuhmann, D.; Kirsch, H.; Arregui, M.; Gaudan, S.; Rynbeek, M.; Stoehr, P. *Nat. Biotechnol.* **2006**, *24*, 902–903.