BOSTON UNIVERSITY

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

Vivek N. Bhatia, David H. Perlman, Catherine E. Costello, Mark E. McComb.
Cardiovascular Proteomics Center, Boston University School of Medicine, Boston, MA

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 the pace of research.

Where Do Protein Lists Come From?

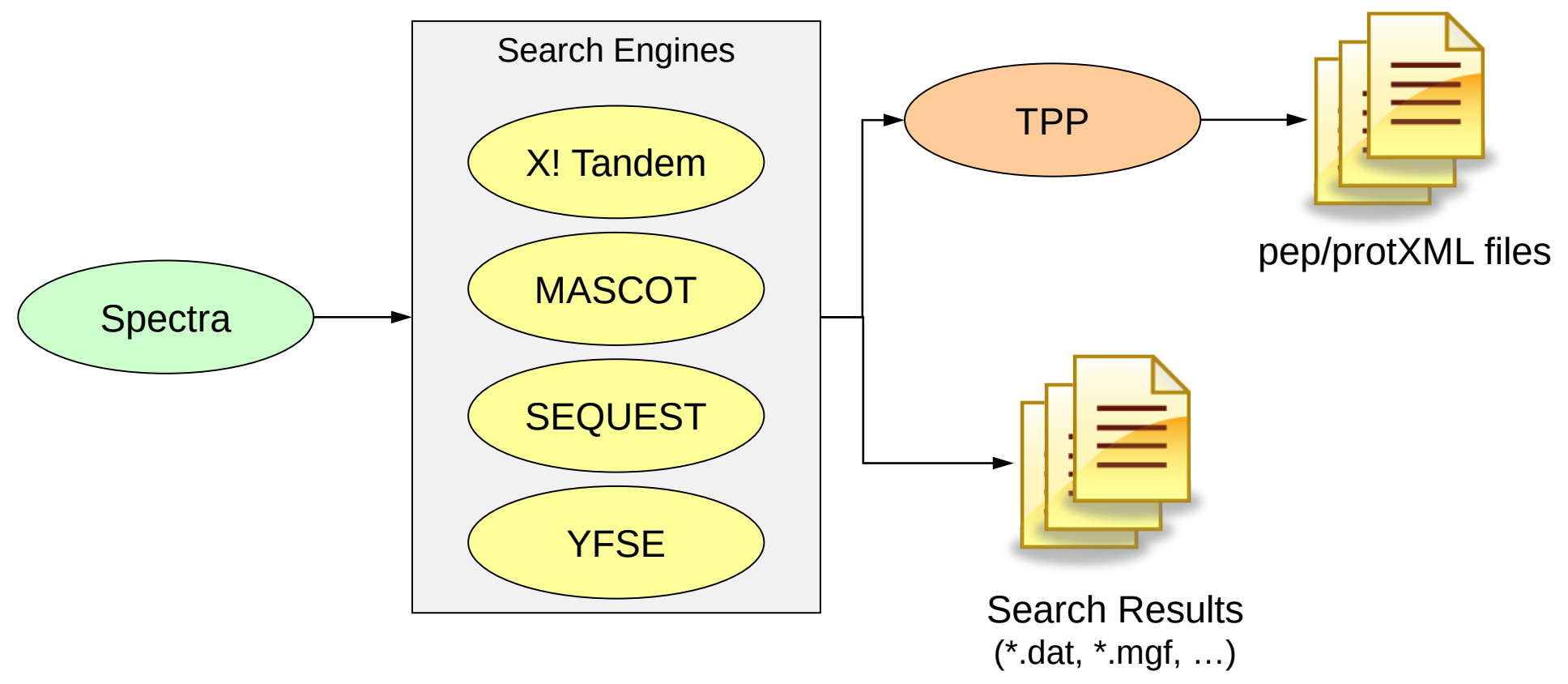

Search Engines
X! Tandem
MASCOT
SEQUEST
YFSE
Spectra
TPP
pep/protXML files
Search Results
(*.dat, *.mgf, ...)

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.

AVAILABLE TOOLS

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

- Lab Information Management System (LIMS)-type software, such as Genologics 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 PIPE from the Institute for Systems Biology, allow for analysis without a complicated setup.

PROBLEMS WITH TODAY'S TOOLS

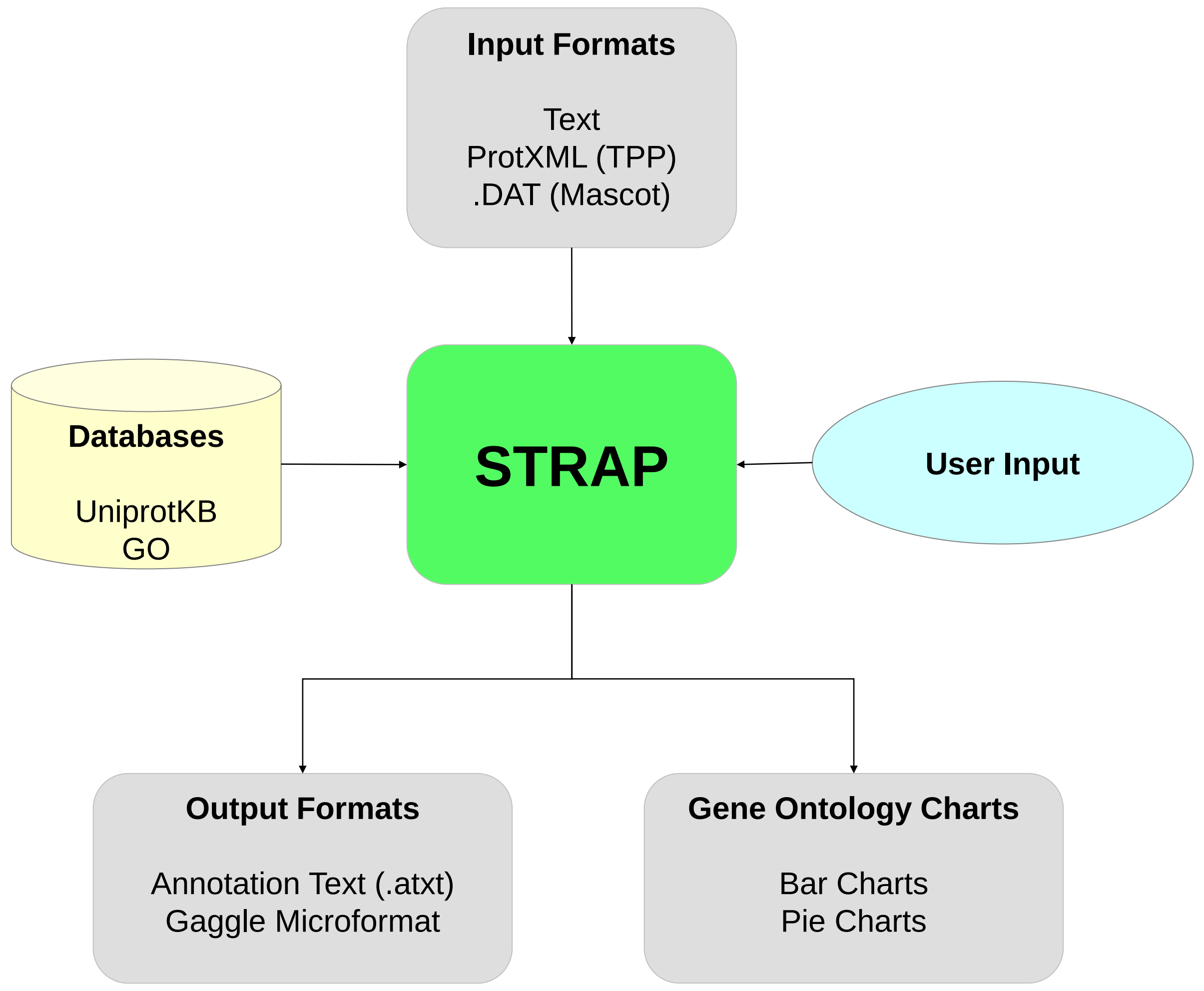
Today's tools help with data interpretations, but are limited because:

- 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 ZedGraph (http://zedgraph.org) and 3D Pie Chart libraries to create charts (http://www.codeproject.com/KB/graphics/julijanpiechart.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, FireGoose (the Gaggle Toolbar for Fire-fox) 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


Input Formats
Text
ProtXML (TPP)
.DAT (Mascot)

Databases
UniprotKB
GO

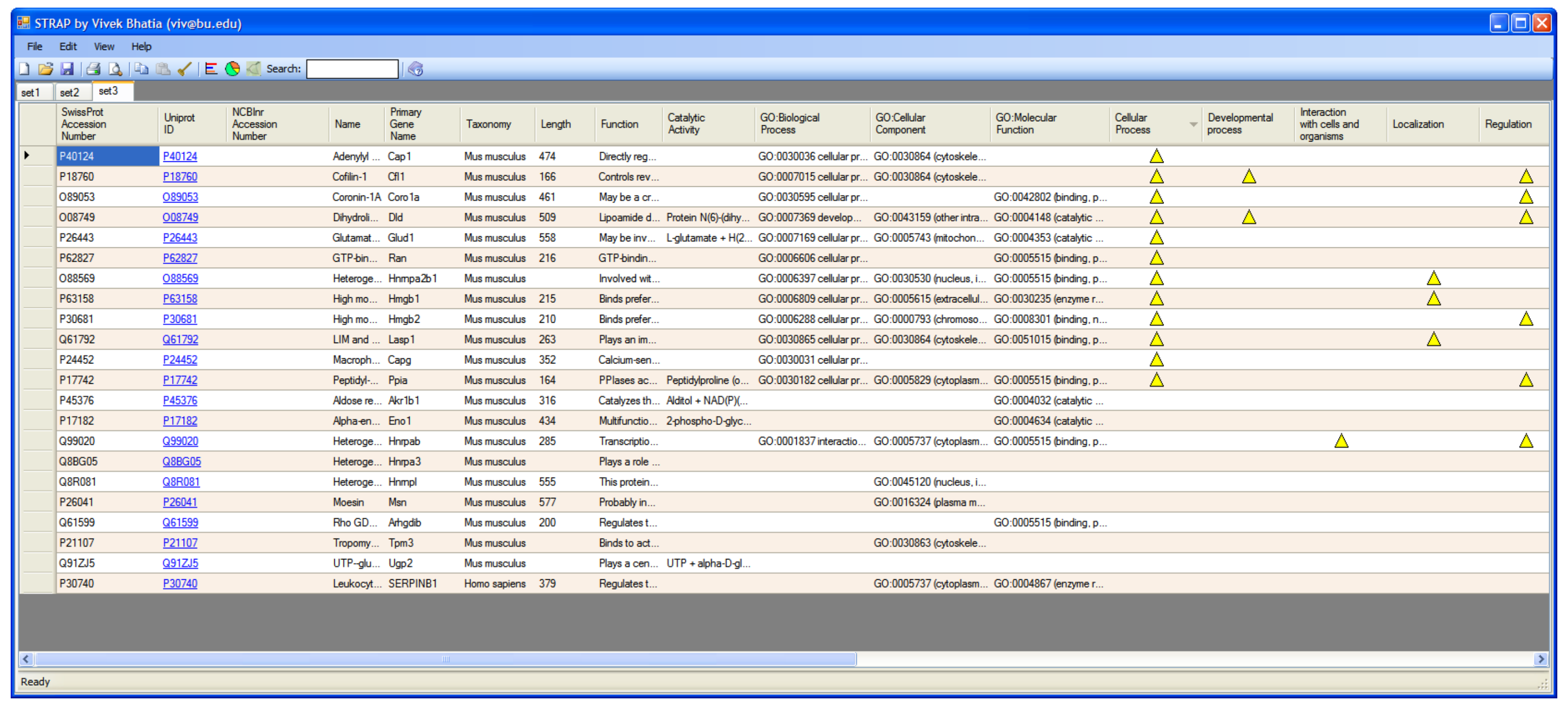
STRAP

User Input

Output Formats
Annotation Text (.atxt)
Gaggle Microformat

Gene Ontology Charts
Bar Charts
Pie Charts

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.


STRAP by Vivek Bhatia (viv@bu.edu)

SwissProt Accession Number	Uniprot ID	NCBInr Accession Number	Name	Primary Gene Name	Taxonomy	Length	Function	Catalytic Activity	GO:Biological Process	GO:Cellular Component	GO:Molecular Function	Cellular Process	Developmental process	Interaction with cells and organisms	Localization	Regulation
P40124	P40124		Adenylyl ...	Cap1	Mus musculus	474	Directly reg...		GO:0030036 cellular pr...	GO:0030864 (cytoskele...		△				
P18760	P18760		Cofilin-1	Cfl1	Mus musculus	166	Controls rev...		GO:0007015 cellular pr...	GO:0030864 (cytoskele...		△	△			△
O89053	O89053		Coronin-1A	Coro1a	Mus musculus	461	May be a cr...		GO:0030595 cellular pr...		GO:0042802 (binding, p...	△				△
O08749	O08749		Dihydroli...	Dld	Mus musculus	509	Lipoamide d...	Protein N(6)-(dihy...	GO:0007369 develop...	GO:0043159 (other intra...	GO:0004148 (catalytic ...	△	△			△
P26443	P26443		Glutamat...	Glud1	Mus musculus	558	May be inv...	L-glutamate + H(2...	GO:0007169 cellular pr...	GO:0005743 (mitochon...	GO:0004353 (catalytic ...	△				
P62827	P62827		GTP-bin...	Ran	Mus musculus	216	GTP-bindin...		GO:0006606 cellular pr...		GO:0005515 (binding, p...	△				
O88569	O88569		Heteroge...	Hnrnpa2b1	Mus musculus		Involved wit...		GO:0006397 cellular pr...	GO:0030530 (nucleus, i...	GO:0005515 (binding, p...	△			△	
P63158	P63158		High mo...	Hmgb1	Mus musculus	215	Binds prefer...		GO:0006809 cellular pr...	GO:0005615 (extracellul...	GO:0030235 (enzyme r...	△			△	
P30681	P30681		High mo...	Hmgb2	Mus musculus	210	Binds prefer...		GO:0006288 cellular pr...	GO:0000793 (chromoso...	GO:0008301 (binding, n...	△				△
Q61792	Q61792		LIM and ...	Lasp1	Mus musculus	263	Plays an im...		GO:0030865 cellular pr...	GO:0030864 (cytoskele...	GO:0051015 (binding, p...	△			△	
P24452	P24452		Macroph...	Capg	Mus musculus	352	Calcium-sen...		GO:0030031 cellular pr...			△				
P17742	P17742		Peptidyl-...	Ppia	Mus musculus	164	PPIases ac...	Peptidylproline (o...	GO:0030182 cellular pr...	GO:0005829 (cytoplasm...	GO:0005515 (binding, p...	△				△
P45376	P45376		Aldose re...	Akr1b1	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 ...					
Q99020	Q99020		Heteroge...	Hnrpab	Mus musculus	285	Transcriptio...		GO:0001837 interactio...	GO:0005737 (cytoplasm...	GO:0005515 (binding, p...			△		△
Q8BG05	Q8BG05		Heteroge...	Hnrpa3	Mus musculus		Plays a role ...									
Q8R081	Q8R081		Heteroge...	Hnrnpl	Mus musculus	555	This protein...			GO:0045120 (nucleus, i...						
P26041	P26041		Moesin	Msn	Mus musculus	577	Probably in...			GO:0016324 (plasma m...						
Q61599	Q61599		Rho GD...	Arhgdib	Mus musculus	200	Regulates t...				GO:0005515 (binding, p...					
P21107	P21107		Tropomy...	Tpm3	Mus musculus		Binds to act...			GO:0030863 (cytoskele...						
Q91ZJ5	Q91ZJ5		UTP--glu...	Ugp2	Mus musculus		Plays a cen...	UTP + alpha-D-gl...								
P30740	P30740		Leukocyt...	SERPINB1	Homo sapiens	379	Regulates t...			GO:0005737 (cytoplasm...	GO:0004867 (enzyme r...					

Ready

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.

STRAP OVERVIEW AND SCREENSHOTS (II)

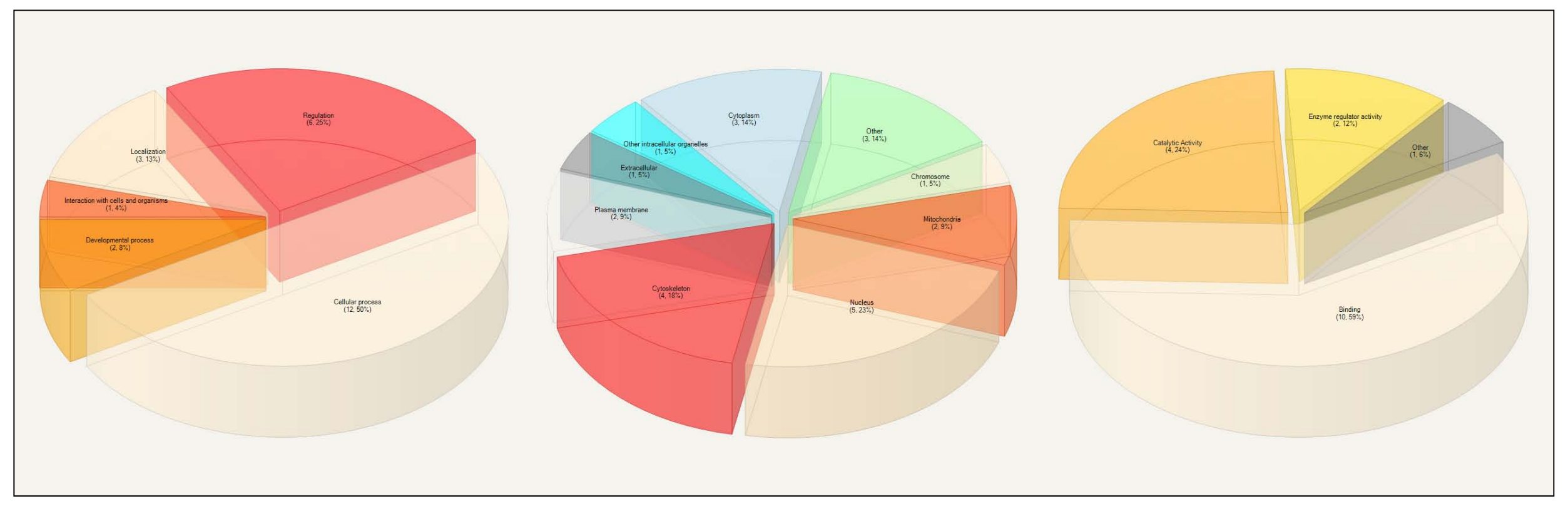
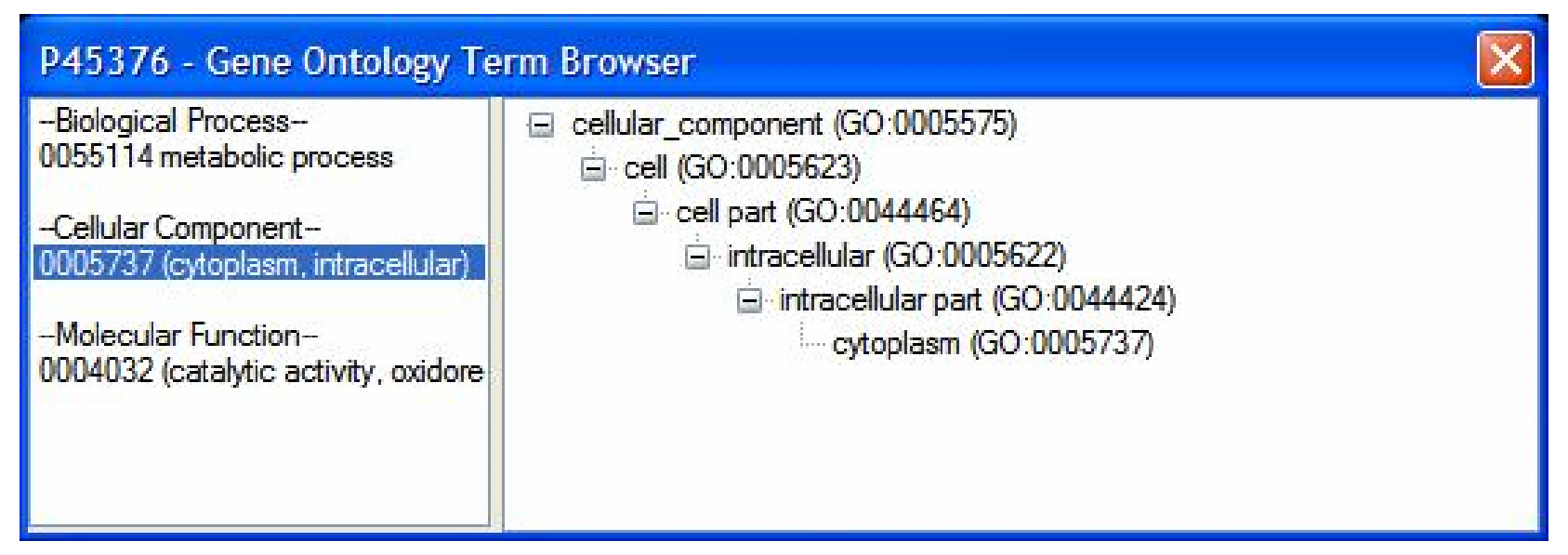


Figure 4. This pie chart was generated from the gene ontology terms of a set of 22 proteins. Each slice represents a high-level category of gene ontology (GO) terms. In order from left to right, the pie charts correspond to biological process, cellular component, and molecular function GO terms.


P45376 - Gene Ontology Term Browser

--Biological Process--
0055114 metabolic process

--Cellular Component--
0005737 (cytoplasm, intracellular)

--Molecular Function--
0004032 (catalytic activity, oxidore

cellular_component (GO:0005575)
cell (GO:0005623)
cell part (GO:0044464)
intracellular (GO:0005622)
intracellular part (GO:0044424)
cytoplasm (GO:0005737)

Figure 6. The 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.

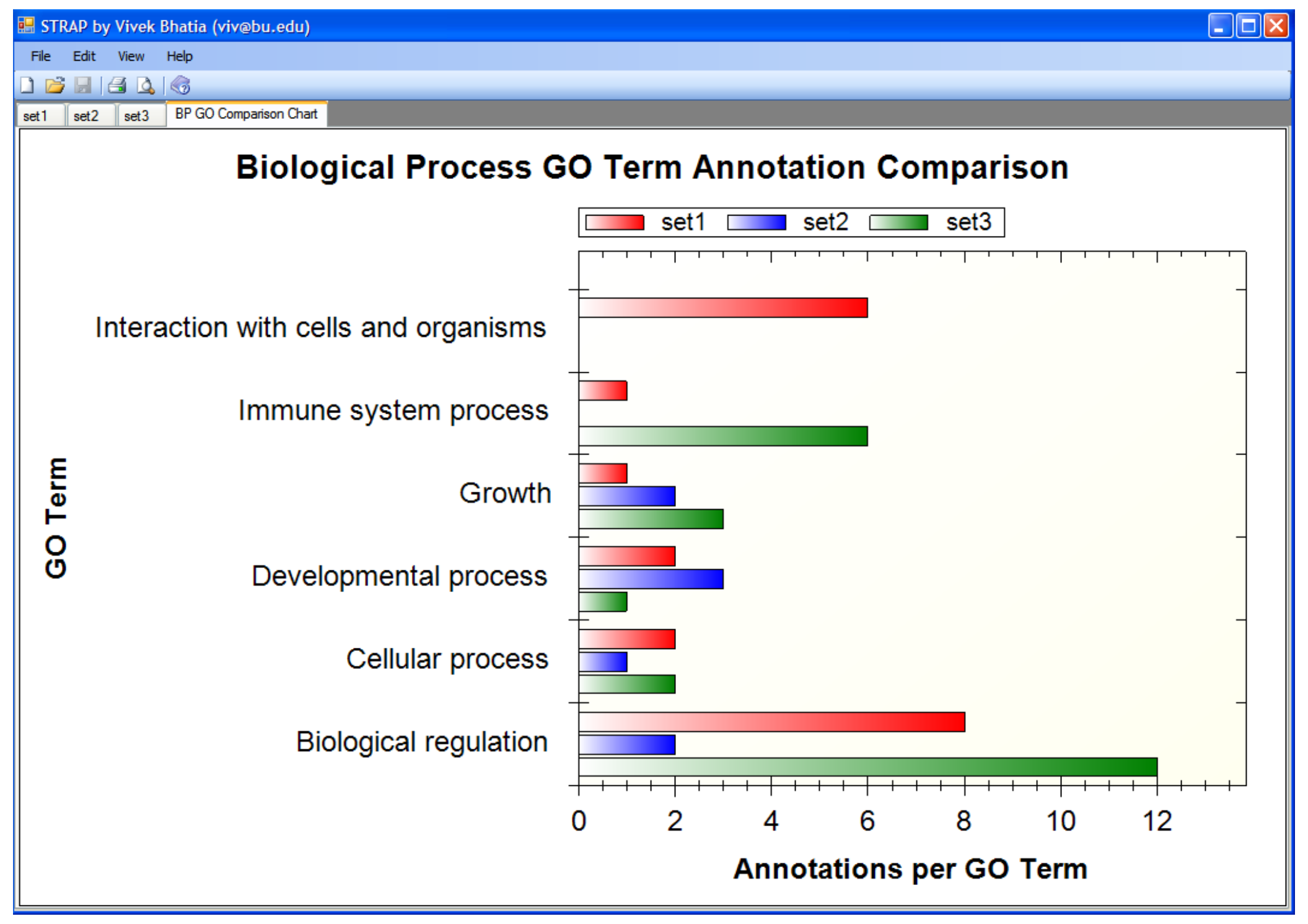


Figure 5. This bar graph compares the amount of biological process GO term annotations between proteins in three different samples.

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.
- Allows manual annotation of GO terms to incorporate in-house expertise
- Provides pie and bar charts of gene ontology terms to aid annotation interpretation
- Bases annotation on the UniProt Knowledgebase database
- 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

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 easy to use PC-based software allows researchers to rapidly parse and annotate large sets of proteins.

AVAILABILITY

STRAP and its C# source code are available under the lesser GNU public license (LGPL) at: http://sourceforge.net/projects/cpctools/.

CURRENT AND FUTURE WORK

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

AKNOWLEDGMENTS

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