

# The Protein Information and Property Explorer: a rich-client web application for the management and functional exploration of proteomic data

Hector Ramos<sup>1</sup>, Paul Shannon<sup>1</sup>, and Ruedi Aebersold<sup>1,2</sup>  
 1, Institute for Systems Biology, Seattle WA, USA  
 2, Institute of Molecular Systems Biology, ETH Zurich, Zurich, Switzerland

## ABSTRACT

**Motivation:** Mass spectrometry experiments in the field of proteomics produce lists containing tens to thousands of identified proteins. With the Protein Information and Property Explorer (PIPE) the biologist can acquire functional annotations for these proteins and explore the enrichment of the list, or fraction thereof, with respect to functional classes. These protein lists may be saved for access at a later time or different location. The PIPE is interoperable with the Firegoose and the Gaggle, permitting wide-ranging data exploration and analysis. The PIPE is a rich-client web application which uses AJAX capabilities provided by the Google Web Toolkit, and server side data storage using Hibernate.

**Availability:** <http://pipe.systemsbiology.net>  
**Contact:** [pshannon@systemsbiology.org](mailto:pshannon@systemsbiology.org)

## INTRODUCTION

After an MS/MS proteomics experiment has completed and the results have been analyzed with a suite of software tools such as the Trans-Proteomic Pipeline (Keller et al.), the end result is typically a list of protein identifiers (IPI, UniRef, UniProt, etc) with varying degrees of certainty assigned each protein on the list. The Protein Information and Property Explorer (PIPE) is a starting point for the next step in analysis of proteomic experimental results such as the functional annotation of the identified proteins and their association with biological processes. It is a launching pad from which simple operations can be performed on this data and messages can be passed to other, more sophisticated analysis software such as the Gaggle (Shannon et al.). The PIPE currently supports Human, Mouse, Rat, Yeast protein identifiers; as well as a few other, less commonly studied organisms.

## FEATURES

### Identifier Mapping

Most biological annotation (GO and KEGG, for example) is provided in terms of Entrez Gene IDs. The first step in annotating proteins, therefore, is to map them to the genes from which they are transcribed and translated. The PIPE currently maps IPI, UniProt, and NCBI protein identifiers to Entrez Gene IDs, and thence to gene symbols, descriptions, and associated Gene Ontology terms. Other mappings are easily added.

### Protein Sequence Lookup

Not all protein identifiers are mapped to genes by the standard bioinformatics authorities. For these cases, the PIPE provides easy point and click access to a protein's sequence and submission to NCBI BLAST. Once the user has examined the blast results and determined which gene, if any, the protein is associated with, he/she may enter and save the new mapping in the PIPE; this assignment may be preserved for future lookups

## SCREEN SHOTS

View of the PIPE after logging in

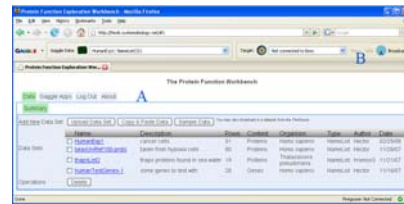


Fig. 1. A) Logging into the PIPE produces a view summarizing all previously entered data sets. B) The Firegoose. Through JavaScript, the PIPE and the Gaggle interchange data at the click of the "Broadcast" button.

The option to add personal annotations to un-annotated proteins

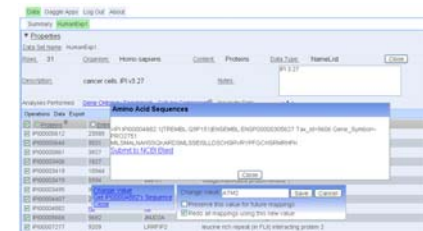


Fig. 2. When no mapping is found for a protein identifier, clicking on the "na" value gives the user the option to lookup the protein sequence or manually fill in the missing value.

### Gene Ontology Enrichment

Functional annotation and the association of the identified proteins with biological processes are crucial to the analysis of proteomics data. This is often achieved by computing the enrichment of proteins in the identified set related to function or other ontology classes. The PIPE employs the Bioconductor R package 'GOstats', running on the backend server, to calculate the relative enrichment of sets of identified proteins in relation to the whole set of proteins with respect to biological process, molecular function, and/or cellular component information. The PIPE generates a 'Gaggle' instance of Cytoscape (Shannon et al.) to display an interactive, hierarchical graph of the enriched GO categories.

Opening an uploaded dataset



Fig. 3. Opening a data set produces a view of the data and a menu bar containing operations which can be performed on the data. Here we have performed an ID Mapping operation from IPI numbers to Entrez Gene ID, gene symbol, and description.

Viewing the results of a GO enrichment operation

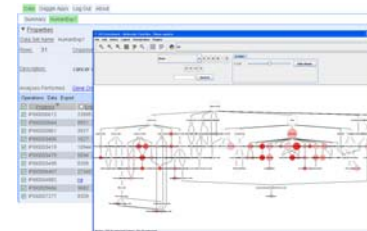


Fig. 4. The results of a Gene Ontology enrichment operation are presented in a Gaggle Cytoscape window.

### Communication with other Software and Websites

By interfacing with the Firegoose (Bare et al.), the PIPE can access several valuable online bioinformatics resources as well as different Java applications running on the user's desktop. Web resources include DAVID, Kegg Pathways, Entrez Gene, and EMBL String. The Firegoose Firefox browser plug-in plays the intermediary between the PIPE and other applications in the Gaggle. The PIPE is thus able to send and receive data from popular Gaggle enabled software applications such as Cytoscape, Data Matrix Viewer, MultiExperiment Viewer, and the R statistical environment.

## Data Persistence

A prominent feature of the PIPE is the ability to save data at any given step in an analysis and in that way provide persistence to the data mining workflow. Because of the nature of web applications, this data can then be accessed at a later time or at a different location.

## TECHNICAL DETAILS

The Google Web Toolkit (GWT) was employed to enable extensive Asynchronous JavaScript and XML (AJAX) in the application. This has an advantage over traditional web applications in that the user is not required to wait for page reloading. Also, in contrast to traditional web applications, the MVC model runs entirely on the client browser, contacting the server only for specific data when required. The GWT allows the developer to code in Java and then compiles the code into JavaScript that runs in the browser. The back end is hosted on a tomcat servlet container and uses Hibernate ORM technology and MySQL for data persistence.

## FUTURE DEVELOPMENT

To enable collaborations between groups and users, a permissions system based on the standard owner-group-world paradigm will be implemented in the PIPE. Furthermore, to continue to complement the Gaggle framework, the PIPE will be extended to support the remaining 4 Gaggle data types, beginning with Networks. Storing the Network data type in the PIPE will facilitate data exploration and network inference in Cytoscape and other networking applications.

## REFERENCES

- Bare, J., Shannon, P., Schmid, A., and Baliga, N. (2007) The Firegoose: a two-way integration of diverse data from different bioinformatics web resources with desktop applications. *BMC Bioinformatics*, 8, 456.
- Keller, A., Eng, J., Zhang, N., Li, X. J., Aebersold, R. (2005) A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Mol. Syst. Biol.* 1, 17.
- Shannon, P., Markiel, A., Ozier, O., Baliga, N., Wang, J., Ramage, D., Amin, N., Schwikowski, B., Ideker, T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13, 2498-2504.
- Shannon P., Reiss, D., Bonneau, R., and Baliga, N. (2006) The Gaggle: an open-source software system for integrating bioinformatics software and data sources. *BMC Bioinformatics*, 7, 176.

## ACKNOWLEDGMENTS

This work was supported by the National Heart, Lung, and Blood Institute, National Institutes of Health, under contract No. N01-HV-28179 (to R.A.).