

Satya S. Sahoo<sup>1</sup>, Daniel B. Weatherly<sup>2</sup>, Lin Lin<sup>3</sup>, James A. Atwood III<sup>3</sup>, Rick Tarleton<sup>2</sup>, Ron Orlando<sup>3</sup>, William S. York<sup>3</sup>, Amit P. Sheth<sup>1</sup>, Krys Kochut<sup>1</sup>, John Miller<sup>1</sup>  
<sup>1</sup>Large Scale Information Distributed Information Systems, <sup>2</sup>Center for Tropical and Emerging Global Diseases, <sup>3</sup>Complex Carbohydrate Research Center

## OBJECTIVES

Develop systems to:

- Track, store, and query experimental glycoproteomics data.
- Automate processing and storage of MS data to database search results
- Statistically validate and combine multiple database search results
- Easily correlate experimental results with biology

## INTRODUCTION

- No commercial software is available that facilitates the high throughput analysis of large glycoproteomic datasets
- No methods exist to annotate glycoproteomic data
- Currently the end result of several hundred proteomic analyses is a static list of protein identifications with little or no additional information
- Robust resources need to be developed to annotate, track, store and disseminate glycoproteomic experimental results in a high throughput and transparent manner

## STEP 1: TRACKING EXPERIMENTAL DATA

- We have created a simple schema to facilitate storage and tracking of essentially any information. The essential idea is to create a database schema that allows the researcher to create an organism/process-specific database for his own needs.
- This "schema-less" schema is created by treating the database tables that would have been designed in a traditional schema as categories and the individual fields in each table as properties of these categories (Figure 1).
- Using a *T. cruzi*-specific example the categories are organism, strain, stage, fractionation, and ms prep while the properties are the names and description fields of each table (Figure 2).

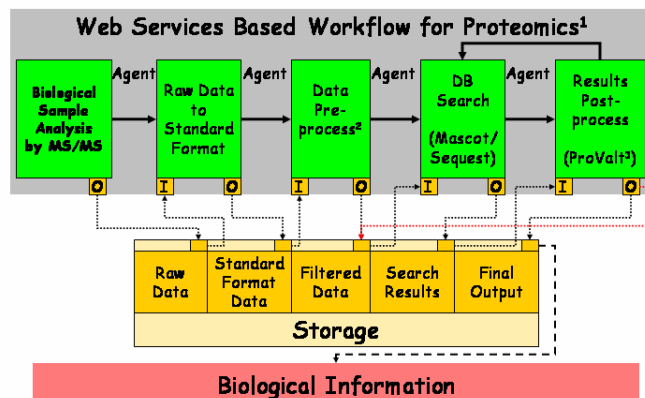
## ACKNOWLEDGMENTS

This work is funded by the NIH grants RO1 AI-033106, PO1 AI-044979 (to R.L.T.) and the NIH Research Resource for Integrated Glycotechnology P41RR005351.

## STEP 2 : AUTOMATED DATA PROCESSING USING WEB SERVICES

- The protocol developed for high throughput glycoproteomics consists of multiple computational resources, located on multiple machines running different operating systems, executed in a specific order with attendant transfer of data (Table 1).
- We deployed **Web Services** that execute these specific tasks in the workflow using the open source SOAP (Simple Object Access Protocol, the XML-based language used by Web Services to communicate) engine Apache Axis, deployed over the open source servlet container Apache Tomcat (Figure 3).

Figure 3.



- Design and Implementation of Web Services based Workflow for proteomics. Journal of Proteome Research. Submitted
- Computational tools for increasing confidence in protein identifications. Association of Biomolecular Resource Facilities Annual Meeting, Portland, OR, 2004.
- A Heuristic method for assigning a false-discovery rate for protein identifications from Mascot database search results. *Mol. Cell. Proteomics*. 4(6), 762-772.

Figure 1. The "schema-less" schema

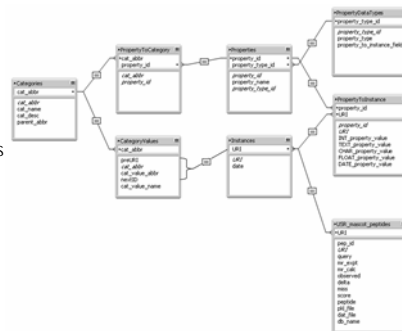


Figure 2. Example DB structure

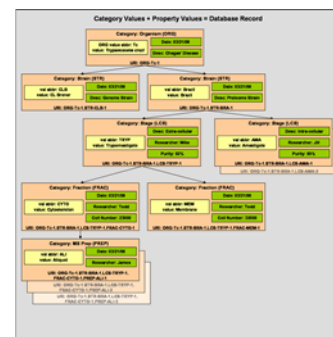


Table 1.

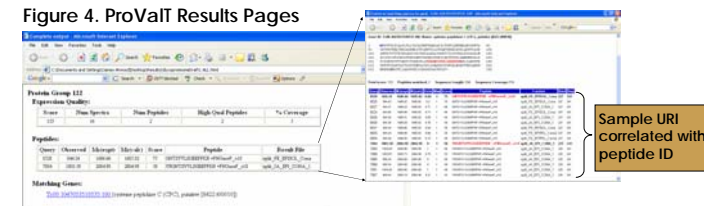
The workflow broadly consists of the following steps:

- Generation of RAW file by a mass spectrometer (outside the scope of the workflow)
- Conversion of RAW file to mzXML format
- Archive one copy of the mzXML file and use the second copy for further processing
- Conversion of mzXML file into peak list format (.pkl)
- Archive one copy of the .pkl file and use the second copy for further processing
- Processing of .pkl file using a specialized software for correction of precursor ion charge state (pSplit)
- Archive one copy of the pSplit file and use the second copy for further processing
- Invoke MASCOT database search for each of the pSplit files
- Accumulate and archive all the database search results for statistical analysis by ProValT
- Processing all database search results using ProValT
- Archive ProValT results

## STEP 3: RESULTS to BIOLOGY

- The peptides from the ProValT results (Figure 4) are then input into the database as a function of the sample from which they originated.

Figure 4. ProValT Results Pages



- The "searchURI" interface allows the researcher to formulate any possible query of the information that is stored in the database (Figure 5). Complex queries can be formed using a combination of search criteria and the boolean operators "AND", "OR", and "NOT".

Figure 5.

