



Jagan Kommineni¹, Simon G. Michnowicz¹ and Robert L. Moritz^{1,2}

¹ Joint ProteomicS Laboratory, Ludwig Institute For Cancer Research and The Walter and Eliza Hall Institute of Medical Research, Royal Melbourne Hospital, Parkville, Victoria, Australia 3050. ² Institute for Systems Biology, Seattle, WA, USA.

APCF Computational Cluster: Post-processing of proteomics search results

Introduction

The APCF's large computational power of over a 1000 CPU cores contained in 128 dedicated computational nodes dedicated to proteomics data analysis via the world-wide-web is managed by multiple interactive systems that direct resource allocation and act as gateway systems to the computational nodes. We have devised a special software abstraction layer called APCF Meta Scheduler (AMS) to link all the 128 computational nodes having 1024 cores and perform computational tasks in an efficient manner which makes the system utilized to the maximum extent and on the other end respond back to the users requests at the quickest possible.

Different proteomics search algorithms such as Mascot, X!Tandem and OMSSA produce results completely different formats and there is no validation mechanism in place. It is not only time consuming to validate results manually but also highly tiresome and difficult as new generation of mass spectrometry instruments produce data enormously high data rates. We have developed a pipeline to use industry-standard statistical tools, PeptideProphet, iProphet and ProteinProphet developed at Institute of Systems Biology (ISB), for not only validation of results but also generating results in a common xml format. These tools perform robust statistical validations for peptides and proteins produced by various search algorithms based on the expectation maximization approach in a unified way and produce output in common xml-format independent of search algorithms result formats.

Although these tools are highly customized to work with Trans-Proteomic Pipeline (TPP) on a standalone computer, the APCF has developed software wrappers to allow the Prophet's to run in APCF cluster environment. These tasks are highly compute intensive depending on the user selected input parameters and hence we are customizing to run these tasks on APCF Compute Cluster. We also provided a virtual environment on the server to perform access controls checks to the data as well as generating dynamic web pages for the results to be viewed with any standard web browser.

APCF Post-processing pipeline with Prophets from ISB

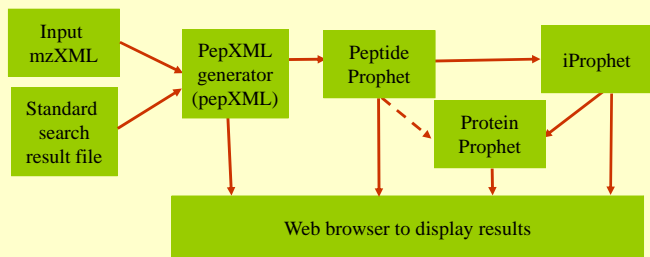


Figure 1 APCF post-process pipeline

APCF Post-processing results

We are using computational resources (compute nodes) in an optimized manner depending on the computational needs of the individual task. The results produced are in standard xml formats, such as pepXML (peptide results) and protXML (protein results) and are converted dynamically in HTML format on fly. The results so produced are viewed by any standard web browser without installing additional software on users system. Consistent with our previous efforts, we perform users credential checks but also adopt access control checks to see that the right user is receiving the right data.

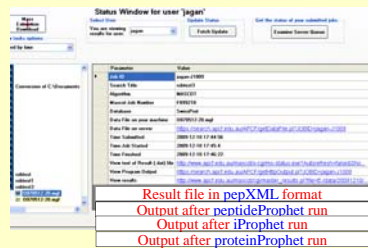


Figure 2 APCF UNITE

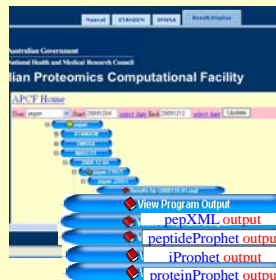


Figure 3 APCF UNIWI

PeptideProphet result analysis and Peptide 3D image

PeptideProphet automatically validates peptide assignments to MS/MS spectra made by database search programs. From each dataset, it learns distributions of search scores and peptide properties among correct and incorrect peptides, and uses those distributions to compute for each result a probability that it is correct.

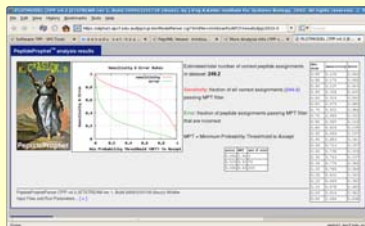


Figure 4 Peptide Prophet result analysis

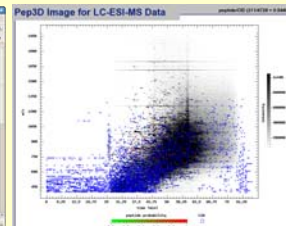


Figure 5 Peptide 3D Image

rank	sequence	score	length	protein	accession
1	SDSDVY	19.94	6	SDSDVY	Q95222
2	SDSDVY	19.94	6	SDSDVY	Q95222
3	SDSDVY	19.94	6	SDSDVY	Q95222
4	SDSDVY	19.94	6	SDSDVY	Q95222
5	SDSDVY	19.94	6	SDSDVY	Q95222

Figure 6 Browser view of Peptide Prophet results

APCF Security and Post-processing

Specially designed web based software that resides on the server hidden from the user, provides a unified, secure and consistent interface across all proteomics algorithms and post-processing tasks by following international standards with 128 bit SSL encryption and signed security certificates. A group security system allows users to share results within group.

The security mechanism installed at the APCF cluster validates user's credentials as well as perform access control checks and also provides secure communication channels between the user system and the facility.

APCF Post-processing highlights

AMS (APCF Meta Scheduler) follows the international standards and obeys APCF policies strictly in allocating resources to uses for returning search and post-process results to users at the quickest possible.

Current post-processing tasks are performed on the server and Integrate well with other APCF developments including cluster environment. Uses APCF security mechanism no separate login is required to view post-process results. Reuses part of existing TPP infrastructure without any modifications and provides unified post-processing interface between APCF UNITE and APCF UNIWI. The new trapping exclusively for post processing helps not only to create dynamic web pages but also allows to inject APCF security mechanism before returning results to users.

Uses standard web technologies, no additional software is required to be installed on the users (clients) system and works independent of any particular operating system. Familiar Interface for users who already had experience with TPP.

We also made available quantitative proteomics tools such as i-Tracker by running these tasks on the computational cluster. Other tools such as the proteomics search engine: Inspect from UCSD and other quantitative tools (e.g., x-Tracker) will be available soon.

The APCF can be accessed by a secure user account which can be obtained from the APCF at www.apcf.edu.au The APCF is open to all Australian and New Zealand researchers with the possibility of expanding the system for use by other countries.

References

- Empirical Statistical Model To Estimate the Accuracy of Peptide Identifications Made by MS/MS and Database Search: Andrew Keller, Alexey I. Nesvizhskii, Eugene Kolker, and Ruedi Aebersold, Anal. Chem., 2002,74(20),pp5383-5392.
- A Statistical Model for Identifying Proteins by Tandem Mass Spectrometry Alexey I. Nesvizhskii, Andrew Keller, Eugene Kolker, and Ruedi Aebersold, Anal. Chem., 2003,75(17),pp4646-4658. Institute for Systems Biology, 1441 North 34th Street, Seattle, Washington 98103.