Development and evaluation of a Galaxy-based one-click multi-stage database searching pipeline for proteogenomics analysis

Praveen Kumar[1], Jim Johnson[2], Pratik Jagtap[3], Timothy Griffin[3]
[1]Biomedical Informatics and Computational Biology, University of Minnesota Twin Cities;
[2]Minnesota Supercomputing Institute, University of Minnesota Twin Cities;
[3]Biochemistry, Molecular Biology, and Biophysics, University of Minnesota Twin Cities

Introduction

With the technology advancement in next-generation genomic sequencing and mass-spectrometry based proteomics, various methods has been developed to characterize genes by integrating genomic and proteomic data. This field of study is called proteogenomics (***Figure 1***).

Genome/Transcriptome sequencing data is used to for the generation of potentially expressed protein variants, such as single-amino acid variants (SAVs), splice junction peptides, etc. This produces a large number of possible sequences to be matched with (MS/MS) data.

Matching the tandem mass spectrometry (MS/MS) data to peptide sequences contained within a database confirms the presence of translated variant proteins in the sample.

One of the challenge proteogenomics analysis faces is the large databases. It increase false-positive identifications, leading to overestimation of FDR and loss of sensitivity for identifying true peptide spectrum matches (PSMs).

Additionally, the translated RNA-Seq database are usually appended to the database of known proteins that have been previously characterized, thus, when variants are matched to this single database there is no FDR estimation that applies specifically to the variants.

We developed the multi-stage database searching (multiDBsearch) method that can help in reducing the false positive identifications, and it may also help in optimizing the database composition being used for matching.

MultiDBsearch has possible advantages not only in proteogenomics, but also related metaproteomics studies, where large databases are used that contain protein sequences from thousands of microbes that may be present in certain samples of interest.

This implementation is available as a workflow on Galaxy-P (***Figure 3***) that can be shared with other researchers and integrated in other pipelines.

Figure 1

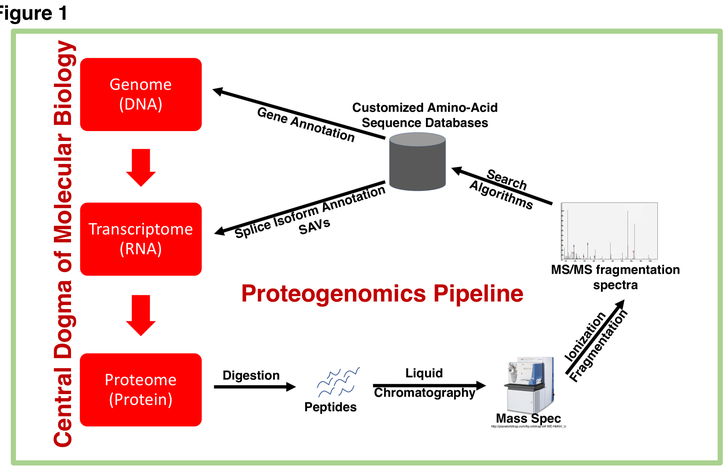

Figure 1: Overview of the proteogenomics pipeline within the context of the central dogma of molecular biology.

Method

In multiDBsearch method the MS/MS data is matched to separate sequence databases (such as the normal sequences, SAVs, frameshifts etc.) sequentially. MS/MS spectra matched successfully to peptide sequences at each successive stage are removed, and the remaining MS/MS spectra are matched to the next database.

To enable extraction of MS/MS spectra, which is identified by scan numbers, we developed a program called MS/MS Extractor that can read scan numbers from a PSMs report file, and remove those scan numbers from the MS/MS data.

The program is then integrated into a multi-step workflow where MS/MS data is matched with the first database. From the results, the matched spectra is removed (MS/MS with PSM). Next, the unmatched MS/MS data is used to match with next database and filter the matched spectra (***Figure 2***).

We also tested two-step search method, where a smaller database is created, based on the multiDBsearch protein identifications, and matched with MS/MS data.

This workflow is implement on Galaxy-P, a web-based proteomics cloud computing platform that enables easy sharing of workflows and datasets with other researchers (***Figure 3***).

<u>Dataset used for testing (from ABRF):</u>
MS/MS Dataset: Human proteins sample along with 4 spiked-in known proteins from other organisms:
(1) Beta-Gal from *Escherichia Coli*
(2) Lysozyme from *Gallus gallus*
(3) Protein-G from *Aspergillus niger*
(4) Amylase from *Streptococcus dysgalactiae*

Figure 2

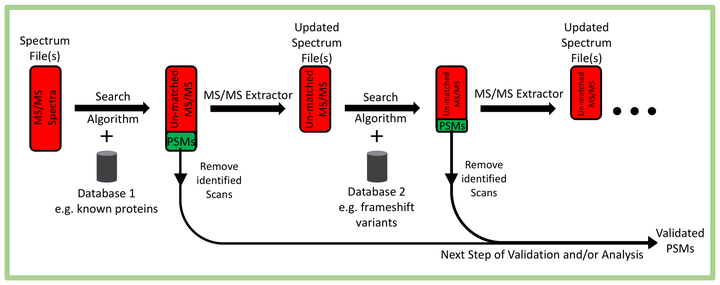

Figure 2: Pipeline describing the matching of MS/MS data against a database, removing the matched spectra, and matching the un-matched MS/MS spectra against next database. **The spectrum file(s) are in .mzML format and the databases are in .fasta format**

Evaluation Approach

MS/MS data matched with databases in 5 different ways:
1) Single step matching with combined Human proteome and 4 spiked-in proteins database
2) Single step matching with combined Human proteome and 4 organisms' proteome database
3) MultiDBsearch with Human proteome then each of the other organisms' proteome
4) Two-step searching method (Database: selected proteins + all human proteins)
5) Two-step searching method (Database: selected proteins + selected human proteins)

Comparing the number protein identification, peptide identification, and PSM identification from each approach.

Ideally, we should identify only the spiked-in proteins when MS/MS data is matched with those organisms' proteome.

Figure 3

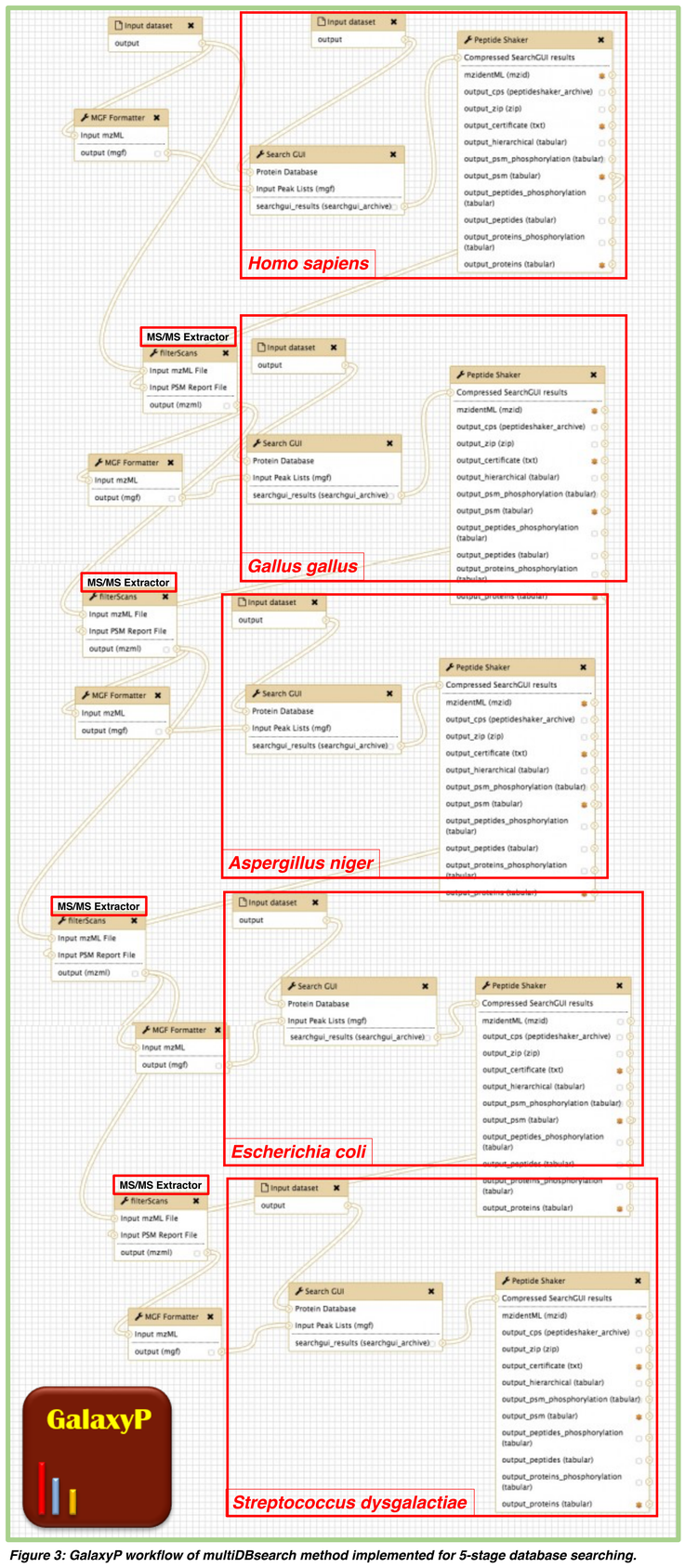

Figure 3: GalaxyP workflow of multiDBsearch method implemented for 5-stage database searching.

Figure 4

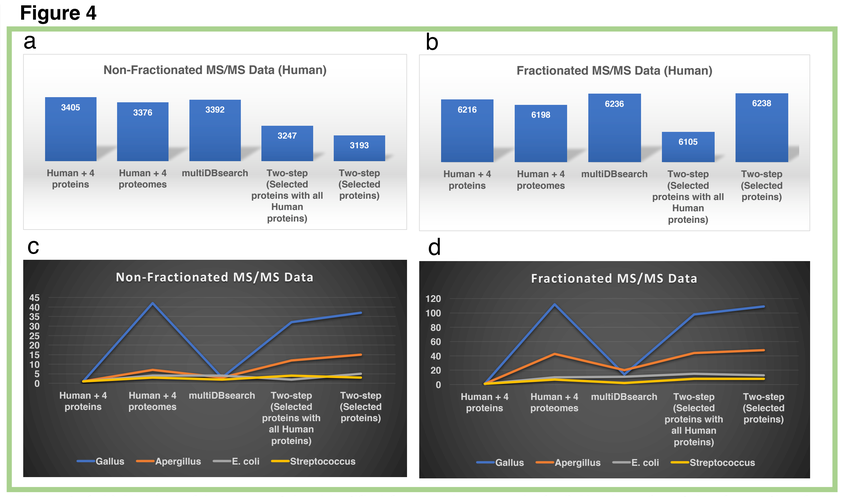

Figure 4: (a,b) Number of human proteins identified from each approach. (c,d) Number of proteins from other four organisms.

Figure 5

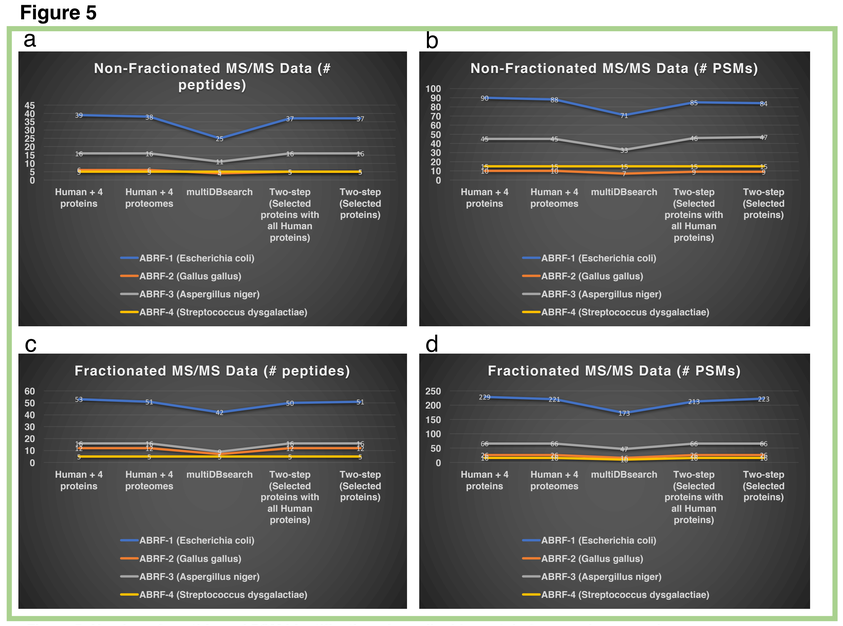

Figure 5: Number of peptide and PSM identifications for spiked-in protein from each approach

Results and Discussion

MultiDBsearch method gives better protein identification results *(Figure 4).*
- More number of human proteins are identified
- Less number proteins identified from other four organisms

MultiDBsearch method seems to hurt the peptide and PSM identification *(Figure 5)* that may lead to lower coverage, and in case of proteogenomics and metaproteomics study, it may miss some identifications

Adding two-step search to multiDBsearch method is helping in rescuing the peptides and PSMs missed by multiDBsearch.

MultiDBsearch with two-step searching method can be a better method to use in a proteogenomics and metaproteomics study.

Future Directions

- Using and evaluating this method on a proteogenomic dataset and metaproteomic datasets.
- More extensive peptide and PSM identification comparison.
- Using "workflow within workflow" to manage the repetitive processes in workflows on Galaxy-P.
- Expand the workflow to integrate two-step searching in a single Galaxy workflow.

Acknowledgements

- Data was generated through the collaborative work of the ABRF Proteomics Research Group (https://abrf.org/research-group/proteomics-research-group-prg).
- This project is supported by National Science Foundation (NSF) grant "1458524" and National Institutes of Health (NIH) grant "U24CA199347".

References

1. Nesvizhskii, A. I. Proteogenomics: concepts, applications and computational strategies. Nat. Methods 11, 1114–1125 (2014).
2. Nesvizhskii, A. I. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. J. Proteomics 73, 2092–2123 (2010).
3. Tanca, A. et al. Evaluating the Impact of Different Sequence Databases on Metaproteome Analysis: Insights from a Lab-Assembled Microbial Mixture. PLoS One 8, e82981 (2013).
4. Woo, S. et al. Advanced Proteogenomic Analysis Reveals Multiple Peptide Mutations and Complex Immunoglobulin Peptides in Colon Cancer. J. Proteome Res. 14, 3555–67 (2015).
5. Jagtap, P. et al. A two-step database search method improves sensitivity in peptide sequence matches for metaproteomics and proteogenomics studies. Proteomics 13, 1352–7 (2013).
6. Sheynkman, G. M. et al. Using Galaxy-P to leverage RNA-Seq for the discovery of novel protein variations. BMC Genomics 15, 703 (2014).
7. Jagtap, P. D. et al. Flexible and Accessible Workflows for Improved Proteogenomic Analysis Using the Galaxy Framework. J. Proteome Res. 13, 5898–5908 (2014).
8. Zhang, K. et al. A note on the false discovery rate of novel peptides in proteogenomics. Bioinformatics 31, 3249–53 (2015).
9. Blakeley, P., Overton, I. M. & Hubbard, S. J. Addressing Statistical Biases in Nucleotide-Derived Protein Databases for Proteogenomic Search Strategies. J. Proteome Res. 11, 5221–5234 (2012).
10. Renuse, S., Chaerkady, R. & Pandey, A. Proteogenomics. Proteomics 11, 620–630 (2011).
11. Sheynkman, G. M., Shortreed, M. R., Cesnik, A. J. & Smith, L. M. Proteogenomics: Integrating Next-Generation Sequencing and Mass Spectrometry to Characterize Human Proteomic Variation. Annu. Rev. Anal. Chem 9, 521–45 (2016).
12. Yates, J. R., Ruse, C. I. & Nakorchevsky, A. Proteomics by mass spectrometry: approaches, advances, and applications. Annu. Rev. Biomed. Eng. 11, 49–79 (2009).
13. Li, J. et al. A bioinformatics workflow for variant peptide detection in shotgun proteomics. Mol. Cell. Proteomics 10, M110.006536 (2011).
14. Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. Nature 422, 198–207 (2003).
15. Vertes, A. Mass spectrometry in proteomics. Med. Appl. Mass Spectrom. 173–194 (2008). doi:10.1016/B978-044451980-1.50010-0
16. Chambers, M. C. et al. A cross-platform toolkit for mass spectrometry and proteomics. Nat. Biotechnol. 30, 918–20 (2012).
17. Vaudel, M. et al. PeptideShaker enables reanalysis of MS-derived proteomics data sets. Nat. Biotechnol. 33, 22–24 (2015).
18. Vaudel, M., Barsnes, H., Berven, F. S., Sickmann, A. & Martens, L. SearchGUI: An open-source graphical user interface for simultaneous OMSSA and X!Tandem searches. Proteomics 11, 996–999 (2011).
19. Van Riper, S. et al. 'An ABRF-PRG study: Identification of low abundance proteins in a highly complex protein sample' at the 64th Annual Conference of American Society of Mass Spectrometry and Allied Topics" at San Antonio, TX

UNIVERSITY OF MINNESOTA
Driven to Discover℠