

Adapting Existing Proteomics Tools for Massively Parallel Peptide Discovery (MPPD)

William D. Nelson¹, Kert Viele² and Bert C. Lynn³

¹Department of Biology, ²Department of Statistics and ³Department of Chemistry, University of Kentucky, Lexington, KY 40506, USA

We have integrated the analysis pipeline of the Computational Proteomics Analysis System (CPAS)^{1,2} with Amazon Web Services' (AWS)³ cloud computing to provide low cost High Performance Computing (HPC) resources for the discovery of modified peptides. The integration uses standard analysis tools, including the Trans Proteomics Pipeline (TPP)⁴ and the X!Tandem⁵ database search engine, to facilitate validation and reusability of the analysis pipeline.

The most common implementation method for cloud computing is to install an entire informatics platform on the cloud's virtual computers and then access the platform remotely through a web interface. Instead, our implementation leaves the core of the informatics platform in-house and only uses the cloud resources to provide HPC cluster nodes for computationally intensive tasks. With this design proteomics researchers have the security and configurability of an in-house system as well as the low cost HPC resources of cloud computing.

The number of variable peptide modifications that can be searched simultaneously with standard protein database search engines is limited because of all the permutations possible with large numbers of modifications in combination. This complexity decreases the number of correct peptide identifications as well as the speed of the searches. Previous work has demonstrated that X!Tandem's search speed can be increased by performing a single search on multiple computers^{6,7} but this will not address the problem of peptide identification errors. Specialized peptide modification discovery tools^{8,9} run separate searches for each peptide modification to decrease peptide identification errors but these tools are not easily integrated into standard analysis pipelines. Running a separate X!Tandem search for every modification using parallel nodes in the cloud addresses both speed and peptide identification issues while facilitating further analysis of the search results with other standardized proteomics analysis tools.

To test the scalability of MPPD we ran the application with 73 PeptideAtlas¹⁰ files from fractions of the yeast proteome¹¹. A benchmark search with oxidized methionine as the single variable modification was run versus an MPPD search with 8 different variable modifications. MPPD, installed on a laptop computer, routed the 73 files through the complete TPP using X!Tandem as the search engine. The search was completed in less than 2 hours, on 584 Amazon nodes, at a cost of less than \$50. Compared to the benchmark, the MPPD search discovered 17% more peptides (1100) and 7.6% additional proteins (73) with a false discovery rate of less than 1%.

References

- (1) Rauch A.; Bellew M.; Eng J.; Fitzgibbon M.; Holzman T.; Hussey P.; Igra M.; Maclean B.; Lin C. W.; Detter A.; Fang R.; Faca V.; Gafken P.; Zhang H.; Whitaker J.; States D.; Hanash S.; Paulovich A.; McIntosh M. W. Computational Proteomics Analysis System (CPAS): an extensible, open-source analytic system for evaluating and publishing proteomic data and high throughput biological experiments. *J. Proteome Res.* **2006** 5,112-21.
- (2) MacLean B.;Malick P. Adaptable Enterprise Pipeline for the Computational Proteomics Analysis System. *ASMS Poster* **2008** <http://www.labkey.com/resources/posters/2008ASMS-AdaptableEnterprisePipeline-1.pdf>.
- (3)Amazon Web Services: <http://aws.amazon.com/> .
- (4) Keller, A., Eng, J., Zhang, N., Li, X. J., and Aebersold, R. A uniform proteomics MS/MS analysis platform utilizing open XML file formats *Mol. Syst. Biol.* **2005** 1, doi:10.1038/msb4100024.
- (5) Craig R.; Beavis R. C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **2004** 20, 1466-1467.
- (6) Halligan, B., Geiger, J., Vallejos A., Greene A., Twigger, S. Low Cost, Scalable Proteomics Data Analysis Using Amazon's Cloud Computing Services and Open Source Search Algorithms *J. Proteome Res.* **2009** 8, 3148–3153.
- (7) Duncan D., Craig R., Link A. Parallel Tandem: A Program for Parallel Processing of Tandem Mass Spectra Using PVM or MPI and X!Tandem *J. Proteome Res.* **2005** 4, 1842-1847.
- (8) Tanner S., Payne S.H., Dasari S., Shen Z., Wilmarth P.A., David L. L., Loomis W. F., Briggs S., Bafna V. Accurate Annotation of Peptide Modifications through Unrestrictive Database Search. *J. Proteome Res.* **2008** 7, 170-181, DOI: 10.1021/pr070444v.
- (9) Creasy D., Cottrell J. Error tolerant searching of uninterpreted tandem mass spectrometry data *Proteomics* **2002** 2, 1426–1434.
- (10) Deutsch E., Lam H., Aebersold R. PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows *EMBO reports* **2008**, 9, 429–434.
- (11) Peng J., Elias J. E., Thoreen C. C., Licklider L. J., Gygi S. P. Evaluation of Multidimensional Chromatography Coupled with Tandem Mass Spectrometry (LC/LC-MS/MS) for Large-Scale Protein Analysis: The Yeast Proteome. *J. Proteome Res.* **2003** 2, 43-50.

