BIOINFORMATICS APPLICATIONS NOTE

Vol. 26 no. 8 2010, pages 1131–1132
doi:10.1093/bioinformatics/btq081

Data and text mining

Advance Access publication February 25, 2010

CPFP: a central proteomics facilities pipeline

David C. Trudgian[1,2,*], Benjamin Thomas[1], Simon J. McGowan[3], Benedikt M. Kessler[2], Mogjiborahman Salek[1] and Oreste Acuto[1]

[1]Sir William Dunn School of Pathology, University of Oxford, South Parks Road, Oxford OX1 3RE, [2]Henry Wellcome Building for Molecular Physiology, University of Oxford, Roosevelt Drive, Oxford OX3 7BN and [3]Computational Biology Research Group, University of Oxford, John Radcliffe Hospital, Oxford OX3 9DS, UK

Associate Editor: John Quackenbush

ABSTRACT

Summary: The central proteomics facilities pipeline (CPFP) provides identification, validation, and quantitation of peptides and proteins from LC-MS/MS datasets through an easy to use web interface. It is the first analysis pipeline targeted specifically at the needs of proteomics core facilities, reducing the data analysis load on staff, and allowing facility clients to easily access and work with their data. Identification of peptides is performed using multiple search engines, their output combined and validated using state-of-the-art techniques for improved results. Cluster execution of jobs allows analysis capacity to be increased easily as demand grows.

Availability: Released under the Common Development and Distribution License at http://cpfp.sourceforge.net/. Demonstration available at https://cpfp-master.molbiol.ox.ac.uk/cpfp_demo

Contact: dctrud@ccmp.ox.ac.uk

Received on November 24, 2009; revised on January 29, 2010; accepted on February 19, 2010

1 INTRODUCTION

Mass spectrometry (MS) has emerged as the dominant technique for the identification and quantitation of peptides and proteins from complex mixtures. A large number of institutions have established core proteomics facilities to provide MS services, sharing equipment and expertise with a wide range of users (Ogorzalek-Loo *et al.*, 2009). With the growth in demand for high-throughput LC-MS/MS analysis of complex samples, and increased interest in quantitative proteomics, effective analysis of data can be challenging. Existing freely available pipelines such as TPP (Keller *et al.*, 2005), CPAS (Labkey Software Foundation), OpenMS/TOPP (Kohlbacher *et al.*, 2007), SwissPIT (Quandt *et al.*, 2008) and MASSPECTRAS (Hartler *et al.*, 2007) support a variety of proteomics search engines and validation tools, but often require users to understand the various parameter formats of the search engines, reconcile differences in post-translational modification (PTM) specifications and manually run multiple searches if more than one search engine is to be used. Commercial applications such as Scaffold (Proteome Software, Portland OR, USA), PEAKS (Bioinformatics Solutions Inc., Waterloo ON, Canada) and Sorcerer (Sage-N Research, Milpitas CA, USA) provide user-friendly interfaces but require significant outlay if multiple licences are required. The central proteomics facilities pipeline (CPFP) aims to provide a simple interface for core facility staff and clients, and to fully automate the analysis of MS/MS data with multiple search engines.

*To whom correspondence should be addressed.

2 FEATURES

CPFP accepts datasets of LC-MS/MS spectra in mzXML, mzML, pkl and mgf formats. Files of unlimited size can be uploaded for analysis via a web browser. Identification of peptides from spectra is performed using Mascot (Matrix Science, London, UK), OMSSA (Geer *et al.*, 2004) and X!TANDEM (Craig and Beavis, 2004). X!TANDEM searches may be performed using the native or *k*-score algorithms (Maclean *et al.*, 2006). A single web form allows submission to all search engines using common parameters. Translation of parameters into the formats required by the individual search engines is performed automatically. PTM definitions may be imported from Unimod (Creasy and Cottrell, 2004) for use with all search engines.

Validation of results is performed using the TPP analysis tools. Peptide identifications from each search engine are validated with PeptideProphet (Keller *et al.*, 2002), and then combined using iProphet (Shteynberg *et al.*, 2008). Finally, protein identifications are inferred using ProteinProphet (Nezvizhskii *et al.*, 2003). Quantitation can be performed using LIBRA for iTRAQ-labelled samples, and ASAPRatio (Li *et al.*, 2003) for heavy isotope-labelled samples. All searches are performed against concatenated target/decoy sequence databases. Results can be viewed at the 1 and 5% false discovery rates (FDRs) as calculated by Peptide/ProteinProphet or estimated empirically from decoy hits (Elias and Gygi, 2007). External functionality such as BLAST searches against identified peptides, and submission of spectra for spectral search against public datasets, is inherited through the use of the TPP PepXML and ProtXML viewers. Graphs indicating the quality of results can be viewed for each search, and users may download result files for further analysis with additional software.

Submission and search details are recorded in a database. User authentication can be integrated with existing systems. Users who submit data may grant others access to view their results. A basic administrative interface is provided to configure sequence databases and quantitation methods based on ASAPRatio.

3 IMPLEMENTATION

CPFP consists of a web application, relational database and collection of pipeline scripts. It is written in Perl using the Catalyst Web Framework and is intended for installation on recent

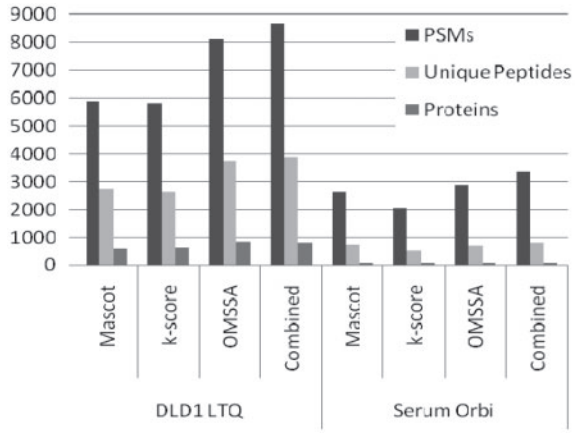

Fig. 1. PSMs, unique peptide identifications and protein identifications for Mascot, X!Tandem k-score, OMSSA and combined searches performed on the test datasets using CPFP.

Linux systems. GridEngine (Sun Microsystems) is used to provide job scheduling across a cluster, allowing computing infrastructure to be scaled as demand increases. Mascot searches are submitted via HTTP, allowing the use of a Mascot server that is not part of the cluster. The pipeline may be used without Mascot if it is not available.

4 RESULTS

The DLD-LTQ and Serum-Orbi datasets described in Ma *et al*. (2009) were analysed using CPFP. Searches used the IPI-Human sequence database v3.64 (Kersey *et al*., 2004) and the parameters given in Ma *et al*. (2009). Results were filtered to an estimated 1% FDR using the target–decoy procedure. Figure 1 shows the number of identifications for Mascot, X!TANDEM k-score and OMSSA searches processed separately, and a combined result merged using iProphet. Combining the results of the three search engines gives a higher number of peptide spectrum matches (PSMs) and unique peptide identifications versus the best performing single search engine (OMSSA in both cases). The increase in PSMs is 6.6 and 14.4% for the DLD1-LTQ and Serum-Orbi datasets, respectively. Submission to and combination of results from all search engines is automated, and required no additional steps for the user versus the use of a single search engine. Timings for the Serum-Orbi dataset on a 32-core cluster were 9 min 47 s for Mascot, 7 min 31 s for X!TANDEM and 11 min 50 s for OMSSA individually, versus 16 min 45 s for all searches in parallel, followed by combination of results.

5 FUTURE DEVELOPMENT

CPFP is in daily use and under active development. The application has been released under an open source licence so that it may be used and adapted by other groups. Work is ongoing to incorporate export of results to PRIDE XML format (Jones *et al*., 2006) and allow integration with a local PRIDE repository. Scripts for the generation of inclusion and exclusion lists will be incorporated into the pipeline, reducing the effort necessary to use multi-injection MS workflows to increase sample coverage. PTM identification and validation is a priority, which will involve the incorporation of non-TPP analysis tools and alternative workflows, similar to those previously implemented in SwissPIT. A desktop Java client for bulk submission of data is under development.

ACKNOWLEDGEMENTS

We wish to acknowledge the Computational Biology Research Group, Medical Sciences Division, Oxford, for use of their services in this project. We thank Zong-Pei Han for computing support and OSSWatch for advice regarding open source licensing.

Funding: E.P. Abraham Cephalosporin Trust (to O.A.); John Fell OUP Research Fund (to B.M.K.).

Conflict of Interest: none declared.

REFERENCES

Craig,R. and Beavis,R.C. (2004) TANDEM: matching proteins with tandem mass spectra. *Bioinformatics*, **20**, 1466–1467.

Creasy,D.M. and Cottrell,J.S. (2004) Unimod: protein modifications for mass spectrometry. *Proteomics,* **4**, 1534–1536.

Elias,J.E. and Gygi,S.P. (2007) Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods*, **4**, 207–214.

Geer,L.Y. *et al.* (2004) Open mass spectrometry search algorithm. *J. Proteome Res*., **3**, 958–964.

Hartler,J. *et al.* (2007) MASPECTRAS: a platform for management and analysis of proteomics LC-MS/MS data. *BMC Bioinformatics*, **8**, 197.

Jones,P. *et al.* (2006) PRIDE: a public repository of protein and peptide identifications for the proteomics community. *Nucleic Acids Res.*, **1**, D659–D663.

Keller,A. *et al.* (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem*., **74**, 5358–5392.

Keller,A. *et al.* (2005) A uniform proteomics MS/MS analysis platform utilizing open XML formats. *Mol. Syst. Biol.*, **1**, 2005.0017.

Kersey,P.J. *et al.* (2004) The International Protein Index: an integrated database for proteomics experiments, *Proteomics,* **4**, 1985–1988.

Kohlbacher,O. *et al.* (2007) TOPP – the OpenMS proteomics pipeline. *Bioinformatics*, **23**, e191–e197.

Li,X.-J. *et al.* (2003) Automated statistical analysis of protein abundance ratios from data generated by stable isotope dilution and tandem mass spectrometry. *Anal. Chem*., **75**, 6648–6657.

Ma,Z.-Q. *et al.* (2009) IDPicker 2.0: improved protein assembly with high discrimination peptide identification filtering. *J. Proteome Res.*, **8**, 3872–3881.

Maclean,B. *et al.* (2006) General framework for developing and evaluating database scoring algorithms using the TANDEM search engine. *Bioinformatics*, **22**, 2830–2832.

Nezvizhskii,A.I. *et al.* (2003) A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.*, **75**, 2279–2287.

Ogorzalek-Loo,R. *et al.* (2009) Association of Biomolecular Resource Facilities Survey: Service Laboratory Funding. *J. Biomol. Tech.*, **20**, 180–185.

Quandt,A. *et al.* (2008) swissPIT: a novel approach for pipelined analysis of mass spectrometry data. *Bioinformatics*, **24**, 1416–1417.

Shteynberg,D. *et al.* (2008) iProphet: improved validation of peptide and protein ids in the trans-proteomic pipeline. Poster session at: *HUPO 7th Annual World Congress*, August 16–20, Amsterdam.