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., 2007). There is a 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.

2 FEATURES

CPFP accepts datasets of LC-MS/MS spectra in mzXML, mzML, pkI, 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), OMSA (Gect 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 (Nezuizhki 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 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. 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


D.C. Trudgian et al.