BIOINFORMATICS APPLICATIONS NOTE Vol. 24 no. 11 2008, pages 1416–1417
doi:10.1093/bioinformatics/btn139

Data and text mining

swissPIT: a novel approach for pipelined analysis of mass spectrometry data

Andreas Quandt[1,*], Patricia Hernandez[1], Alexandre Masselot[1,2], Céline Hernandez[1], Sergio Maffioletti[3], Cesare Pautasso[4], Ron D. Appel[1,5] and Frederique Lisacek[1]

[1]Swiss Institute of Bioinformatics, Proteome Informatics Group, Geneva, [2]Geneva Bioinformatics (GeneBio) S.A., Geneva, [3]Swiss National Supercomputing Centre, Manno, [4]University of Lugano, Faculty of Informatics, Lugano and [5]University of Geneva, Computer Science Department, Geneva, Switzerland

Received on March 14, 2008; revised and accepted on April 10, 2008

Advance Access publication April 23, 2008

Associate Editor: Jonathan Wren

ABSTRACT

The identification and characterization of peptides from tandem mass spectrometry (MS/MS) data represents a critical aspect of proteomics. Today, tandem MS analysis is often performed by only using a single identification program achieving identification rates between 10–50% (Elias and Gygi, 2007). Beside the development of new analysis tools, recent publications describe also the pipelining of different search programs to increase the identification rate (Hartler *et al*., 2007; Keller *et al*., 2005).

The Swiss Protein Identification Toolbox (swissPIT) follows this approach, but goes a step further by providing the user an expandable multi-tool platform capable of executing workflows to analyze tandem MS-based data. One of the major problems in proteomics is the absent of standardized workflows to analyze the produced data. This includes the pre-processing part as well as the final identification of peptides and proteins. The main idea of swissPIT is not only the usage of different identification tool in parallel, but also the meaningful concatenation of different identification strategies at the same time. The swissPIT is open source software but we also provide a user-friendly web platform, which demonstrates the capabilities of our software and which is available at http://swisspit.cscs.ch upon request for account.

Contact: andreas.quandt@isb-sib.ch

1 INTRODUCTION

Mass spectrometry (MS) has become a method of choice for analyzing complex protein samples (Aebersold and Mann, 2003). In recent years, algorithms have been developed to match experimental spectra with sequence entries in databases such as Uniprot/Swissprot and Uniprot/TrEMBL. In this procedure, the major part of the spectra set is often not identified, e.g. due to their low quality, and/or the existence of unexpected modifications (Palagi *et al*., 2006). It was frequently suggested to use multiple tools to increase the identification rate (Hernandez *et al*., 2006). With the Swiss Protein Identification Toolbox (swissPIT), we introduce the first integrated analysis platform that combines several programs and search strategies in analysis workflows for identifying proteins with MS/MS spectra and present results in a unified visualization.

2 CAPABILITIES

Applying multiple search tools on a data set is recommended for producing more confident matches by cross-validating the matches of each search engine. The strategy of running parallel searches with commercial and non-commercial products was validated in recent publications resulting in open-source pipelines such as the Trans-Proteomic Pipeline (Keller *et al*., 2005) and commercial products such as Sage N and Scaffold. However, the combination of identification tools with different strategies analyzing workflows is a novel approach. Currently, swissPIT integrates four identification programs: the classical search tools Phenyx (Colinge *et al*., 2003) and X! Tandem (Craig and Beavis, 2003) as well as the open-modification search approaches Popitam (Hernandez *et al*., 2003) and InsPect (Tanner *et al*., 2005), which are optimized to identify spectra with unexpected modifications. As an integrated analysis system, swissPIT supports various data formats, such as mgf, dta, pkl, mzXML and mzData, as well as the data upload of compressed archives in zip and tar.gz format. A further advantage of swissPIT is the common database type, database version and taxonomy that all identification programs use in their calculation process. This makes the comparison of results easier for the user. swissPIT runs at present with the UniProt databases SwissProt (Release 50.7) and TrEMBL (Release 32.6). The taxonomy entries cover the main groups root, bacteria, mammalia, fungi, viruses and also offer access to several related subgroups.

3 WORKFLOWS

At the moment, our web platform provides two pre-defined workflows. They implement the two following scenarios: (A) an independent data analysis with multiple tools and (B) a sequential analysis where the classical search tools are applied in a first stage followed by the open-modification search tools which process the information gained in the previous step (e.g. the accession numbers, list of unmatched spectra).

*To whom correspondence should be addressed.

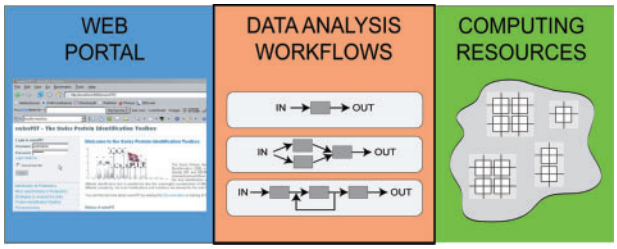

Fig. 1. The three components of our web platform (http://swisspit.cscs.ch): the web interface for user interaction and result visualization, the execution of analysis workflows and the usage of cluster computing resources.

Additionally, the user can select programs to be run in the workflow that was chosen for analyzing the data. We control the execution of workflows with JOpera as workflow manager (Pautasso *et al.*, 2006). In order to incorporate commercial software such as Phenyx, adapter classes for JOpera execute all identification tools. New software can easily be incorporated in any workflow only by writing a specific adapter class that will handle the program execution in its batch mode.

4 USAGE

For each program the user is expected to select a parameter file to be used for processing. All parameters and their values are contained in a template file in a form directly usable by the program. Default templates are provided for each of the programs integrated into swissPIT. Alternatively, since result files generated with Phenyx or X! Tandem contain the parameters used during the processing, any of such files can be uploaded. After the job submission, the user can access the job page where each job submission creates a new entry in the personal file space of the user. A status bar shows the user if submissions are still in process or if results are accessible. After a job is successfully finished, the user can review the original result and parameter files for each program within its corresponding subfolder of the job directory.

To also provide a unified visualization of results across the identification programs, we are using parts of the commercial Phenyx software by importing all results in its web interface and referring to them with external links from our web interface. The single result can be accessed by clicking on the status box of each identification program when processing is over. The results of multiple programs within a job submission can be compared while using the status box as a hyperlink to a global view of all matches for each program.

5 RESULTS

With our platform (Fig. 1), the user can process sets of spectra with multiple tools in a single submission. On the one hand, the performance of programs can be compared and on the other hand, the combination of different search strategies such as strict search and open-modification search can be tested to increase the identification rate. In spite of the fact that swissPIT is still a young project, it is already a service assisting scientist in analyzing MS data by providing access to different identification tools, calculation of processes in parallel and standardization of several parameters shared between the programs. The final outcome is a more meaningful comparison of calculated results across different programs.

6 DEPENDENCIES

swissPIT consists of two parts: a core package and a user-friendly web interface both are written in Java. The core package requires the JDK 1.6 and is distributed a single jar file. The web interface part uses techniques such as JSP, Ajax and Struts and can be deployed with an application server such as Tomcat (http://tomcat.apache.org). As swissPIT interacts with the Phenyx web interface (GeneBio, S.A., http://www.genebio.com) to unify the result visualization, we encourage its installation in order to use this user-friendly feature. Otherwise, results are shown in their original format (text or XML). Furthermore, all identification tools and programs combined in workflows have to be available and configured in swissPIT's property file. The workflows are executed and controlled by JOpera, a recent workflow manager not tied to a specific research subject or infrastructure. These tasks are complicated enough for us to recommend the use of the web interface that was precisely designed to save the average user of all this work.

ACKNOWLEDGEMENTS

We would like to thank Peter Kunszt (CSCS), Heinz Stockinger, and Ioannis Xenarios (both Vital-IT) for many fruitful discussions and advices.

Funding: This work was supported by the Swiss National Science Foundation (FNS grant 3100A0-113456).

Conflict of Interest: none declared.

REFERENCES

Aebersold,R. and Mann,M. (2003) Mass spectrometry-based proteomics. *Nature*, **422**, 198–207.

Colinge,J. *et al.* (2003) OLAV: towards high-throughput tandem mass spectrometry data identification. *Proteomics*, **3**, 1454–1463.

Craig,R. and Beavis,R.C. (2003) A method for reducing the time required to match protein sequences with tandem mass spectra. *Rapid Commun. Mass Spectrom*, **17**, 2310–2316.

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.

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

Hernandez,P. *et al.* (2003) Popitam: towards new heuristic strategies to improve protein identification from tandem mass spectrometry data. *Proteomics*, **3**, 870–878.

Hernandez,P. *et al.* (2006) Automated protein identification by tandem mass spectrometry: issues and strategies. *Mass Spectrom Rev.*, **25**, 235–254.

Keller,A. *et al.* (2005) A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Mol. Syst. Biol.*, **1**, msb4100024-E1–msb4100024-E8.

Palagi,P.M. *et al.* (2006) Proteome informatics I: bioinformatics tools for processing experimental data. *Proteomics*, **6**, 5435–5444.

Pautasso,C. *et al.* (2006) *Autonomic Computing for Virtual Laboratories, Dependable Systems: Software, Computing, Networks*, Springer, Heidelberg.

Tanner,S. *et al.* (2005) InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. *Anal. Chem.*, **77**, 4626–4639.