Journal of Biotechnology 261 (2017) 126–130

Contents lists available at ScienceDirect

Journal of Biotechnology

journal homepage: www.elsevier.com/locate/jbiotec

Computational proteomics tools for identification and quality control

Dominik Kopczynski[a], Albert Sickmann[a,b,c], Robert Ahrends[a,*]

[a] Leibniz-Institut für Analytische Wissenschaften – ISAS – e.V., Dortmund, Germany
[b] College of Physical Sciences, University of Aberdeen, Meston Building, Old Aberdeen, United Kingdom
[c] Medizinische Fakultät, Ruhr-Universität Bochum, Bochum, Germany

ARTICLE INFO

Keywords:
Computational proteomics
Protein inference
Peptide identification
Peptide spectra library database

ABSTRACT

Computational proteomics is a constantly growing field to support end users with powerful and reliable tools for performing several computational steps within an analytics workflow for proteomics experiments. Typically, after capturing with a mass spectrometer, the proteins have to be identified and quantified. After certain follow-up analyses, an optional targeted approach is suitable for validating the results. The de.NBI (German network for bioinformatics infrastructure) service center in Dortmund provides several software applications and platforms as services to meet these demands. In this work, we present our tools and services, which is the combination of SearchGUI and PeptideShaker. SearchGUI is a managing tool for several search engines to find peptide spectra matches for one or more complex MS^2 measurements. PeptideShaker combines all matches and creates a consensus list of identified proteins providing statistical confidence measures. In a next step, we are planning to release a web service for protein identification containing both tools. This system will be designed for high scalability and distributed computing using solutions like the Docker container system among others. As an additional service, we offer a web service oriented database providing all necessary high-quality and high-resolution data for starting targeted proteomics analyses. The user can easily select proteins of interest, review the according spectra and download both protein sequences and spectral library. All systems are designed to be intuitively and user-friendly operable.

1. Introduction

During the past decade, proteomics has gained more and more impact on the field of analytical biochemistry. The constantly growing number of mass spectrometry based analytical data for proteomics experiments creates a demand for powerful and reliable tools for analysis. Typical experimental workflows have roughly the same structure. Depending on the scientific question, proteins from a desired biological system are being extracted and purified. Several analytical devices like liquid chromatography, mass spectrometry or ion mobility spectrometry are separating complex mixtures of peptides according to certain chemical and physical properties. Typically, a combination of liquid chromatography and tandem mass spectrometry (LC–MS/MS) is utilized in the proteomics field. Here, bottom-up experiments, where the proteins are dissociated by tryptic digestion into small peptides, are currently more utilized in the field of proteomics than top-down experiments (Gillet et al., 2016), where the proteins are analyzed as a whole. After data acquisition, the proteins have to be identified. Therefore, the peptides are identified within the MS^2 spectra, which is one of the most time consuming computational parts. Subsequently, so-called peptide spectra matches (PSMs) are used, to deduce corresponding proteins. In an additional step, proteins can be relatively or absolutely quantified which is not only depending on scientific question but also on time and cost. If several measurements exist, all measurements have to be aligned with each other, subsequently. According to the scientific question, several follow-up analyses can be performed.

An optional step is the validation of the results by performing a targeted proteomics approach for quality control. Several of the listed steps require a high amount of computation, for instance protein inference, where the presence of proteins is determined based on the identified peptides. Many analysis tools have been developed to perform the computation of all particular steps within the listed workflow. These tools often lack the fact that they are proof-of-concept programs or provided in a way that only a limited number of domain experts can launch or operate. In the following, the services provided by the de.NBI service center BioInfra.Prot partner in Dortmund are introduced, which are in particular the interplay of the in-house developed SearchGUI and PeptideShaker for protein identification and the interactive database for peptide reference spectra which is necessary for targeted proteomics and absolute quantification. Finally, we give an outlook of a web based

* Corresponding author.
E-mail address: robert.ahrends@isas.de (R. Ahrends).

http://dx.doi.org/10.1016/j.jbiotec.2017.06.1199
Received 17 February 2017; Received in revised form 13 June 2017; Accepted 21 June 2017
Available online 01 July 2017
0168-1656/ © 2017 Elsevier B.V. All rights reserved.

protein identification platform as the next service we are currently establishing for a broad user base. All services were designed to be as intuitively and user-friendly as possible and yet offer a broad functionality to close the gap between the developers and the end users.

2. Identifying proteins with SearchGUI and PeptideShaker

Having an MS^2 measurement of a proteomics experiment, a first computer assisted analysis step is to identify all proteins in the measurement. Therefore, SearchGUI and PeptideShaker were in-house developed to perform this task consecutively in a two-stage approach. Since proteins are digested before being analyzed by the mass spectrometer, only the spectra of peptides are being acquired. In order to achieve full protein identification, SearchGUI is utilized to identify peptides within single MS^2 spectra (resulting in PSMs) as the first step of the computational analysis. In a second step, PeptideShaker is using the identification results and building a consensus identification on the protein level.

SearchGUI (Vaudel et al., 2011) is a management tool for peptide search engines, namely X!Tandem (Fenyö and Beavis, 2003), MS-GF+ (Kim and Pevzner, 2014), MS Amanda (Dorfer et al., 2014), MyriMatch (Tabb et al., 2007), Comet (Eng et al., 2012), Tide (Diament and Noble, 2011), Andromeda (Cox et al., 2011), and OMSSA(Geer et al., 2004) or Novor (Ma, 2015) and DirecTag (Tabb et al., 2008) for *de novo* identification. It enables a fast and intuitive selection of the tools for the analysis. Global parameters that are used within all search engines only need to be adjusted once whereas SearchGUI also allows adjusting special parameters required by specific search engines. SearchGUI is a bundle of the tool itself and all dependent search engines. Thus, no time consuming program search and implementation on the host system is necessary. Since most of the listed tools are command line based, the usage of SearchGUI shortens the startup time significantly. SearchGUI itself is built on the compomics-utilities (Barsnes et al., 2011) providing a broad range of proteomics functions. Although SearchGUI is written in Java and thus platform independent, its dependencies are mostly platform dependent binaries. SearchGUI provides an intuitive graphical user interface as well as an extensive command line interface. The latter feature is in particular suitable for a usage in a high-throughput context or in a workflow system. SearchGUI reads MS^2 files in Mascot generic format (MGF) (Deutsch, 2012) which is a text based file format to store MS^2 spectra along with related meta information on MS^2 level, like spectrum ID, precursor mass, charge, or retention time. As a requirement for the database search engines, SearchGUI also reads FASTA files which is a quasi-standard for providing protein sequences. It is a text based file and contains the protein sequences coded in standard IUB/IUPAC amino acid codes.[1] Once a search is finalized, SearchGUI stores all results in a ZIP file that can easily be forwarded to PeptideShaker.

PeptideShaker (Vaudel et al., 2015) is a protein inference tool with several features. As the main function, PeptideShaker can read the results of all common database search engines as well as from *de novo* tools and re-analyze the results in a combined manner to create a consensus result. It also has both an intuitive graphical user interface as well as an extensive command line interface allowing headless operation of all analysis steps. Reading the search results from SearchGUI, PeptideShaker automatically adopts all parameters from the peptide identification without redundant submission including the location of all input files. However, an additional adjustment of the parameters is still possible. PeptideShaker is able to read both FASTA and MGF files and the result files of all mentioned peptide search engines. Even raw Mascot result files (DAT) from the Mascot search engine (Perkins et al., 1999) (which is the most common commercial peptide database search engine) can be parsed. The results of the protein inference are presented in a user-friendly graphical interface as illustrated in Fig. 1. Identified proteins can be reviewed in the top window. All peptides mapped to a selected protein are shown in the center window whereas all spectra that match to a selected peptide are listed in the bottom window along with the annotated spectrum on the right hand side. For increasing the confidence of the search results, PeptideShaker uses the target-decoy approach, where half of the search space i.e. the protein database contains true protein sequences and the other half contains so-called decoy sequences. Typical decoy sequences are the reversed sequences of all proteins.

By tracking which peptides are being mapped to a decoy, a false discovery rate can be set as a confidence measure for an additional filtering to achieve high-quality controlled results. PeptideShaker has several export functions and predefined report forms at several stages like on protein or peptide level. It also supports exporting results in file formats for follow-up analyses e.g. for Progenesis QI.[2] Furthermore, PeptideShaker can provide the identification results in the open standard mzIdentML format[3] which is a mandatory file for public proteomics dataset repositories like PRIDE (Perez-Riverol et al., 2016; Vizcaíno et al., 2014) or can be used for creating spectral libraries which are necessary for targeted proteomics as explained in Section 4. Accordingly, several different analyses e.g. an integrative data analysis or pathway network analysis using the protein set enrichment analysis method (Lavallée-Adam et al., 2014) or a correlation based method (Köberlin et al., 2015) can be performed, consecutively.

As a unique feature among all protein inference tools, PeptideShaker has the ability to re-analyze public proteomics datasets that were published in repositories like PRIDE. Published datasets have the potential to increase the value of extracted information of the data especially when re-analyzing under a different biological question or purpose. PeptideShaker allows starting in 'reshake' mode. Accordingly, the complete PRIDE dataset list is being extracted. The user has the opportunity to re-start the analysis by launching SearchGUI and PeptideShaker with the selected public datasets. The corresponding spectra and search parameters are downloaded for re-analysis. Due to coping with all utilized open standard file formats, PeptideShaker can handle the datasets readily and reduce the complexity for re-analysis into a few clicks or commands. This outstanding feature is unique among all protein identification and analysis tools. Both programs SearchGUI[4] and PeptideShaker[5] are open source, free of charge, and licensed under the Apache2 license.

3. Interactive archive for peptide reference spectra

After a successful analysis of the proteomics data, the complete sample analysis workflow contains a last step, namely the validation of the results. Here, targeted proteomics is a suitable method for validation. Methods such as selected reaction monitoring (SRM) or parallel reaction monitoring (PRM) have been recently developed to measure in the range of several dozen up to hundreds of proteins in one sample with a high sensitivity (Abell et al., 2011; Ahrends et al., 2014; Lange et al., 2008; Peterson et al., 2012; Picotti et al., 2008). These methods are developed for quantification. Adding internal standards (i.e. heavy isotope labeled peptides with fixed concentrations) into the sample even permits absolute quantification.

A typical targeted proteomics workflow contains several steps that are i) determining the set of proteins being monitored, ii) creating a transition list (that is a list of precursor mass/fragment mass pairs for each protein), iii) measure a sample with the created transition list to determine which transitions occur, iv) determine retention time frames for each transition to increase the sensitivity and finally v) measure

[1] http://www.bioinformatics.org/sms2/iupac.html (February 02, 2017)

[2] http://www.nonlinear.com/progenesis/qi-for-proteomics (February 02, 2017).

[3] http://www.psidev.info/mzidentml (February 02, 2017).

[4] http://compomics.github.io/projects/searchgui.html (February 02, 2017).

[5] http://compomics.github.io/projects/peptide-shaker.html (February 02, 2017).

Download English Version:

https://daneshyari.com/en/article/6451847

Download Persian Version:

https://daneshyari.com/article/6451847

Daneshyari.com