Biochimica et Biophysica Acta xxx (2013) xxx–xxx

Contents lists available at ScienceDirect

Biochimica et Biophysica Acta

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

Review

Introduction to opportunities and pitfalls in functional mass spectrometry based proteomics ☆

Marc Vaudel [a,b,*], Albert Sickmann [a,c], Lennart Martens [d,e]

[a] Leibniz-Institut für Analytische Wissenschaften — ISAS — e.V., Dortmund, Germany
[b] Proteomics Unit (PROBE), Department of Biomedicine, University of Bergen, Bergen, Norway
[c] Medizinisches Proteom-Center (MPC), Ruhr-Universität, Bochum, Germany
[d] Department of Medical Protein Research, VIB, Ghent, Belgium
[e] Department of Biochemistry, Ghent University, Ghent, Belgium

ARTICLE INFO

Article history:
Received 17 December 2012
Received in revised form 5 June 2013
Accepted 25 June 2013
Available online xxxx

Keywords:
Proteomics
Data interpretation
Online resource
Pathway
Protein function
Quantification

ABSTRACT

With the advent of mass spectrometry based proteomics, the identification of thousands of proteins has become commonplace in biology nowadays. Increasingly, efforts have also been invested toward the detection and localization of posttranslational modifications. It is furthermore common practice to quantify the identified entities, a task supported by a panel of different methods. Finally, the results can also be enriched with functional knowledge gained on the proteins, detecting for instance differentially expressed gene ontology terms or biological pathways.

In this study, we review the resources, methods and tools available for the researcher to achieve such a quantitative functional analysis. These include statistics for the post-processing of identification and quantification results, online resources and public repositories. With a focus on free but user-friendly software, preferably also open-source, we provide a list of tools designed to help the researcher manage the vast amount of data generated. We also indicate where such applications currently remain lacking. Moreover, we stress the eventual pitfalls of every step of such studies. This article is part of a Special Issue entitled: Computational Proteomics in the Post-Identification Era.

1. Introduction

The progress in the application of mass spectrometry to biological compounds has revolutionized the field of biology: the large scale identification of proteins provides a unique snapshot of a biological system of interest at a given time point [1,2]. The study of proteins and their modifications in a single sample, or differentially between samples dramatically increased our understanding of living cells and allowed the setup of ambitious experiments [3–5] opening new opportunities for biomedical research [6]. The canonical example of the latter is the comparison of the proteomes of a disease affected population against those of a control population [7]. Such studies aim to identify biomarkers – an easily detectable indicator of a biological state – for the targeted disease [8]. However, the efficiency of statistical comparison between metrics associated to a biological entity is questioned in the literature [9–12]. Indeed, such studies suffer from the high variance inherently found in biological systems [9], from the low number of replicates typically analyzed [10], and from experimental artifacts, errors and missing values [13–15]. As a result, the fine nuances of the proteomic variations are often not statistically significant when compared to the global variance of the system.

In order to tackle these issues, the proteomics community has started an ambitious systematic sharing of resources [16]. The rationale is the following: when bringing knowledge from previous experiments and other fields like genomics and transcriptomics together, one will have a better understanding of the results and might be able to extract patterns of interest from the crowd [17]. As a result, the community saw the emergence of quantitative biological pathway or protein interaction analyses. Such systemic approaches aim at providing a fine grained picture of the biological features of interest, hoping at identifying pathology specific disturbances undetectable otherwise.

This process can typically be subdivided into four main tasks: (1) the identification of the biological entities, (2) their absolute or relative quantification, (3) the functional analysis of these entities, and (4) the public dissemination of the results in standardized formats. Starting from the identified and quantified peptides and proteins results canonically obtained from a shotgun proteomics experiment [18], we thus

Abbreviations: PTM, posttranslational modification; FDR, false discovery rate; IEF, isoelectric focusing; OMSSA, Open Mass Spectrometry Search Algorithm; PEP, Posterior Error Probability; FNR, False Negative Rate; PRIDE, PRoteomics IDEntifications database; PICR, Protein Identifier Cross-Referencing; GO, Gene Ontology

☆ This article is part of a Special Issue entitled: Computational Proteomics in the Post-Identification Era.

* Corresponding author at: PROBE, Building for Basic Biology, University of Bergen, Jonas Liesvei 91, N-5009 Bergen, Norway. Tel.: +47 55 58 63 68; fax: +47 55 58 63 60.
E-mail address: marc.vaudel@biomed.uib.no (M. Vaudel).

1570-9639/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.bbapap.2013.06.019

detail in the present review the current follow-up resources available in proteomics. Focusing on free and user-friendly tools, we list the applications – when existing – allowing researchers to reach these objectives. We also list the potential pitfalls involved in these post-identification steps.

2. Global PTM, peptides and protein identification

The typical outcome of a proteomic identification process is a list of identified peptides and proteins with posttranslational modifications (PTMs) mapped onto the sequence. However, these identifications typically contain a certain proportion of false positives [19]. Tremendous progress has been achieved in the monitoring of error rates in proteomics, notably with the use of target-decoy databases [20] — as comprehensively reviewed by Nesvizhskii AI [21]. It has therefore become possible to filter a dataset of interest at a desired False Discovery Rate [22] (FDR) independently from the specific scoring used by the search engine — a demarche common to other scientific fields [23]. For example, in a previous study [24], where three isoelectric focusing (IEF) fractions of the same sample were analyzed using OMSSA [25], a canonical FDR threshold set at 1% required the filtering of all hits with an e-value higher than 0.15, 0.20 and 0.19 for fractions 3, 9 and 20, respectively.

When comparing the target and decoy distribution of hits at these scores, it is possible to estimate an unbiased quality metric, the Posterior Error Probability (PEP) [26]. Concretely, among hundred hits with a PEP of 25%, one expects 25 false positives, 75 true positives. The complement of the PEP (100%–25% = 75%) hence indicates a confidence in the identification. Note that the threshold scores used for the IEF fractions example, although very similar, correspond to a confidence of 72%, 96% and 84%, respectively. This variability in confidence between fractions at fixed FDR shows the heterogeneity found in proteomic results and highlights the necessity for thorough statistical post-processing of the identification matches.

As schematized in Fig. 1A, proteomic experiments typically consist of several samples that are measured in replicates (technical and biological) and that may each be further fractionized. Peptides are inferred from the obtained mass spectra, and proteins are then in turn inferred from the peptides. In the example of the IEF fractions above, we illustrate the importance of processing sets of spectra specifically: the PEP at a given OMSSA score differs from one fraction to the other. Using the OMSSA score for the merged PSM set hence results in a lower identification rate (6084 PSMs at 1% FDR, in orange Fig. 1A) compared to the same set scored using a fraction specific PEP (in black Fig. 1A, 6247 PSMs at 1% FDR: +2.6%). Such specific processing, comparable to charge and modification specific scoring [27], is also mandatory when different mass spectrometers or experimental workflows are used on the same sample. A critical point is then to ensure a statistically relevant size of the subgroups of PSMs retained for scoring [28]. Statistical processing hence makes it possible to filter identification matches at a given quality threshold with a high accuracy [29] and merge the results a posteriori.

However, as illustrated in Fig. 1B with the concatenation of three hypothetical datasets, merging results obtained on different replicates substantially increases the share of false positives since false positive identifications are more likely to differ between replicates than true identifications — a problem well known to affect searches using multiple search engines [30] and peptide and protein inference approaches [31]. In this simple example, where every dataset was filtered to 1% FDR, 25% of the correct matches were unique to a particular dataset, while all false positives were unique. The final FDR therefore reached 1.7% across the datasets: it is hence vital to monitor the quality level of the final result set. Crucially, as illustrated in Fig. 1C, a peptide or a protein can score moderately (in orange) in each of the replicates (like protein D), preventing it from being validated at a quality driven FDR within that replicate. However, its presence among all replicates may make it more confident than another protein scoring well in only one replicate (like protein C). Keeping all identifications from all datasets when creating the merged results and subsequently filtering the combined set thus allows rescuing such peptides and proteins, reducing the False Negative Rate (FNR).

Finally, as illustrated in Fig. 1D, when a peptide is shared between different proteins (e.g., peptide 2 that is shared between proteins A and B), it is not always possible to resolve the correct protein identification; this is the well-described but often underestimated protein inference problem [32], which has particularly strong incidence on the quantification of proteins. In the illustrative example Fig. 1D, the uniquely matched peptide 1 scores well in the first replicate and gives evidence for the presence of protein A. In replicate 2 however, this peptide now receives a poor score and would not pass a stringent quality threshold, thus impairing the protein inference within this replicate. This kind of situation typically occurs when proteins are identified using different fractionation methods or different mass spectrometers. It is hence crucial to consider all peptide candidates across all replicates for protein inference as well.

In summary, an ideal post-identification workflow for proteomics treats identification results accounting for specificities of fractions and replicates (technical and biological) while taking advantage of the study design in its whole to reduce the FNR at a controlled FDR. However, such a statistical analysis is complicated by the fact that identification results between fractions and replicates are not independent. Moreover, there is no guarantee that the target/decoy strategy holds when merging replicates. On the contrary, it can lead to similar issues as multi-stage search strategies [33]. Finally, since such experimental designs can easily lead to the generation of several millions of spectra — as evidenced by some of the biggest datasets [34–37] submitted to the PRoteomics IDEntifications (PRIDE) repository [38], global analysis of complex proteomic studies is also challenging in terms of processing time, computational space, and data management. Several free packages exist that allow the processing of large sets of spectra [39–43], these offer different solutions for the final compilation of the results between replicates, most notably, the MaxQuant/Perseus [39] tool combination allows combination of large datasets, statistical analysis and interaction with external resources.

Fig. 1. Taking advantage of the experimental design. (A) A typical proteomics experiment consists of several samples analyzed in replicates. Here, we take the simple example of three measurements of Isoelectric Focusing fractions from which we want to infer peptide and protein identifications. For every fraction, we represent the target/decoy derived Posterior Error Probability (PEP) at a given score. For the merged result set of Peptide-Spectrum Matches (PSMs), the number of PSMs is plotted at a given False Discovery Rate (FDR) when sorted against the OMSSA score (orange) and against the inferred PEP (black). (B) Processing all samples separately and merging the results increases the FDR substantially: considering an example where 25% of the proteins identified in a sample are unique to that sample, and this includes all false positives (numbers in red). When merging these three datasets (that are each filtered at 1% FDR), a final dataset is obtained with an FDR equal to 1.7%. (C) When considering six proteins identified in the three datasets at different confidence levels (indicated by red, orange and green for bad, medium and good confidence, respectively), it can for instance be seen that protein D is found in all samples yet is not validated due to its moderate score in each sample. The fact that it is found in all samples however, makes it quite likely that it should in fact be included in the global set of identifications. Indeed, although a false negative in all datasets, this protein could be rescued by scoring the identifications globally. Similarly, protein B is not validated in sample 3 but its presence in the global identification suggests that the peptides found in sample 3 should be used for quantification. (D) In proteomics it is sometimes impossible to infer the presence of a protein due to the absence of an identified unique peptide as illustrated here for replicate 1. While protein sequences A and B can be distinguished by peptide 1, this peptide does not receive a high enough score (orange) for identification, and will therefore not be used for protein inference. In replicate 2 however, peptide 1 receives a higher score (green) allowing the unambiguous identification of protein A. Yet if protein A is confidently identified in the second replicate, it is likely to have been in the first replicate as well. A suitable study design can thus help resolve protein inference by analyzing the data globally.

Please cite this article as: M. Vaudel, et al., Introduction to opportunities and pitfalls in functional mass spectrometry based proteomics, Biochim. Biophys. Acta (2013), http://dx.doi.org/10.1016/j.bbapap.2013.06.019