Biochimica et Biophysica Acta xxx (2013) xxx–xxx

Contents lists available at SciVerse ScienceDirect

Biochimica et Biophysica Acta

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

Data processing methods and quality control strategies for label-free LC–MS protein quantification ☆

Marianne Sandin, Johan Teleman, Johan Malmström, Fredrik Levander *

Department of Immunotechnology, Lund University, BMC D13, 22184 Lund, Sweden

ARTICLE INFO

Article history:
Received 14 September 2012
Received in revised form 18 January 2013
Accepted 8 March 2013
Available online xxxx

Keywords:
Protein quantification
Quantitative proteomics
Quality control
Label free
LC–MS
Selected reaction monitoring

ABSTRACT

Protein quantification using different LC–MS techniques is becoming a standard practice. However, with a multitude of experimental setups to choose from, as well as a wide array of software solutions for subsequent data processing, it is non-trivial to select the most appropriate workflow for a given biological question. In this review, we highlight different issues that need to be addressed by software for quantitative LC–MS experiments and describe different approaches that are available. With focus on label-free quantification, examples are discussed both for LC–MS/MS and LC-SRM data processing. We further elaborate on current quality control methodology for performing accurate protein quantification experiments. This article is part of a Special Issue entitled: Computational Proteomics in the Post-Identification Era.

1. Introduction

Liquid Chromatography–Mass Spectrometry (LC–MS) is emerging as one of the main methods for quantification of proteins and consequently the technique is currently a cornerstone of proteomics. Modern LC–MS setups identify thousands of proteins in a single injection of a complex sample [1–4] and enable measurement of low-abundance proteins previously only accessible by expensive and time-consuming antibody-based methods. However, selecting the right MS setup, along with the right informatics solution for analysis of the generated data, remains complex.

Several experimental setups exist for protein quantification using LC–MS. They can in principle be divided into LC–MS/MS, where all peptide ion signals are measured on MS level with MS/MS performed on an either undirected (shotgun) or directed (inclusion list) basis and targeted MS using LC-Selected Reaction Monitoring (SRM) [5]. These can in addition be combined with several conceptually different labeling schemes, including isobaric peptide labels like iTRAQ [6], metabolic labeling e.g. SILAC [7] and stable isotope internal standards like AQUA [8], reviewed in [9]. Labeling techniques facilitate data analysis by minimizing technical bias [10], while label-free methods have the advantage of experimental simplicity and applicability to all numbers and types of samples. In this review we will focus on label-free methods, but most of the considerations will also apply to workflows employing labels.

In parallel with the development of new experimental strategies, a continuous stream of software has emerged to tackle the task of deconvoluting the Gb/h streams of data generated by the instruments. The availability of software has been assessed in several recent reviews [11–14] and it is clear that significant progress has been made in the field. Still, for newcomers and experts alike, it is far from trivial to choose which software solution to use and even more difficult to know if one has performed the data processing in an optimal way. There is thus need for methods to evaluate the results of a selected workflow [15]. In the present review, we will delineate the specific steps of LC–MS/MS as well as LC-SRM data processing and highlight the potential error sources of each step before elaborating on generic issues like normalization and protein inference. Furthermore, we give an overview of quality control methodology for the label-free workflow, highlighting the need for standardized methods and datasets when comparing software setups as well as parameter settings.

2. Label-free LC–MS workflows

The typical quantitative LC–MS workflow starts with extraction and optional separation of proteins, followed by digestion to peptides using a specific endoprotease like trypsin. The peptide mix is then analyzed using one of the available LC–MS techniques. In general, shotgun LC–MS/MS is used for discovery experiments, while more targeted approaches are used in validation experiments, as reviewed in [5]. As the present review is focused on data processing we will not

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

* Corresponding author. Tel.: +46 462223835; fax: +46 462224200.
E-mail address: fredrik.levander@immun.lth.se (F. Levander).

1570-9639/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.bbapap.2013.03.026

further discuss the choice of experimental setup here, but only discuss the different workflows from a data handling perspective, as outlined in Fig. 1.

2.1. LC–MS/MS

The most frequent way of acquiring LC–MS/MS data is using data-dependent acquisition (DDA), where the most abundant peptide peaks at each time point are selected for MS/MS. For a comprehensive overview, see [5]. In order to avoid repeated sampling of the same peak, they are temporarily put on an exclusion list after selection. The resulting data will be a mixture of MS and MS/MS spectra and both types of spectra can be used for quantitative purposes. Spectral counting (SC) uses only MS/MS spectra and protein abundances are estimated based on the number of peptide spectra that are identified for a particular protein. SC has been shown to give adequate estimates of protein quantities [16–18], especially with new algorithmic developments for absolute quantification [19] as well as combination of different SC metrics [20]. However, SC is limited by the MS/MS sampling frequency of the instrument and the quantitative resolution is low for low-abundance proteins where only single peptide identifications can be found. In contrast, precursor-based quantification can potentially quantify a hundred thousand peptides in a single run using a standard high-resolution setup [21]. This number is further increasing as instruments acquire data at higher resolution [22], and both relative and absolute quantification can be performed using the method (examples of the latter are [18,23,24]). We will therefore focus on precursor-based data processing, which is also the most computationally challenging. Precursor-based quantification using LC–MS/MS data involves several computational steps (Fig. 1). Different software algorithms can perform each of these steps and we will discuss potential pitfalls in the following sections. For the specific LC–MS/MS data processing steps, the focus will be on feature detection (extracting potential peptides and their quantities from the data) and alignment (the correction of elution time drifts between LC–MS/MS runs to facilitate differential expression analysis). Several computational platforms that allow for complete data processing by integration of algorithms for these different computational steps exist and an overview of such platforms is given in Table 1.

2.2. LC-SRM

Targeted proteomics using SRM has become a powerful tool due to its ability to repeatedly measure selected peptides in a highly specific and accurate manner. SRM in the standard setup uses a triple quadrupole mass spectrometer to select peptide ions of interest, fragments them and then measures specific fragment ions. To enable this however, one needs knowledge of what peptides to target, what collision energy to use and which fragments to measure. Peptides need to be unique for the protein of interest within the measured proteome and also have properties that make them suitable for MS analysis. Peptide uniqueness is straightforward to predict by in silico digestion of the target genome, but although some efforts have been made to predict peptide suitability for MS (e.g. [39,40]), the typical setup involves selection of peptides that have previously been observed in LC–MS/MS experiments. The generation of SRM assays is thus often closely coupled to MS/MS analysis (Fig 1). SRM post-acquisition data processing can be divided into four steps (Fig. 1) — I) detection of possible peptide peaks (features) in the chromatograms, II) assessment of the quality of these peaks and selection of the best (for example most probable) candidate, III) integration of the peak signal to find a relative intensity and IV) combination of peptide quantities into a protein quantity. We will discuss the data processing methods used by current software for each of these steps and highlight advantages as well as points for improvement.

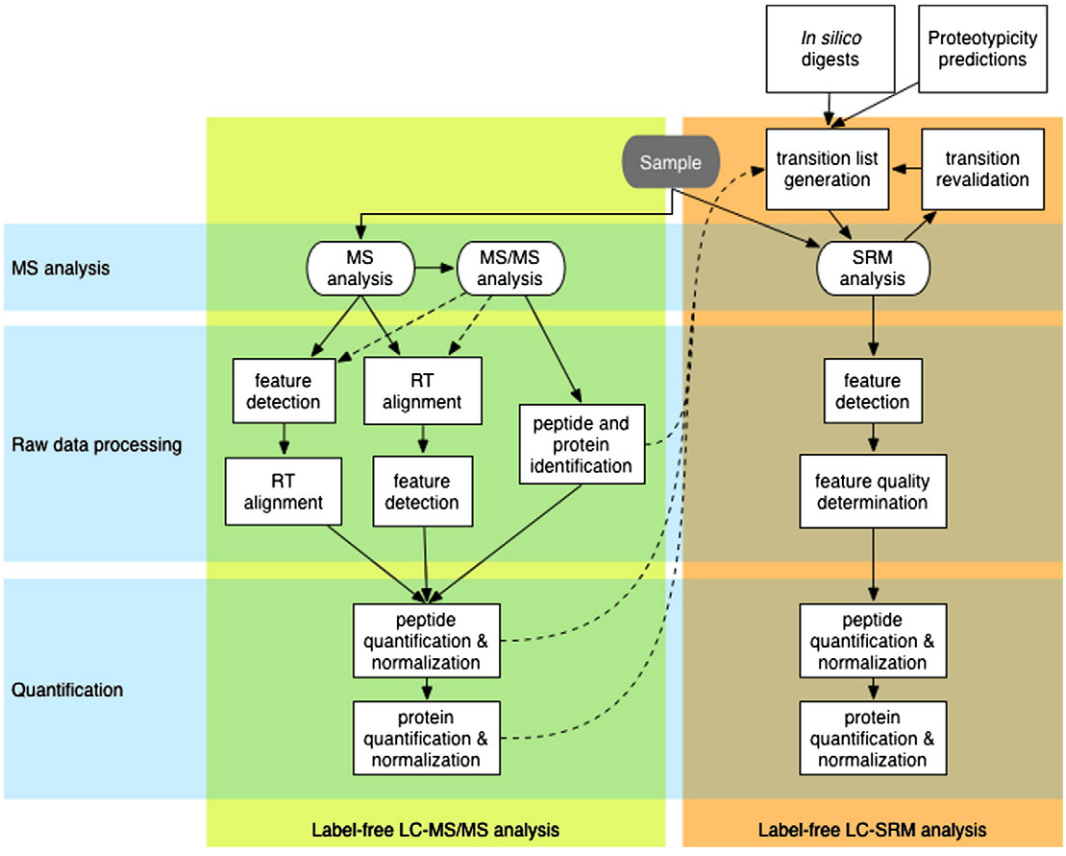

Fig. 1. Information flow during data processing in label-free LC–MS/MS and LC-SRM.

Please cite this article as: M. Sandin, et al., Data processing methods and quality control strategies for label-free LC–MS protein quantification, Biochim. Biophys. Acta (2013), http://dx.doi.org/10.1016/j.bbapap.2013.03.026