Biochimica et Biophysica Acta 1864 (2016) 883–895

Contents lists available at ScienceDirect

Biochimica et Biophysica Acta

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

# Thousand and one ways to quantify and compare protein abundances in label-free bottom-up proteomics☆

Mélisande Blein-Nicolas, Michel Zivy *

*GQE-Le Moulon, INRA, Univ Paris-Sud, CNRS, AgroParisTech, Université Paris-Saclay, F-91190 Gif-sur-Yvette, France*

ARTICLE INFO

*Article history:*
Received 5 November 2015
Received in revised form 21 January 2016
Accepted 24 February 2016
Available online 3 March 2016

*Keywords:*
Mass spectrometry
Data processing
Statistics
Peptide

ABSTRACT

How to process and analyze MS data to quantify and statistically compare protein abundances in bottom-up proteomics has been an open debate for nearly fifteen years. Two main approaches are generally used: the first is based on spectral data generated during the process of identification (*e.g.* peptide counting, spectral counting), while the second makes use of extracted ion currents to quantify chromatographic peaks and infer protein abundances based on peptide quantification. These two approaches actually refer to multiple methods which have been developed during the last decade, but were submitted to deep evaluations only recently. In this paper, we compiled these different methods as exhaustively as possible. We also summarized the way they address the different problems raised by bottom-up protein quantification such as normalization, the presence of shared peptides, unequal peptide measurability and missing data. This article is part of a Special Issue entitled: Plant Proteomics– a bridge between fundamental processes and crop production, edited by Dr. Hans-Peter Mock.



## 1. Introduction

Liquid chromatography coupled to tandem mass spectrometry (LC–MS/MS) has emerged as a core tool for proteomics, particularly for the identification and characterization of large and complex sets of proteins extracted from biological samples [1]. In this technique, protein-derived analytes are separated by LC and ionized by electrospray before entering the mass spectrometer where they can be submitted to two stages of mass analysis (referred as MS1 and MS2 scans) separated by a stage of selection and a stage of fragmentation in a collision cell. Three strategies for LC–MS/MS-based proteomics have been developed. In the bottom-up and middle-down approaches, analytes are peptides resulting from a complete or limited proteolytic digestion, while in the top-down approach, intact proteins are analyzed.

Bottom-up proteomics is currently at the basis for much of the protein research undertaken in biology (reviewed in [2]). One of its main applications is protein quantification that is said absolute when it aims at estimating intracellular protein concentrations, and relative when it aims at comparing protein abundances between different samples. The main difficulty of bottom-up proteomics is to reconstitute information about the proteins from peptide mixtures that remain generally complex, even after protein and/or peptide fractionation and LC separation. Therefore, two main questions arise when designing an experiment in bottom-up proteomics: (*i*) How to extract relevant MS data from these peptide mixtures? (*ii*) How to make these MS data talk? These two questions relate to the acquisition and processing of MS data, both of which are under constant development.

MS data acquisition is controlled through a set of parameters, such as cycle time or resolution of MS1 and MS2 scans, which altogether define a method of acquisition. Over the past 15 years, the panel of acquisition methods available to scientists for analyzing protein samples by bottom-up proteomics has been greatly enriched. These methods can be grouped in three main approaches. In data-dependent acquisition, the precursor ions isolated for fragmentation are selected depending on their signal intensity. In inclusion list-driven acquisition, only the precursor ions representing peptides used as surrogates for proteins of biological interest are selected and fragmented. This approach mainly refers to two methods of targeted proteomics called selected reaction monitoring (also known as multiple reaction monitoring) and parallel reaction monitoring (reviewed in [3]). Lastly, in data independent acquisition, including methods such as SWATH and $MS_E$, all precursor ions within a given range of mass-to-charge ratios are fragmented without selection (reviewed in [4]). These developments in data acquisition have been driven by the rapid advances made in LC–MS/MS instrumentation, particularly with regard to performances of mass spectrometers in terms of resolution, mass accuracy, scanning speed and sensitivity [5] and by the development of new configurations of analyzers in MS/MS (reviewed in [6]). In a complementary way, many advances have been made to process MS data.

During a LC–MS/MS run, three types of data are acquired: the retention time, the mass-to-charge ratios and the intensities of all ions scanned at a given chromatographic time (precursor ions in MS1 scans and

*Abbreviations:* FDR, false discovery rate; LC–MS/MS, liquid chromatography coupled to tandem mass spectrometry; SC, spectral count; XIC, extracted ion current.

☆ This article is part of a Special Issue entitled: Plant Proteomics — a bridge between fundamental processes and crop production, edited by Dr. Hans-Peter Mock.

* Corresponding author.
*E-mail address:* zivy@moulon.inra.fr (M. Zivy).

http://dx.doi.org/10.1016/j.bbapap.2016.02.019
1570-9639/

fragment ions in MS2 scans). Peptide identification is performed by inferring amino-acid sequences from the fragmentation patterns of precursor ions on MS2 spectra (intensity *versus* mass-to-charge ratio at a given retention time) either by sequence database searching, spectral library searching or *de novo* sequencing [7]. This is done automatically by a number of identification softwares such as X!Tandem [8] or Mascot [9]. Protein identification is subsequently achieved by assigning peptide sequences to proteins (reviewed in [10]).

In label-free bottom-up proteomics, quantification can be performed from two types of data: identification results and ion intensities. In the early 2000s, Washburn et al. [11] indeed observed that the number of identified peptides per protein increased with increasing codon adaptation index (which is a measure of codon usage bias) and thus with protein abundance since the codon adaptation index is considered a predictor of protein abundance [12]. At the same time, it was shown that signal intensity from electrospray ionization was correlated with ion concentration [13] and that the chromatographic peak area or peak height was linearly correlated to the protein concentration [14]. Since then, several tools such as MassChroQ [15], MaxQuant [16], Skyline [17] or Progenesis (Nonlinear Dynamics, Newcastle, UK) have been developed to detect, quantify and match chromatographic peaks in MS1 or MS2 extracted ion current (XIC, intensity *versus* retention time for a given mass-to-charge ratio).

High level LC–MS/MS data processing includes inferring and comparing protein abundances from information collected at the peptide level. This is a daunting task, regardless of the type of data considered. First, because the structure of the data is highly complex, due to a high number of missing data arising from different mechanisms: random (due, for instance, to ionization efficiency or ion-suppression effects), intensity-dependent (low abundance peptides are more likely to present missing values), performance-dependent (variations in the instrument performances may affect the total number of peptides quantified in the samples) or database-dependent (a peptide may be missed in a given genotype if the genetic polymorphism is not taken into account in the searched database). Second, because of the presence of peptides shared by different proteins. These peptides are necessarily more abundant than the proteotypic peptides belonging to the same proteins and their variations between samples depend on the possible variations of all the proteins that share them. Although they constitute a valuable source of information [18], shared peptides are generally discarded because of the difficulty to properly deconvolve the information they carry. Third, because peptides in equal amounts may not provide the same intensity value. This is due to several factors, such as digestion efficacy, peptide hydrophobicity and ionization potential, which affect the peptide measurability [19]. A wide variety of methods have been proposed to infer relative protein abundances from data obtained by label-free bottom-up proteomics and to detect abundance changes between biological samples of interest (reviewed in [20–22]). Several methods for absolute protein quantification using label-free bottom-up proteomics have also been proposed. Briefly, these methods consist in dividing the relative abundance of a given protein by its number of theoretically observable peptides, so that the resulting index theoretically allows comparisons between different proteins within a sample. Absolute quantification in the sense of obtaining the concentration of a protein in a sample can be obtained by label free methods only by comparing the measurement observed in the sample to measurements performed with the same method on a concentration range of known concentrations of the same protein. In any case, no gold standard method has emerged so far, and how to best infer and statistically compare protein abundances remains an open question which is regularly addressed in the literature.

In this paper, we provide an up-to-date compilation of the different methods used to relatively quantify and statistically compare protein abundances in label-free bottom-up proteomics. These methods are summarized in Table 1. We also provide a review of their relative performances and summarize the way they address the different problems raised by bottom-up protein quantification such as normalization, presence of shared peptides, unequal peptide measurability and missing data.

## 2. Identification-based methods

### 2.1. Count-based methods

#### 2.1.1. Peptide counting

The first identification-based abundance feature, called Protein Abundance Index (PAI), was proposed by Rappsilber et al. [23] as the ratio between the number of peptides-charge identified for a given protein (or peptide count) and the number of theoretical tryptic peptides of this protein that fall within a given mass range of the mass spectrometer. The Protein Abundance Index, as originally defined, was not considered as an accurate measure of protein amount, but rather as a guide for relative classification in abundant and less abundant proteins [23]. Indeed, with its denominator, the Protein Abundance Index takes into account that large proteins generate more detectable peptides than small proteins [11], therefore allowing comparisons between different proteins. An exponentially modified version of the Protein Abundance Index, called emPAI, was later reported by Ishihama et al. [24], that is better correlated to protein concentration. The Protein Abundance Index and/or the exponentially modified Protein Abundance Index are implemented directly in certain search engines like Mascot [9] as well as in various softwares for post-processing identification results like emPAI Calc [25], Scaffold [26] or Crux [27]. Similar to the Protein Abundance Index, Sun et al. [28] proposed the modified Spectral Count Index (mSCI), where number of observed peptides per protein is divided by the protein relative identification possibility (RIPpro). The protein relative identification possibility measures the potential of a specific technique's ability to identify protein expression in a large-scale study. Its calculation is based on the molecular weight, the isoelectric point and the hydrophobicity of the proteins.

#### 2.1.2. Spectral counting

Spectral count (SC) refers to the number of MS2 spectra assigned to one protein. It includes all redundancy of peptide identification such as modifications, charge state, missed cleavages and multiple detection of the same peptide due to expired dynamic exclusion. SC was introduced as a protein abundance feature by Liu et al. [29] from the work of Pang et al. [30] and Gao et al. [31] showing that SC is correlated to protein abundance. Many LC–MS/MS instrument parameters can have a strong impact on SC, the most important being the dynamic exclusion duration, which can significantly affect the detection number of peptides and spectra for lower abundance proteins [32]. To correct for fluctuations in total SC between samples, normalization can be performed by dividing the SC of a protein by the total SC of the sample.

The SC was shown to be more correlated to relative protein abundance than the peptide count used in the Protein Abundance Index [29]. However, like the peptide count, the SC is strongly correlated to the protein length or molecular weight. In addition, proteins with high SC also have higher statistical significance of protein abundance change [33]. To take this into account, some researchers proposed to divide the SC by the protein length (Normalized Spectral Abundance Factor, NSAF [34]) or the molecular mass (Protein Abundance Factor, PAF [35,36]). Others applied machine learning techniques accounting for protein size, sequence properties, ionizability and other properties influencing MS detectability to estimate the number of unique tryptic peptides expected for a protein [37,38]. The value obtained was used together with the probability of correctly identifying the protein to compute a SC-based abundance feature named Absolute Protein EXpression (APEX). The Absolute Protein EXpression is implemented in The APEX Quantitative Proteomics Tool, a free open source Java application [39] as well as in aLFQ [40].

In the methods cited above, the presence of peptides shared by several proteins in SC computation is not clearly addressed. Generally, the