Journal of Proteomics xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

Journal of Proteomics

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

The benefits (and misfortunes) of SDS in top-down proteomics

Carolyn Kachuk, Alan A. Doucette *

Department of Chemistry, Dalhousie University, Halifax, Nova Scotia, Canada

ARTICLE INFO

Article history:
Received 8 November 2016
Received in revised form 19 January 2017
Accepted 3 March 2017
Available online xxxx

Keywords:
Top-down proteomics
Sodium dodecyl sulfate
SDS depletion
Sample preparation
SDS-protein binding
ESI suppression

ABSTRACT

Top-down proteomics (TDP) has great potential for high throughput proteoform characterization. With significant advances in mass spectrometry (MS) instrumentation permitting tandem MS of large intact proteins, a limitation to the widespread adoption of TDP still resides on front-end sample preparation protocols (*e.g.* fractionation, purification) that are amenable to MS analysis of intact proteins. Chromatographic strategies are improving but pose higher risk of sample loss. Gel-based separations (*e.g.* GELFrEE) may alleviate this concern but at the expense of requiring sodium dodecyl sulfate (SDS). While this surfactant maintains protein solubility during fractionation, the advantage is short-lived, as the detergent must ultimately be depleted to avoid MS signal suppression. To do so requires overcoming strong interactions between SDS and protein. Adding to the challenge, one must now consider upholding the solubility of purified protein(s) in the absence of SDS. This review explores uses of SDS in TDP workflows, addressing front-end strategies that reduce matrix interferences while maximizing recovery of intact proteins in MS-compatible formats.

Significance: The benefits of employing SDS in a TPD workflow can easily outweigh the disadvantages. Several SDS depletion strategies are available, though not all are equally amenable to TDP. This review provides a comprehensive and critical accounting of SDS in TDP, demonstrating methods that are suited to MS analysis of intact proteins.

1. Introduction

For decades, the terms “bottom-up” and “top-down” have been applied across multiple disciplines (economics, sociology, biology, *etc.*) to describe complementary strategies of information gathering. It was not until the late 1990′s that the terms appeared in the literature in relation to proteomic workflows. Kelleher et al. first used “top-down” in 1998 to describe their process of identifying a protein through accurate mass measurement coupled with tandem mass spectrometry (MS/MS) of the intact protein [1]. The term “bottom-up” followed one year later by the same group when describing the complementary, established workflow incorporating peptide-level MS analysis [2]. Though top-down proteomics (TDP) is perhaps seen as evolving from the bottom-up approach, it has long been a goal to directly sequence larger proteins by MS/MS. The development of soft ionization techniques (ESI [3] and MALDI [4]) removed a limitation of large molecule ionization, permitting direct analysis of intact proteins with collisionally activated dissociation (CAD) to fragment the molecule. In 1990, using a conventional triple quadrupole instrument, Loo, Edmonds and Smith were able to sequence ribonuclease A as the intact protein (MW ~ 14 kDa) [5], and quickly extended this work to much larger albumin proteins (~66 kDa) only a year later [6]. The challenges of large molecule dissociation by CAD, together with interpreting the complex fragmentation spectrum generated from a low resolution platform were acknowledged in their report. It was no surprise that in 1993, when Henzel et al. [7] and other groups [8–11] independently proposed a simple approach to identify proteins through MS but without MS/MS (*i.e.* peptide mass fingerprinting, PMF), the technique rapidly gained in popularity. The limitations of PMF to characterize protein mixtures are evident and so shortly thereafter, computational algorithms became available to automatically interpret tandem spectral data, including the MOWSE algorithm (*a.k.a* MASCOT) by Pappin et al. [12], and SEQUEST by Eng et al. [13]. What truly solidified the shotgun, bottom-up approach were developments in front-end separations, allowing complex proteome mixtures (in the form of digested peptides) to be automatically fed to the mass spectrometer. The coupling of two dimensional chromatography with MS (*e.g.* Washburn et al.'s MUDPIT [14]) or pairing reversed phase chromatography with isoelectric focusing [15] are two examples which exemplify high throughput analysis of complex peptide mixtures.

Of course, the desire to employ a similar strategy with intact proteins continued to advance TDP. MS instrumentation, once acting as a major limitation to the size of the protein that could be characterized [16], were seeing improvements in resolution, sensitivity, scan speed, ionization, and fragmentation [17–19]. These improvements lent the characterization of proteins as large as 200 kDa [20]. One advantage of TDP is

* Corresponding author at: PO Box 15000, 6274 Coburg Road, Halifax, Nova Scotia B3H 4R2, Canada.
E-mail address: alan.doucette@dal.ca (A.A. Doucette).

http://dx.doi.org/10.1016/j.jprot.2017.03.002
1874-3919/

that it theoretically provides 100% sequence coverage [21] allowing characterization of post-translational modifications (PTM) [22], and other protein variations [23,24]. While bottom-up typically employs CAD or the related HCD (high-energy collision dissociation) to fragment ions, cleaving bonds at the lowest activation energy – often the PTMs bond [25], TDP often favors non-ergodic techniques such as electron capture dissociation (ECD) or electron transfer dissociation (ETD). These modes of fragmentation cleave along the protein backbone and preserve PTM's [25,26]. Shaw et al. recently introduced a novel approach to fragment proteins *via* ultraviolet photodissociation [27]. By preserving labile PTMs, this allows direct quantification of proteoforms [28], mapping of modifications with full sequence coverage, discovery of unexpected modifications, identification of positional isomers, and the determination of the order of multiple modifications [25,29,30]. Any discrepancy between the exact mass obtained by top-down MS and the predicted MW based on the DNA sequence is attributed to a modification [20]. The type and location of the modification is determined through the fragment ion masses [2,6,21,26,31]. TDP has been employed to identify potential tumor biomarkers from breast cancer xenografts [32] and from salivary glands [33]. Toby et al. provide a comprehensive review of the current status of TDP and its application to proteoform and PTM identifications [23]. As instrument limitations become less obstructive, TDP is seen as an opportune alternative to bottom-up for MS-based protein identification [32,34,35].

A true high-throughput TDP workflow must also make use of front-end proteome fractionation. In 2007, Sharma et al. identified 715 proteoforms by coupling 2D liquid chromatography (LC) separation to MS, in the form of weak anion exchange and reversed phase LC [36]. In 2013, Ansong et al. identified 1665 proteoforms using a single dimension of separation by taking advantage of a shallow (250 min) RPLC gradient [37]. Employing a single dimension of separation, this approach reduces sample loss and thus is highly suited to low quantities of starting material. By incorporating GELFrEE, an electrophoretic platform similar to SDS PAGE where proteins migrate through a polyacrylamide gel 'column' [38,39], and subsequent protein precipitation to deplete SDS, Kelleher's research group identified over 5000 proteoforms *via* TDP in the same year [40]. They also obtained similar identifications in a 2016 follow up study [28], though at a lower false positive rate, and provided quantitative information on the intact proteins identified. Also in 2016, Shortreed et al. identified over 8600 proteoforms belonging to 1178 different protein families (*i.e.* unique genes) [41]. This study did not employ MS/MS, but rather relied on accurate mass measurements of intact proteins together with a count of the lysine residues per protein (obtained *via* SILAC labelling [42]). Durbin et al. have improved the process of MS data acquisition, guiding MS/MS fragmentation to maximize identification of lower abundant proteoforms [43].

With the vast amount of data that is obtained through mass spectrometry of proteins, data processing has become an essential part of any proteomic workflow [44–47]. Online bioinformatics platforms such as DAVID (Database for Annotation, Visualization, and Integrated Discovery) [48,49] provide functional annotation tools, particularly the discovery of biological themes (molecular function, cellular component, or biological processes), using the controlled vocabulary of Gene Ontology (GO) terms [50,51]. Another useful tool in the classification of proteins relates to hydrophobicity. Using Grand Average of Hydropathy (GRAVY) [52], numerical scores are assigned to every amino acid (hydrophilic are <0). The higher the GRAVY score, the more hydrophobic proteins tend to be, which suggests that they are likely associated with the membrane. Membrane protein topology can then be predicted using algorithms such as TMHMM (Tied Mixture Hidden Markov Model) [53] and AmphipaSeek [54], which identify transmembrane helices and in-plane membrane anchors, respectively, within protein sequences. These types of data processing tools are important for detecting remote sequence homologies [44], sub-cellular localization [55], and protein-protein interactions [56]. Additionally, Kelleher's research group have contributed developments in data processing (C-score and ProSight PTM) that improve interpretation of TDP MS data [57–61]. In conjunction with algorithms such as THRASH [62], the high throughput classification of large molecules is facilitated.

Despite these many improvements in MS instrumentation and data processing, it is realized that technological challenges continue to limit the potential of TDP. In particular, front-end manipulations of intact protein mixtures must consider the presentation of purified analyte to the mass spectrometer without biased recovery, and with sufficient resolution to maximize data output. This becomes increasingly important as deep proteome mining, and quantitative TDP are now being realized.

Quantitative approaches measure changes in abundance and facilitates the comparison of samples with changing environments [63]. Paša-Tolić et al. were one of the first to implement quantification using capillary isoelectric focusing (CIEF) FTMS [64] and the field has grown to include labeled (ICAT [65], SILAC [42], iTRAQ [66]) and non-labeled quantification protocols [67]. These methods have been thoroughly reviewed elsewhere [63,68] and are an integral part for the complete mapping of proteomes [69].

1.1. Challenges in TDP

To a first approximation, as the digestion step is avoided one would presume a mixture of intact proteins to be inherently less complex, owing to a reduced number of components in the sample [24]. Unfortunately, the diversity of intact proteins far outstrips any advantage gained in lowering the number of components. There are approximately 22,000 non-redundant protein families (*i.e.* coded by unique genes) in humans [70], but protein complexity dramatically increases due to allelic variations, post-translational modifications (PTMs), alternative splicing events, and degradation [24,71,72]. Estimates of the number of chemically distinct protein forms (termed proteoforms [73]) range from 0.1 million [74] to between 0.6 and 6 million [75] proteins. Proteins are known to be modified by >100 known chemical groups [21], often multiple times on a given protein, creating proteoforms which may be nearly identical in terms of physical or chemical properties (*e.g.* molecular weight, solubility, abundance), but differing significantly in biological function. These modifications further complicate sample preparation strategies for intact proteins [21,24].

In TDP, simply maintaining the solubility of all components during fractionation is a primary concern. Unlike bottom-up, where one or more peptides from a given protein may be lost during separation, with top-down there is no "backup" molecule and so if a given protein suffers poor extraction efficiency from a cell or low recovery from a chromatographic column, the impact on MS will be felt. Concerns of protein recovery are further exaggerated as one considers the inherent loss of MS sensitivity when dealing with the multiple charge series envelope of intact protein ions [21]. By incorporating effective front-end separation techniques to reduce sample complexity ahead of MS, these challenges can be overcome.

2. Protein separation techniques

Intact protein separation by chromatography is inherently more difficult than peptide separation [76] as proteins tend to interact with chromatographic stationary phases in undesirable ways [77–83]. Ionic interactions are common, resulting in adsorption of protein [84], shifts in retention time [85], peak tailing or asymmetry [86], and changes to the 3D structure of the protein [78,87]. Non-binding electrostatic interactions can also occur, resulting in "ion-exclusion" that prevents the proteins from interacting with the pores of the column, thus eluting sooner than predicted [79]. Hydrophobic interactions can also play a role, leading to increased retention or on-column denaturing [88].

Despite these challenges, progress has been made in chromatographic separation. Over the past couple of years, Ying Ge's group has worked on effectively separating intact proteins with HIC, RPLC, and IEC chromatographic approaches [89,90]. They include ammonium

Please cite this article as: C. Kachuk, A.A. Doucette, The benefits (and misfortunes) of SDS in top-down proteomics, J Prot (2017), http://dx.doi.org/10.1016/j.jprot.2017.03.002