Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Chemical Biology

Developing top down proteomics to maximize proteome and sequence coverage from cells and tissues

Dorothy R Ahlf[1], Paul M Thomas[2] and Neil L Kelleher[2,3]

Mass spectrometry based proteomics generally seeks to identify and characterize protein molecules with high accuracy and throughput. Recent speed and quality improvements to the independent steps of integrated platforms have removed many limitations to the robust implementation of top down proteomics (TDP) for proteins below 70 kDa. Improved intact protein separations coupled to high-performance instruments have increased the quality and number of protein and proteoform identifications. To date, TDP applications have shown >1000 protein identifications, expanding to an average of ~3–4 more proteoforms for each protein detected. In the near future, increased fractionation power, new mass spectrometers and improvements in proteoform scoring will combine to accelerate the application and impact of TDP to this century's biomedical problems.

Addresses
[1] Department of Chemistry and Biochemistry and the Harper Cancer Institute, University of Notre Dame, Notre Dame, IN, United States
[2] Department of Chemistry, and the Proteomics Center of Excellence, Northwestern University, Evanston, IL, United States
[3] Department of Molecular Biosciences, and the Proteomics Center of Excellence, Northwestern University, Evanston, IL, United States

Corresponding author: Kelleher, Neil L (n-kelleher@northwestern.edu)

Current Opinion in Chemical Biology 2013, **17**:787–794

This review comes from a themed issue **Analytical techniques**

Edited by **Milos V Novotny** and **Robert T Kennedy**

For a complete overview see the Issue and the Editorial

Available online 27th August 2013

1367-5931/$ – see front matter,

http://dx.doi.org/10.1016/j.cbpa.2013.07.028

Introduction

Proteomics: from inception to enduring goals

The analysis of proteins has undergone a major revolution over the past 20 years from the earliest days of amino acid analysis and Edman sequencing to today's sophisticated mass spectrometry platforms. The successes of the human genome project have inspired similar efforts within the context of the proteome and have thus led the rapid development of high-throughput methods for proteomics [1,2]. Characterizing the chemical state of these proteins provides valuable biological information. The complexity of proteomics, a 'global cellular view', arises when all combinatorial patterns are taken into account across a variety of cell types. To date, bottom-up proteomics has proven ineffective to detect combinatorial proteomics, unless the modifications are co-located on one peptide.

In many regards, the human proteome is more complex than its genome. Each somatic cell in the human body encodes the same genetic information in $\sim 3 \times 10^9$ base-pairs of DNA. However, the human proteome cannot be defined this trivially. The proteoform content of a cell changes with cell type, over time and in response to external stressors. While the human genome contains just over 20 000 protein-expressing genes, RNA processing alone increases the number of possible base sequences to perhaps >100 000 in most cells. Finally, proteins may also be highly modified with differential combinatorial patterns of post-translational modifications (PTMs) [3,4]. Extensive studies of singly, highly modified proteins (e.g. histones) show that though these multitudes of modification combinations are possible, only a limited number modified forms are observed [5–7].

A word on language and protein databases

During the development of mass spectrometry-based proteomics, many new terms have entered the scientific vernacular. One sequence translated from a gene in the Universal Protein Resource, or UniProt, is selected as the 'canonical sequence', and variations to the base amino acid sequence are referred to as isoforms. However, this term fails to capture the complexity of highly post-translationally modified proteins that may also have base sequence changes. As different isoforms may be modified differently from each other, it is important to have language to differentiate the level at which one is speaking, analogous to the levels of protein higher order structure. The term 'proteoform' encapsulates the combinatorial combination of a set of modifications on a particular UniProt isoform (stably identified with a hyphen and then an integer, e.g. *-1 for the canonical*, *-2*, *-3* and so on) [8••]. The proteoform term includes all site specific features such as coding single nucleotide polymorphisms, mutations, or PTMs that map to the same gene. One isoform may have many different possible proteoforms. Note also that the UniProt KnowledgeBase is a gene-centric database, and, if used precisely with database search engines, can provide better clarity on the lingering issue of protein inference for bottom up; top down technology achieves gene-specific identification for proteins and thus has no such inference problem.

Mass spectrometry methods for proteomics: top down and bottom up

From the earliest days of proteomics (even before it was termed as such) two main types of mass spectrometric

analysis were performed. The primary method for protein identification is bottom-up, where peptides, generated from enzymatic proteolysis of proteins, are analyzed in a mass spectrometer [9,10]. To increase dynamic range, many groups have employed polyacrylamide gel electrophoresis (SDS-PAGE), either in one dimension, separating by molecular weight, or in two dimensions with a primary isoelectric focusing component. As excising proteins from a gel is labor intensive, many groups have preferentially turned to on-column separation techniques such as Multidimensional Protein Identification Technology (MudPIT) or other separation strategies [11,12]. Digestion of proteins requires the researcher to infer the identity of a protein from smaller peptides in a robust, relatively easy, and rapid fashion. Further analytical techniques have been based around this method to give quantification and identify modified proteins by class [13]. However, a major limitation of these enrichment protocols is their potential to alter observed stoichiometry. Rarely do the peptides detected provide information covering the entire protein because certain peptides may not be detected (particularly true for low abundance proteins). Finally, as with many scientific methods generating 'big data', researchers continue to optimize the most correct statistical methods of reporting identifications and false discovery rates [14–16].

To complement the speed and sensitivity of bottom-up proteomics, top-down proteomics introduces intact proteins into the mass spectrometer and then fragments whole protein ions directly [17•]. When the complete intact protein is present and measured at high mass accuracy, 100% sequence coverage is obtained and PTM combinations are preserved, leading to precise identification and characterization of specific genes, isoforms and proteoforms. However, due to inherent difficulties in both the separation and detection of intact proteins, there is low proteome coverage per injection compared with peptide-based analyses [18]. Also, the cost of mass spectrometers required to obtain high mass accuracy measurements is prohibitive to many groups. Moving forward, benchtop style instruments will bring this capability to more research groups than in past years [19–21]. With this and further development on high-throughput methods for intact proteins, the barriers to implementation of the top-down approach will drop substantially over the coming years [22,23••]. The full platform recently developed by the Kelleher lab combines all the elements discussed in the following sections to obtain high proteome coverage (Figure 1). For this reason, it will serve as the focus of this perspective, along with selected other methods discussed in the sections below.

A platform for top down proteomics on a high throughput basis

Mass-based fractionation of intact proteins

Once protein samples have been obtained from many different available methods, the next downstream step can be a mass-based separation. This approach allows the researcher to sequester proteins into similar ranges of molecular weight and apply a few adjustments to downstream analytical methods for low (>30 kDa), medium (30–70 kDa), and high (>70 kDa) mass proteins [24]. Many previous researchers had attempted to use mass-based separation for intact proteins, with limited success [3,25]. A special gel band elution device can be used, but few papers exist due to its low recovery of intact proteins [4].

Tube gel electrophoresis overview and theory

Tube gel electrophoresis operates upon the same separation principles of SDS-PAGE gel electrophoresis; however, in the Gel Elution Liquid-based Fractionation Entrapment Electrophoresis (GELFrEE) device and other similar devices, proteins elute through the gel and into solution (Figure 2). Tube gel separation, therefore, gives higher sample recovery and is amenable to other separations either before or afterwards. Depending on the cross-sectional area of the separation tube, much greater sample amounts can be separated than in a single lane of a SDS-PAGE slab gel. Similar to gel electrophoresis, the separation can be optimized for an expected mass range by changing the degree of gel crosslinking. Each time-based fraction harvested correlates to a specific expected mass range which one may optimize with standard proteins and lysates for reproducible results [26–29]. Some highly hydrophobic proteins can be maintained in solution with surfactants present (even integral membrane proteins with up to ~8 transmembrane domains). GELFrEE allows the researcher to obtain protein fractions in a time-based manner, although the sample harvesting is currently manual [28•,30]. Since the publication of the initial paper in Analytical Chemistry, this technology has been commercialized as the GELFREE 8100 Fractionation System. Each particular sample may present unique challenges; yet the GELFrEE device allows many parameters to be optimized such as stacking gel length, loading amount, and collection time. Many different types of protein sample have been coupled to this separation platform due to the ease of use and its similarity to SDS-PAGE [26,31,32].

Reversed-phase liquid chromatography (RPLC) and online separations

Liquid chromatography (LC) is among the most popular method of separation for peptides and intact proteins. Reverse phase liquid chromatography, RPLC, in particular is among the most common separation before mass spectrometry. This technique separates proteins based on hydrophobicity, with the most hydrophilic molecules eluting first. In large part due to the popularity of this technique, a wide range of materials are available and numbers are continuing to grow. In addition, even though challenges still exist for nanocapillary-based RPLC of whole proteins, many research groups are using this for