Sequencing covalent modifications of membrane proteins

Julian P. Whitelegge1,*, Arthur Laganowsky2, John Nishio2, Puneet Souda1, Huamin Zhang3 and William A. Cramer3
1 The Pasarow Mass Spectrometry Laboratory, The Jane and Terry Semel Institute for Neuroscience and Human Behavior, The Brain Research Institute and The Molecular Biology Institute, David Geffen School of Medicine, University of California, 405 Hilgard Avenue, Los Angeles, CA 90095, USA
2 College of Natural Sciences, California State University, Chico, California, USA
3 Department of Biological Sciences, Purdue University, West Lafayette, Indiana, USA

Received 10 April 2004; Accepted 13 February 2006

Abstract

A number of strategies have successfully extended plant proteomics into the bilayer domain. Important benefits can be afforded by including a well-resolved intact protein mass spectrum alongside peptide identification experiments; recent studies of thylakoid membranes have yielded new information on the primary structure and covalent post-translational modification of many of the integral proteins. Intact mass proteomics is advancing through the development of core technologies in separations and mass spectrometry, with the goal of providing comprehensive primary structure coverage that includes transmembrane domains with reliable measurements of protein abundance and turnover. To address the limitations associated with separation technologies such as 2D-gel electrophoresis, alternative systems are being investigated and 2D-liquid chromatography of thylakoid membrane proteins, using both denaturing and non-denaturing first dimensions, has been successful, extending separation space and providing intact protein solutions for electrospray-ionization mass spectrometry and top-down proteomics. High-resolution conventional, and Fourier-transform, mass spectrometry is bringing increasing resolution to tandem mass spectrometry allowing for ‘top-down’ mass spectrometry of intact proteins. Thus the core chromatographic technologies already developed for intact mass proteomics of integral membrane proteins also allow their ‘top-down’ analysis. Thylakoid membrane proteins with one and two transmembrane helices have been analysed demonstrating the ability of collision-activated dissociation (CAD) to sequence through transmembrane domains. When analysed by top-down proteomics, the small subunits of the cytochrome b6f complex from Arabidopsis thylakoids reveal novel post-transcriptional/translational modifications including the presence of glutamic acid at position 2 of PetL instead of the proline residue predicted from the gene sequence.

Key words: 2D-chromatography, bilayer, chloroplast, integral membrane protein, proteome, proteomics, thylakoid, top-down.

Introduction

Biological systems employ numerous lipid/protein membranes that play critical roles in many aspects of life. The proteins of the membrane bilayer fall into two general classes; the integral or intrinsic proteins that constitute part of the membrane mass itself and must be extracted by its solubilization, and the peripheral or extrinsic proteins that are membrane-associated and can be displaced without its destruction. The integral proteins are divided into two general (known) structural classes; those possessing the transmembrane ß-barrel porin-type fold and those with one or more transmembrane α-helical domains. Transmembrane helices can be quite accurately predicted and analysis of assigned genomic open reading frames (ORFs) indicates that proteins with this motif make up around 30% of the proteome. Taking the transmembrane porins, short transmembrane unassigned ORFs, and peripheral membrane proteins into consideration, it is clear that membrane-associated proteins constitute as much as half of the proteome. The critical role of membrane proteins in life is further emphasized by the estimate that 70% of drug targets fall in the transmembrane category. Consequently,
proteomics technologies must address the bilayer proteome (Whitelegge et al., 1999). Unfortunately, integral membrane proteins tend to have awkward physico-chemical properties resulting in their general under-representation in most proteomics experiments. The origin of this problem probably arises from the tendency of hydrophobic transmembrane domains to aggregate and precipitate upon removal from the bilayer, exacerbated by the presence of reactive thiols of reduced cysteine residues.

Modern biological mass spectrometry is driving proteomics, and the significance of soft ionization techniques for large biological macromolecules was recognized with the award of the Chemistry Nobel prize to Fenn and Tanaka in 2002 for their discoveries. Since proteins are conveniently identified by short, unique, internal sequences, early proteomics protocols have cleaved intact proteins into groups of shorter peptides for mass spectrometry. Informatics tools are then used to match data from one or more peptides to the genome for identification of the parent gene. Integral membrane proteins are often amenable to this type of analysis as a result of loop regions that yield soluble peptides for convenient mass spectrometric analysis. Intact mass (Whitelegge et al., 1998) or ‘top-down’ (Kelleher et al., 1999) proteomics starts with an intact protein mass measurement prior to intact protein tandem mass spectrometry providing primary structure information for a description of the entire protein. The realization of this goal for integral membrane proteins will allow post-translational modification to be investigated within the bilayer domain.

The separation challenge in proteomics

There are typically thousands of open reading frames in a genome such that the living organism comprises several thousand different proteins of diverse size, sequence, and chemical properties. The most widely used large-scale separation technology in proteomics is the two-dimensional gel electrophoresis (2D-gel) which provides a useful visual display of a subset of the abundant components of a complex protein mixture (Hoogland et al., 2004). Isoelectric focusing used in the first dimension is limited to non-ionic or neutral zwitterionic detergents for solubilizing proteins with their native charge under denaturing conditions. This results in some integral membrane proteins exhibiting a tendency to precipitate and thus become poorly represented in 2D-gel studies. Alternative choices of chaotrope and detergents have addressed the problem and 2D-gels remain quite widely employed in membrane protein research (Molloy et al., 2000). Alternatives replace the first dimension, going to sodium-dodecylsulphate polyacrylamide gel electrophoresis (SDS-PAGE) for the second dimension after using organic solvents to pre-fractionate the membrane system (Ferro et al., 2003; Friso et al., 2004) or non-denaturing separations for the separation of integral membrane protein complexes (Herranen et al., 2004).

Concomitantly, ‘shotgun’ proteomics has appeared, improving throughput for the analysis of the proteome (Washburn et al., 2001). In the shotgun approach the proteome is fragmented into peptides using a specific cleavage agent, such as the enzyme trypsin–this hydrolyses the peptide bond C-terminal to Arg and Lys amino acid residues, provided that the next residue is not Pro. The resulting mixture presents a phenomenal separation challenge because of the presence of hundreds of thousands of different peptides with abundances varying over several orders of magnitude. Consequently, separation space has been expanded through the use of two-dimensional chromatography with strong-cation exchange generating fractions for sequential reverse-phase runs coupled to mass spectrometry (Multi-dimensional protein identification technology; MuDPIT) (Peng et al., 2003).

Modification of shotgun protocols has been required to address the integral membrane proteins (Wu et al., 2003) and good coverage can now be achieved. Recovery of peptides from transmembrane domains is restricted, illustrating the disadvantage of shotgun proteomics: proteins become represented by the peptides most easily recovered and many studies rely on single peptide recoveries for a large proportions of the proteins identified, making false-positive rates a significant issue (Whitelegge, 2002).

Intact mass measurement of integral membrane proteins

In order to perform intact mass or ‘top-down’ proteomics, samples of intact proteins must be prepared that are at a reasonable concentration, quite pure, and dissolved in solvents compatible with ESI-MS. In the case of integral membrane proteins, a suite of protocols has been developed, allowing the mass determinations of proteins with 1–15 transmembrane domains and masses exceeding 100 kDa (Whitelegge et al., 2003; Whitelegge, 2004). An organic acid (typically formic) is used to solubilize the sample and liquid chromatography is used to purify the proteins in aqueous/organic solvent mixtures for protein mass measurements achieving 0.01% accuracy (±3 D at 30 kDa) on traditional quadrupole instruments, demonstrating that the integral membrane proteome is fully compatible with ESI-MS (Whitelegge et al., 1998). In this way, a Fourier-transform mass spectrum (FT-MS) was recorded of bacteriorhodopsin, which has seven transmembrane helices with accuracy exceeding 10 ppm, demonstrating for the first time that ‘top-down’ proteomics by FT-MS is feasible for the integral membrane proteome (Whitelegge, 2003). An intact protein mass measurement is useful because it ‘defines the native covalent state of a gene product and associated heterogeneity’ (Whitelegge et al., 1998). Based on the measured mass of the cytochrome b subunit of the cytochrome b6f complex, the presence of a covalently bound haem was suggested (Whitelegge et al., 2002), subsequently
confirms the following year by high-resolution structures from two groups (Kurisu et al., 2003; Stroebel et al., 2003).

There are limits to the efficiency of separations and mass spectrometers that restrict their ability to accommodate complex protein mixtures. Analysis of the cytochrome b$_{6}$f complex from photosynthetic membranes demonstrated that around ten different proteins could be resolved by size-exclusion chromatography ESI-MS (SEC-MS) (Whitelegge et al., 2002), with the limit for reverse-phase chromatography ESI-MS (RPC-MS) in the range of 50–100 polypeptides (Gómez et al., 2002). Furthermore, elution efficiency of some integral proteins from reverse-phase columns can be quite low, requiring subsequent regeneration of the column. Intact integral membrane proteins can be recovered from a reversed-phase column for downstream ESI-MS using a hybrid chromatography system (Whitelegge, 2005). When a sample of bacteriorhodopsin (50 µg protein, 10 µl) was dissolved in formic acid (90%, 90 µl) and subjected to RPC-MS using organic buffers containing isopropanol as originally reported (Tarr and Crabb, 1983) and developed (Whitelegge, 2004), approximately 60% of the protein eluted during the RPC-MS. To elute the residual protein, the reverse-phase column is coupled to a size-exclusion column and the system equilibrated in chloroform/methanol/aqueous 1% formic acid (4/4/1; by vol.). An injection of formic acid (90%) releases the remaining protein to the mobile phase thereby regenerating the column for further experiments without the possibility of ‘ghost’ peaks. By including the size-exclusion column the protein can be separated from small molecule contamination allowing the mass spectrum to be recorded (Whitelegge, 2005). While this recovery/regeneration scheme is most useful, it does not overcome the limitations of spatial separation. Therefore, these protocols have been applied to purified integral membrane protein complexes (Whitelegge et al., 2002) and membrane subdomains (Gómez et al., 2002, 2003). Application to complete membrane systems requires at least one more dimension of chromatography to provide reasonable coverage.

**Two-dimensional chromatography**

Generating liquid samples for ESI-MS makes liquid chromatography preferable to gel-based technologies because recovery of intact proteins from gels is inefficient and covalent side-reactions modify the protein. 2D-chromatography is being evaluated for dissection of the membrane proteome and since compatibility of this study’s liquid chromatography systems to ESI-MS has been demonstrated, it is logical to reserve these for the second dimension and consider first dimensions that separate the different protein complexes of the membrane. Since native gel experiments have demonstrated that some membrane systems can be solubilized with non-denaturing detergents, size-exclusion chromatography was investigated under non-denaturing conditions for the separation of total membrane complexes with some success. For example, Photosystem I (PS1) could be enriched from membrane fractions of *Synechocystis* sp. PC 6803 allowing a second dimension analysis of the two large reaction-centre subunits, PsaA and PsaB, each with 11 transmembrane helices and masses exceeding 80 kDa (Whitelegge et al., 2003). However, chromatographic resolution in the first dimension was insufficient for analysis of the complexes of lesser abundance without contamination by PS1, the most abundant membrane complex. Reasonably well-resolved separations of this membrane system have been achieved using Blue-native gels (Herranen et al., 2004), but the interface of this technology with liquid chromatography in the second dimension is challenging. An alternative 2D-chromatography system uses chromatofocusing in the first dimension step, which is performed under similar, denaturing, conditions to the isoelectric focusing of 2D-gels (Whitelegge, 2005). Fractions collected at pH steps during elution from chromatofocusing are separated by reverse-phase chromatography with absorbance detection at 214 nm. While providing quite reproducible 2D-chromatograms for quantitative proteomics, it is apparent that larger integral membrane proteins tend to precipitate before or during the first dimension and that they can be difficult to recover from the reverse-phase columns used in the second. Furthermore, protein–protein interaction information related to complexes is lost in the denaturing first dimension.

**Top-down proteomics: technology**

Once solubilized in the aqueous organic solvent, mixtures used for both reverse-phase and size-exclusion chromatography, electrospray-ionization (ESI), as well as matrix-assisted laser desorption ionization (MALDI), mass spectrometry is routine. Recorded ESI mass spectra yield protein mass measurements of resolution and accuracy
comparable to similar measurements on water-soluble proteins (Whitelegge et al., 1998). Following on from this it has been shown that integral membrane proteins are amenable to Fourier-transform mass spectrometry with the associated benefits to resolution and accuracy (Whitelegge, 2003). Collision-activated dissociation (CAD) can be applied to intact integral membrane protein ions and was used for tandem mass spectrometry describing the
primary structure of the four small subunits of the cytochrome $b_{6}f$ complex from both pro- and eukaryotes (Whitelegge et al., 2002). Interestingly, CAD did not yield any sequence information for six or seven C-terminal residues of PetG from *Mastigocladus laminosus* illustrating the need for alternative dissociation mechanisms. Electron-capture, and the related electron-transfer dissociation (ECD, ETD; Zubarev et al., 1998, 1999; Breuer et al., 2004; Syka et al., 2004; Zubarev, 2004) have not yet been applied to published studies of integral membrane proteins, although this is feasible (JP Whitlegge, unpublished data).

Top-down analysis by mass spectrometry was originally described for FT-MS only. However, any mass spectrometer with sufficient resolution to measure the spacing on peptide/protein carbon isotopes, and thus charge, can be employed. A quadrupole-time-of-flight instrument was used to perform a top-down analysis on the integral membrane proteolipid that forms the major proton channel for ATP synthase (AtpH) from *Arabidopsis thaliana* (Whitelegge, 2005) demonstrating $b$- and $y$-series ions that read through both transmembrane helices.

Hydroponic growth systems for *Arabidopsis* are yielding large plants with ample thylakoids for biochemical analysis (A Laganowsky, J Nishio, unpublished data) including the preparation of highly purified active complexes such as the cytochrome $b_{6}f$ complex (H Zhang, W Cramer, unpublished data). The latter preparation was analysed by LC-MS+ as described previously (Whitelegge et al., 2002) with top-down tandem mass spectrometry of the small subunits revealing new information on the *Arabidopsis* complex. Table 1 shows the measured intact masses of the small units, and their assigned identities and post-transcriptional/translational modifications which were confirmed by tandem mass spectrometry (Fig. 1; Table 2). In this study, ProsightPTM was used, a web-based informatics tool developed by Neil Kelleher, for top-down proteomics (LeDuc et al., 2004), in order to identify the proteins and provide a probabilistic computation of the most likely primary structure to account for the intact mass and tandem mass spectrometry data. In all cases, the subunits were correctly identified using ProsightPTM with high confidence. Table 2 shows the probability scores obtained for a number of different searches to probe ProsightPTM performance. The second column shows the success of ProsightPTM in identifying the correct subunits from the entire *Arabidopsis* database; in all cases the matches and probability scores listed were the top hits in the database search. In the case of PetL the search was repeated at 25 ppm stringency instead of 100 ppm, resulting in a smaller probability score indicating a stronger hit. This nicely illustrates the advantages of high-resolution mass spectrometers for top-down proteomics and more precise protein identification. However, complete primary structure assignment in all cases required manual intervention with the correct sequence and covalent modifications being entered into the manual feature of ProsightPTM. Thus the optimal number of $b$- and $y$-ions was matched, again increasing assignment confidence (smaller probability scores in Table 2, column 4) for all subunits except PetL.

<table>
<thead>
<tr>
<th>Protein</th>
<th>db Search$^a$</th>
<th>Actual sequence$^b$</th>
</tr>
</thead>
<tbody>
<tr>
<td>PetL</td>
<td>6.29E-14 (100 ppm) 4.23E-18 (25 ppm)</td>
<td>1.04E-14 (250 ppm) 4.23E-18 (25 ppm)</td>
</tr>
<tr>
<td>PetM</td>
<td>3.21E-09 (100 ppm)</td>
<td>7.80E-18 (100 ppm)</td>
</tr>
<tr>
<td>PetG</td>
<td>1.24E-08 (100 ppm)</td>
<td>1.37E-11 (100 ppm)</td>
</tr>
<tr>
<td>PetN</td>
<td>3.08E-09 (100 ppm)</td>
<td>4.42E-30 (100 ppm)</td>
</tr>
<tr>
<td>PetNox</td>
<td>4.61E-12 (100 ppm)</td>
<td>5.34E-17 (100 ppm)</td>
</tr>
<tr>
<td>PetL</td>
<td>2.53E-29 (100 ppm)</td>
<td>1.79E-27 (25 ppm)</td>
</tr>
</tbody>
</table>

a Crude Pscore for matches of peak lists with database entries for the cognate protein. Search stringency is given in parentheses. Peak lists were obtained by running Bayesian peptide reconstructions on tandem mass spectra (Analyst, Applied Biosystems). In all cases the identified protein was the top hit for each db search.

b Single protein mode was used to match peptide data with the final assigned protein sequences including post-transcriptional/translational modifications.

Sequencing modifications of membrane proteins

---

Table 2. ProsightPTM search results for tandem mass spectra from top-down analysis of Arabidopsis thaliana cytochrome $b_{6}f$ complex small subunits

---

The initial database search was only successful when the measured mass of the protein was practically ignored, such that the much larger gene product was included in the search window. Success in identification was achieved because the C-terminus was correctly reported allowing many y-ion matches. Proteins showing both N- and C-terminal processing will likely be tougher to identify by top-down proteomics. The analysis of the small subunits of *Arabidopsis* cytochrome $b_{6}f$ complexes revealed that the chloroplast-encoded subunits all retained their initiating
formylMet residues, as previously reported for spinach (Whitelegge et al., 2002). In spinach, residue 2 of PetL was found to be Phe instead of the Ser residue predicted from the sequence, whereas in Arabidopsis Glu was found to replace Pro. The consistency of modification at position 2 of PetL provides further evidence for an RNA editing event, of unknown functional consequence.

The signal peptide cleavage site of PetM (Figs 1, 2) is consistent with the cleavage site observed for the spinach subunit.

Fig. 2. Identification of integral membrane proteins using top-down proteomics and ProsightPTM. A tandem mass spectrum obtained by collision-activated dissociation of a triple-charged ion (1397.4) from the 4198 kDa subunit was deconvoluted to the zero-charge state using the Bayesian peptide reconstruct tool (Analyst, Applied Biosystems). The peak list was exported and submitted to ProsightPTM (https://prosightptm.scs.uiuc.edu) along with the intact mass and used to interrogate an Arabidopsis thaliana database for potential matches. When the mass tolerance for the parent was extended to 20 kDa to include the possibility of long signal sequences, the PetM protein was identified by a y-series of ions with a probability score of 3.20693e-09 using ProsightPTM, set at 100 ppm for daughter-ion matches. Based upon the correct C-terminal fragments and the intact protein mass it was concluded that the signal peptide directing this nuclear-encoded gene product from cytoplasm to the thylakoid was cleaved after position 85. The mature PetM sequence was then entered into the manual feature of ProsightPTM allowing for inclusion of b-ions raising the identification confidence to 7.79893e-18, also at 100 ppm daughter tolerance. Note that 26 out of the 33 assigned ions achieve matches at less than 20 ppm.
Hybrid ion-trap Fourier-transform mass spectrometers for improved top-down analysis

Top-down Fourier-transform mass spectrometry has traditionally been performed by direct sample infusion thereby avoiding the changes in concentration that are intrinsic to liquid chromatography. The development of new ion-trap FT-MS instrumentation that regulates the number of ions transferred to the FT-ICR cell is very exciting because it largely overcomes these earlier issues (Page et al., 2005). Using this technology Wu and coworkers achieved 95% sequence coverage of a human growth hormone with a 200 fmol sample in an LC-MS top-down experiment (Wu et al., 2004). The authors’ studies have shown that data such as this can be searched at 5 ppm tolerance with ProsightPTM, providing further gains in accuracy of protein identification (JP Whitelegge, unpublished results). Top-down proteomics is coming of age and has been reviewed recently (Kelleher, 2004; Bogdanov and Smith, 2005; Meng et al., 2005).

Summary

Integral membrane proteins are amenable to analysis by electrospray-ionization tandem mass spectrometry, once purified in an aqueous organic solvent mixture compatible with their solubility. ProsightPTM is proving to be a useful informatics tool towards automating top-down data interpretation, although manual intervention is required for optimal primary structure assignment. Transmembrane domains are conveniently sequenced by tandem mass spectrometry of the intact proteins.

Acknowledgements

Richard LeDuc, Yong-Bin Kim, Andrew Forbes, and the rest of the software development team in Neil Kelleher’s Laboratory are thanked. JPW acknowledges support from NIH (U01 DE016275-01, PO1 NS049134-01, PO1 HL80111-01).

References


