Positional proteomics: selective recovery and analysis of N-terminal proteolytic peptides

Lucy McDonald[1], Duncan H L Robertson[1], Jane L Hurst[2] & Robert J Beynon[1]

Bottom-up proteomics is the analysis of peptides derived from single proteins or protein mixtures, and because each protein generates tens of peptides, there is scope for controlled reduction in complexity. We report here a new strategy for selective isolation of the N-terminal peptides of a protein mixture, yielding positionally defined peptides. The method is tolerant of several fragmentation methods, and the databases that must be searched are substantially less complex.

Bottom-up proteomics operates at the level proteolytic peptides, generated from single proteins or from complex mixtures of proteins[1]. These peptides, generated by exhaustive proteolysis to limit peptides *in vitro*, are then analyzed by various mass spectrometric methods. Mass spectrometric analysis yields either the masses of a formally connected set of peptides that were all derived from a single protein (peptide mass fingerprinting) or, by tandem mass spectrometry, sequence-derived information that can identify the parent protein of a single peptide[2,3]. It can be argued that more peptides are analyzed than strictly necessary, and comprehensive proteomic analysis should focus on the minimal number of peptides that are required for protein identification. Methods such as ICAT[4] implicitly adopt this principle, in as much as the selective chemistry recovers only those peptides that contain at least one cysteine residue. Cysteine-mediated peptide recovery, however, is likely to abstract more than one peptide for each protein, and it is not possible to target the recovered peptide(s) positionally, as cysteine residues can occur anywhere in the protein sequence.

Positionally defined peptides would yield a substantial information gain in protein identification strategies. Most obviously, the two positional locations within every protein are the extreme ends—the N-terminal and the C-terminal peptides. Methods for recovery of C-terminal peptides have been reported, predominantly based on the ability of a catalytically disabled trypsin, anhydrotrypsin, to selectively bind peptides that terminate in a lysine or arginine residue[5,6]. There are several reports that indicate routes to selective recovery of N-terminal peptides, including specific N-terminal sequencing by mass spectrometry of gel-separated and blotted proteins[7], selective modification of N-terminal serine or threonine residues[8], modification of the hydrophobicity of a peptide mixture to preferentially expose N-terminal peptides by

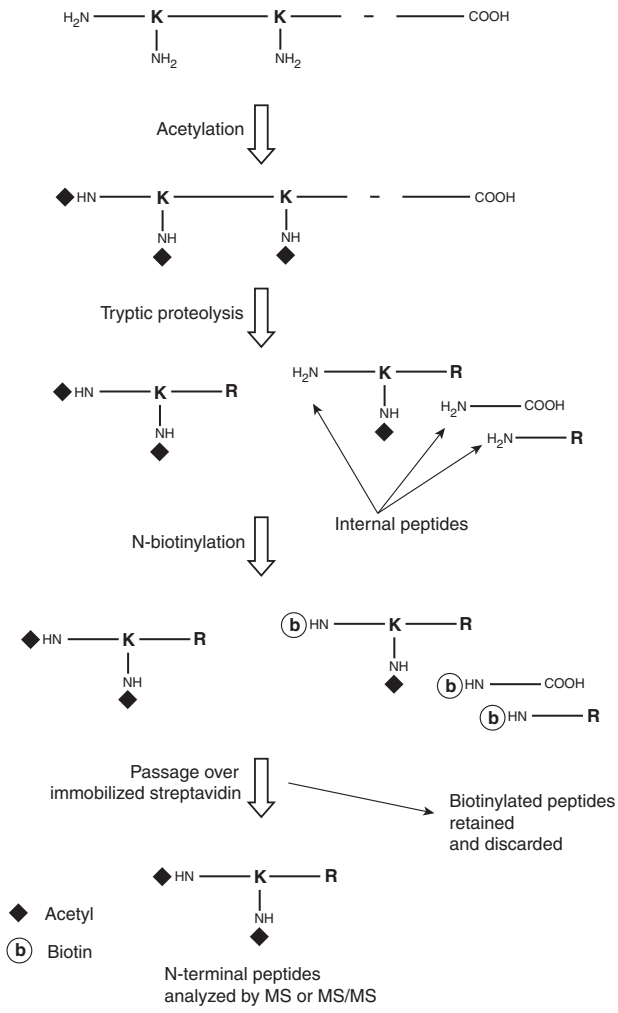

Figure 1 | Protocol for recovery of N-terminal peptides in a proteome. Free α- and ε-amino groups are acetylated before proteolysis (trypsin in the figures, but potentially any other fragmentation method), which is followed by biotinylation of proteolytically exposed α-amino groups. Subsequent subtractive binding to immobilized streptavidin creates a preparation enriched in those peptides that were originally derived from the N terminus, blocked by acetylation and therefore refractory to biotinylation.

[1]Protein Function Group and [2]Mammalian Behavior and Evolution Group, Faculty of Veterinary Science, University of Liverpool, Crown Street, Liverpool L69 7ZJ, UK. Correspondence should be addressed to R.J.B. (r.beynon@liv.ac.uk).

RECEIVED 23 AUGUST; ACCEPTED 3 OCTOBER; PUBLISHED ONLINE 18 NOVEMBER 2005; DOI:10.1038/NMETH811

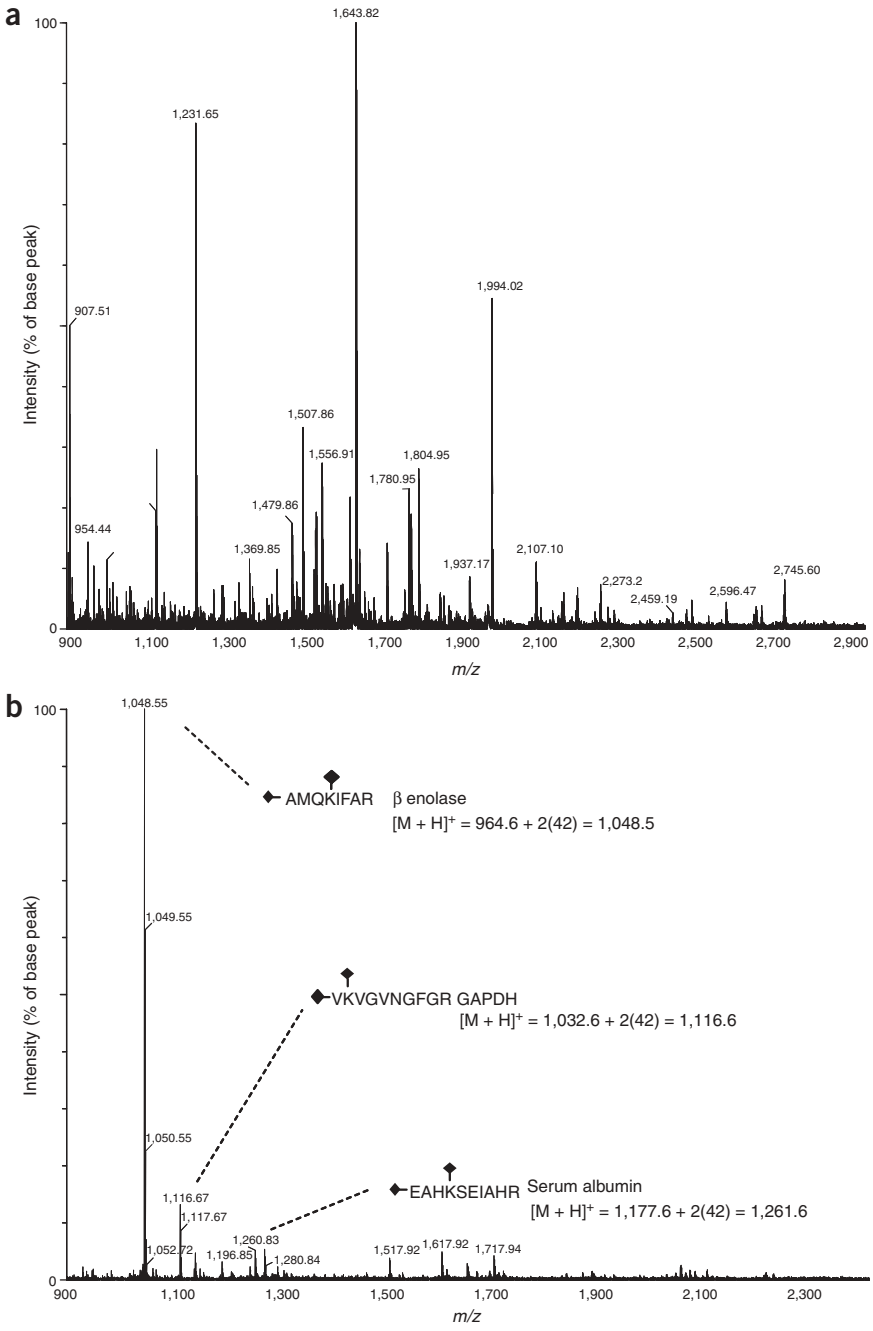

Figure 2 | Isolation of N-terminal peptides from soluble proteins of mouse skeletal muscle. (**a**,**b**) Skeletal muscle was homogenized in 10 ml of 20 mM sodium phosphate buffer (pH 8.0) and centrifuged for 45 min at 13,000*g*. The resultant supernatant fraction was used without further purification for preparation of N-terminal peptides (acetylation, tryptic proteolysis, N-biotinylation and subtractive purification). Detailed protocols are available in **Supplementary Methods** online. The entire tryptic digest of the mixture was analyzed by MALDI-ToF mass spectrometry (**a**). After application of the positional simplification protocol, the MALDI-ToF mass spectrum (**b**) contained major ions labeled in the figure, mass shifted by 42 Da through the addition of acetyl groups.

diagonal chromatography[9,10] and selective capture of all non–N-terminal peptides by amine scavenging beads[11,12]. We report here a new approach to selective recovery of the N-terminal-most peptides of a complex protein mixture. The method, based on subtractive removal of internal peptides, is not reliant on any particular endopeptidase cleavage—a flexibility that can compensate for the limitations in N-terminal peptide size distributions. Moreover, in contrast to other approaches[7,11,12], protein N termini that are naturally acetylated are automatically included in the analyte set. Indeed, if stable isotope–labeled acetic anhydride was used, it would be possible to identify and discriminate between naturally and chemically acetylated peptides. In brief, all available amino groups are blocked by acetylation. Subsequently, proteolysis generates new peptides, and all but the N-terminal peptide (whether naturally or artificially acetylated) expose a new amino group that is subsequently biotinylated. These biotinylated internal peptides are removed by recovery onto immobilized avidin or strepatividin, leaving behind the set of N-terminal peptides (**Fig. 1** and **Supplementary Methods** online, which contains a protocol for N-terminal peptide recovery and a description of the analysis of protein databases).

One of our major interests is in proteome dynamics in skeletal muscle[13–15]. The tryptic digest of the soluble protein fraction of mouse skeletal muscle contains peptides derived from a large number of proteins, and a matrix-assisted laser desorption/ionization–time-of-flight (MALDI-ToF) spectrum on an instrument of medium-level performance (resolution 12,000 FHWM (full width at half maximum)) yielded a detailed but complex mass spectrum (**Fig. 2a**). Owing to the complexity of the peptide mixture, we were unable to identify any N-terminal peptides in the spectrum. We passed the N-acetylated, trypsin-digested, biotinylated mixture over immobilized streptavidin. The unbound eluate gave a much simpler mass spectrum (**Fig. 2b**), and we were able to assign the highest intensity signals to true N-terminal peptides, confirmed by tandem mass spectrometry (**Supplementary Fig. 1** online). To test the method with a more complex mixture, we applied the same protocol to the soluble proteins of mouse liver. After purification of N-terminal peptides, the MALDI-ToF spectrum remained complex. By liquid chromatography–tandem mass spectrometry, and even without optimized separation or mass spectrometric analysis, many peptides (over 90) could immediately be assigned as N termini of mouse proteins (**Supplementary Fig. 2** and **Supplementary Table 1** online). As predicted, all terminated at C-terminal arginine residues. Moreover, the data were consistent with known or inferred N-terminal

processing (removal of initiator methionine, loss of signal peptide or propeptide) but in other cases have provided new information on N-terminal processing of liver proteins. All identifications were from a search of the entire database of mouse proteins rather than a restricted N-terminal database—there were virtually no peptides identified as internal sequences.

An analysis of extracted N-terminal peptides from mouse entries in Swissprot (**Supplementary Fig. 3** online) confirmed that over 85% of all proteins yielded an informative N-terminal peptide with trypsin digest, and that this value rose to almost 90% if we used two endopeptidases (trypsin and endopeptidase Glu-C). Thus, a substantial fraction of proteins in a proteome can be uniquely identified simply by the mass of the N-terminal peptide, using one or multiple endopeptidase digests. But the complexity of most N-terminal peptide preparations would require liquid chromatography–electrospray tandem mass spectrometry or liquid chromatography–MALDI tandem mass spectrometry. Even partial sequence data considerably enhance identification, and remove the need for multiple proteolytic digests—the residual 4% unidentifiable proteins represent sequences in the database that are either replicated entries or which represent paralogous proteins. There has recently been discussion about the uncertainty of 'one hit wonders' in proteomics, and we conjecture that part of the uncertainty relates to the lack of information about the location of the peptide in the parent proteins. A positional proteomics strategy anchors the peptides at a precise location within the parent protein, greatly reducing the search space for identification algorithms. As an average protein might be predicted to yield 50 tryptic peptides, the approximate reduction in search space is also 50-fold. Further, selective isolation and partial sequencing of N- and C-terminal peptides would allow virtually full length PCR-amplification of the cDNA corresponding to an expressed protein sequence.

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

Supported by grants to R.J.B. & J.L.H. from the Natural Environment Research Council and the Biotechnology and Biological Sciences Research Council. We are grateful to M. Doherty for assistance with the mass spectrometry.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Published online at http://www.nature.com/naturemethods/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Bogdanov, B. & Smith, R.D. *Mass Spectrom. Rev.* **24**, 168–200 (2004).
2. Standing, K.G. *Curr. Opin. Struct. Biol.* **13**, 595–601 (2003).
3. Steen, H. & Mann, M. *Nat. Rev. Mol. Cell Biol.* **5**, 699–711 (2004).
4. Smolka, M.B., Zhou, H., Purkayastha, S. & Aebersold, R. *Anal. Biochem.* **297**, 25–31 (2001).
5. Kasai, K. *J. Chromatogr.* **597**, 3–18 (1992).
6. Sechi, S. & Chait, B.T. *Anal. Chem.* **72**, 3374–3378 (2000).
7. Yamaguchi, M. *et al. Anal. Chem.* **77**, 645–651 (2005).
8. Chelius, D. & Shaler, T.A. *Bioconjug. Chem.* **14**, 205–211 (2003).
9. Gevaert, K. *et al. Nat. Biotechnol.* **21**, 566–569 (2003).
10. Martens, L. *et al. Proteomics* **5**, 3193–3204 (2005).
11. Kuhn, K. *et al. J. Proteome Res.* **2**, 598–609 (2003).
12. Kuhn, K. *et al. Proteomics* **5**, 2364–2368 (2005).
13. Doherty, M.K., Whitehead, C., McCormack, H., Gaskell, S.J. & Beynon, R.J. *Proteomics* **5**, 522–533 (2005).
14. Doherty, M.K. *et al. Proteomics* **4**, 2082–2093 (2004).
15. Hayter, J.R., Robertson, D.H., Gaskell, S.J. & Beynon, R.J. *Mol. Cell. Proteomics* **2**, 85–95 (2003).