Microbial proteomics: how far have we come?

Stuart J Cordwell[1,2]

[1] School of Molecular Bioscience
Building GO8, Maze Crescent, The
University of Sydney, NSW 2006
[2] Discipline of Pathology
School of Medical Sciences, The
University of Sydney, NSW 2006
Tel 02 9351 6050
Email
stuart.cordwell@sydney.edu.au

Introduction

It is now over 15 years since the beginning of the microbial genome era. At that time, great interest was invested in the idea of understanding bacterial genome dynamics via the analysis of their protein complements or 'proteomes'[1,2]. As bacterial analysis had driven the early advances in genomic sequencing, these organisms were amongst the first subjected to protein-based studies. It is fair to suggest in hindsight that the original hype did not match the outcomes. At the time the term proteome was coined[3], *en masse* protein analysis meant little more than two-dimensional electrophoresis gels and mainly Edman sequencing for protein identification. As such, only the most abundant cellular constituents could be analysed and interest soon turned for many researchers towards microarray-based transcriptomics where a much greater percentage of the genome could be surveyed. Despite this, substantial and continuing evidence has demonstrated that transcript and protein levels (and, of course, the associated substrates and products of the predicted protein functions) often do not correlate[4,5]. Therefore, a truly 'systems biology' organism-wide approach is necessary, where genomics, transcriptomics, proteomics and metabolomics are integrated to understand how microbes respond to changes in their genetic or physiological environments, and thus generate new hypotheses for functional understanding. This article will examine how proteomics technology has evolved to the stage at which the total proteome can be elucidated, and provide a guide for 'best practice' for undertaking successful proteomics analyses.

A near-to-total proteome?

One of the perceived major shortcomings of proteomics is the widespread belief that the technology is not capable of surveying enough of the genome to provide an organism-wide appraisal. Phenomenal technology advances in the field of mass spectrometry (MS) have now enabled the proteome generation[6,7]. The combination of gel-free approaches, utilising peptide separation by one or two dimensions of liquid chromatography (LC) coupled to tandem-MS (MS/MS) for peptide identification ('shotgun' proteomics) is now capable of generating information on a very significant proportion of a microbial proteome (75–80% expressed ORF coverage). In this scenario, protein samples are proteolytically digested with an enzyme (or to improve coverage, a series of proteases with different specificities) and the resulting complex peptide mixtures extensively separated using several alternative LC strategies[6,7]. At the end of the day, the amount of identifications are based purely on the MS 'space', that is the available instrument time (to run the number of LC fractions used to separate the original complex peptide mixture; weeks of instrument time may be required to achieve near-complete proteome depth) and the scan speed of the chosen instrument (new instruments can generate several tens of MS/MS events per MS scan; reviewed in 8). Increased LC separation can aid in overcoming abundance (peptides from abundant proteins dominating the resulting data files) and ionisation (some peptides ionise better than others and are thus repeatedly chosen for MS/MS fragmentation) issues.

Proteome coverage is inversely proportional to the overall number of predicted genes/proteins, which means that organisms with relatively smaller genomes tend to have higher percentages covered using proteomics. For example, approximately 90% (620 of 700 predicted) of *Mycoplasma pneumoniae* proteins have been identified[9,10], 70% of the proteome of *Campylobacter jejuni*[11], and above 50% of *Bacillus subtilis*[12,13], *Mycobacterium tuberculosis*[14] and *Staphylococcus aureus*[15] amongst others, have now been identified. The issue is now becoming how best to 'close' the proteome. One technique that has been applied for this approach is selective reaction monitoring (SRM)[8,16]. SRM is an approach generally employed to undertake relative or absolute quantification of a chosen peptide/protein for validation of data generated by high-throughput proteomics approaches. In addition to this, SRMs can be designed for proteotypic peptides representing proteins not previously identified in a high-throughput screen (for example, by gel-free shotgun proteomics). Unlike shotgun proteomics, where the goal is to identify as many peptides as possible to increase proteome coverage, an SRM assay is targeted to look for only a handful of well-defined peptide species, based on mass, LC retention time and either previously observed or predicted fragment ions. The precursor and fragment ion mass pairs are referred to as 'transitions'. The relative intensities of each set of transitions (usually based on 2–5 per proteotypic peptide) compared across samples can be used to provide relative quantitation (Figure 1), or in the presence of a 'spiked' synthetic peptide of known concentration, absolute quantitation[17]. In the context of closing the proteome, proteotypic peptides representing previously unidentified but predicted proteins are assayed to determine whether the protein itself is expressed. While these assays are limited by the available MS space, such that in most configurations only 50–100 peptides can be examined, once established, quantitative data can also be generated.

Once a near-to-total proteome is identified, the next issue is how best to utilise and interrogate those data to provide real biological information. This step requires quantification of proteins compared across samples ('relative' protein quantification) or within a sample ('absolute' quantification)[7]. In relative quantification, peptide samples from one condition (for example, 'control') are labelled with a mass tag (for example, in stable isotope labelling of amino acids in cell culture [SILAC][18] or isobaric tags for relative and absolute quantitation [iTRAQ]), while those from a second condition (for example, 'test') are labelled with a different tag. Samples are mixed together and separated using LC. Quantitation then occurs in the MS scan (for SILAC) or by release of the mass tag reporter fragment ions in MS/MS (for iTRAQ). In this way the relative abundances of peptides/proteins can be discerned between samples. In absolute quantification, isotopically labelled synthetic peptides representing proteins of interest can be spiked into a complex sample and signal from the synthetic peptide of known quantity compared to that from the non-labelled peptide from the original sample. Once that ratio is determined, the copy number per cell can be generated provided the total number of cells in the starting sample is known. Absolute quantification, therefore, allows determination of the stoichiometric relationship between proteins in a complex or cell, and if post-translationally modified, between the native and altered forms of a peptide. Recent studies have successfully employed this approach to examine the *Leptospira interrogans* proteome[17,19].

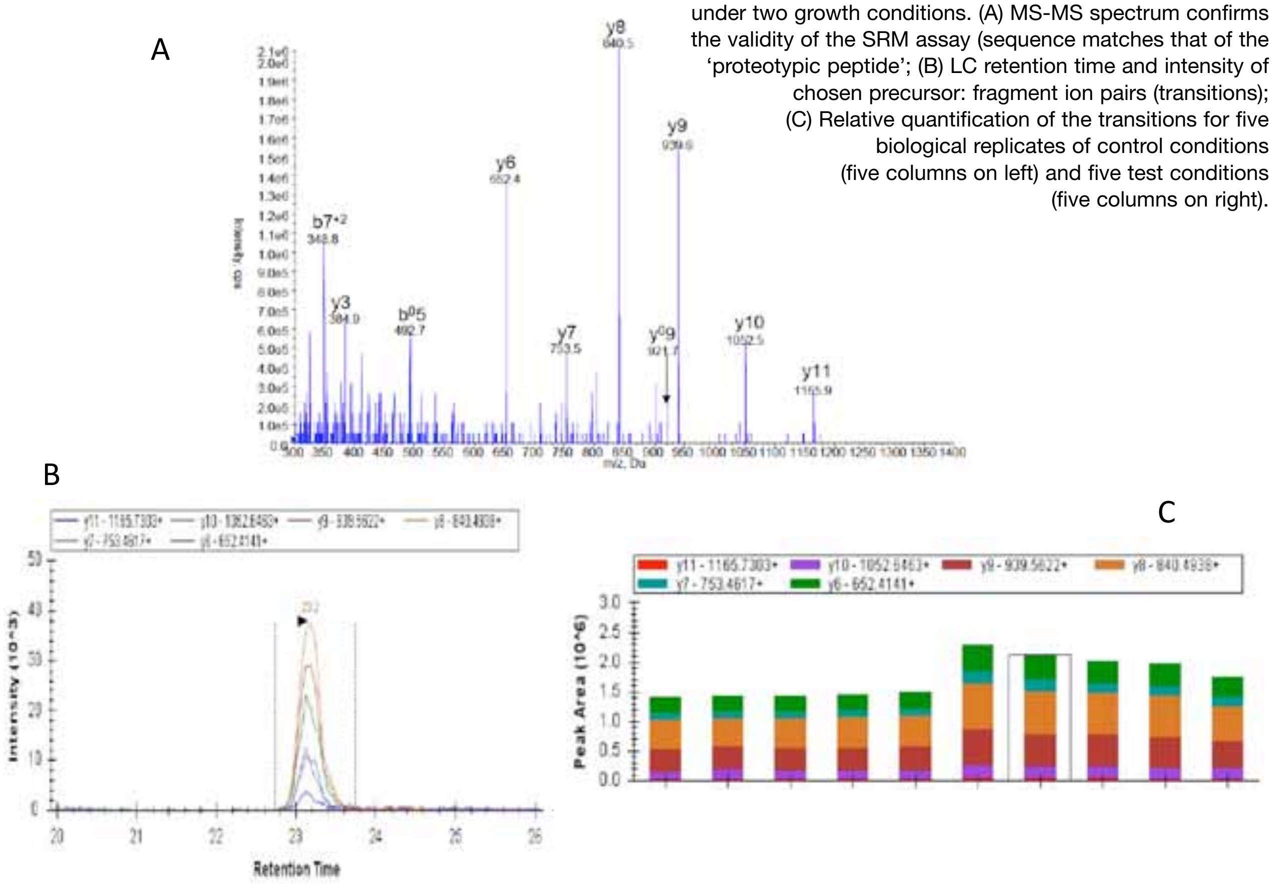


Figure 1. Selected reaction monitoring (SRM) of a peptide compared across five biological replicates of a bacterial cell under two growth conditions. (A) MS-MS spectrum confirms the validity of the SRM assay (sequence matches that of the 'proteotypic peptide'; (B) LC retention time and intensity of chosen precursor: fragment ion pairs (transitions); (C) Relative quantification of the transitions for five biological replicates of control conditions (five columns on left) and five test conditions (five columns on right).

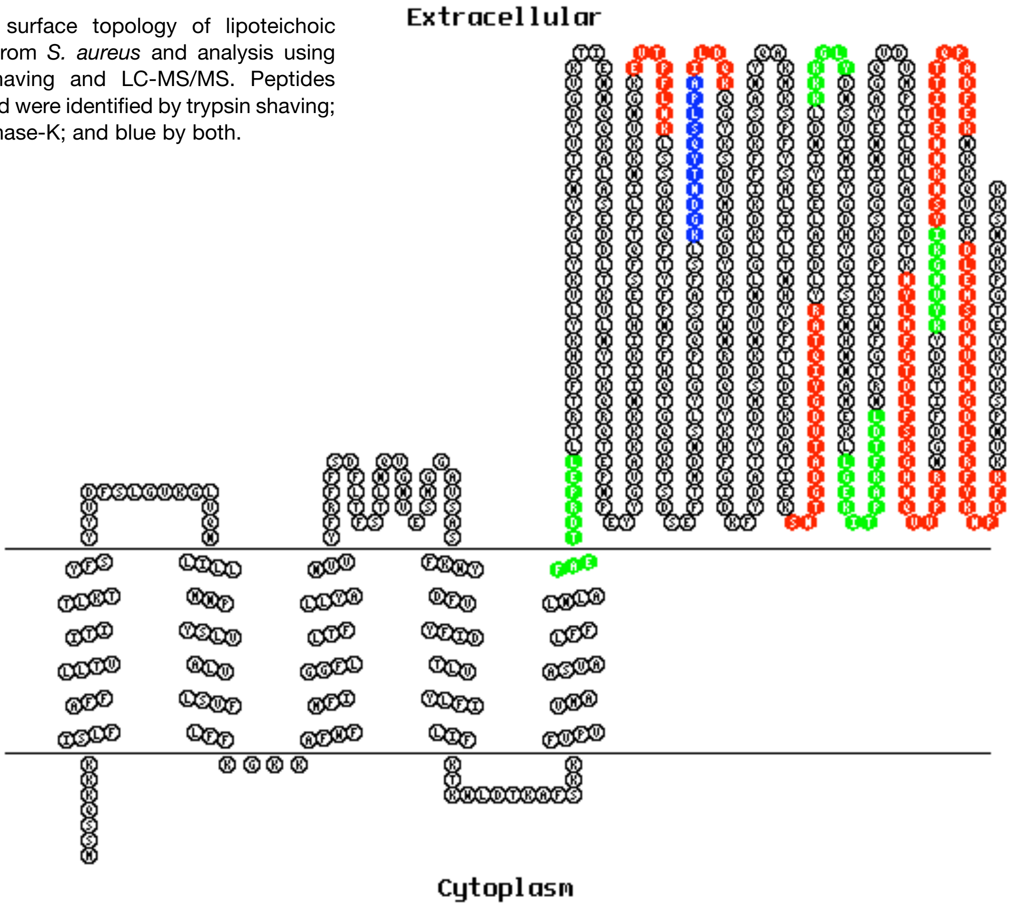
The elusive bacterial membrane proteome

In gel-based approaches, the poor solubility of many membrane proteins, particularly integral membrane proteins with several transmembrane-spanning regions, meant they were severely under-represented in proteome studies. Gel-free approaches, however, overcome this issue since hydrophilic and soluble regions within otherwise insoluble proteins can be readily cleaved by the protease of choice and identified by LC-MS/MS (Figure 2). Coverage of membrane and membrane-associated proteins is, therefore, now governed by the above rules regarding peptide analysis, provided there are sufficient hydrophilic peptides (non-transmembrane spanning regions) for analysis. Beyond this, there has been a significant push for understanding protein orientation and epitope mapping of membrane-associated proteins. A technique that is particularly useful, at least in Gram positives, is cell surface shaving[20]. In this approach, whole cells are briefly incubated with a proteolytic enzyme resulting in the cleavage of any surface-exposed peptides. These are then amenable to analysis by LC-MS/MS for identification (Figure 2). The technique has shown great promise in generating new vaccine targets against *Streptococcus pyogenes*[21] and *S. aureus*.

Post-translational modification of bacterial proteins: now a reality

Bacteria have traditionally been thought of as proteomically 'simple', in that each gene is likely to encode for only a single protein. This is unlike eukaryotic systems, where extensive post-translational modification of proteins increases the proteomic complexity. In more recent times, however, MS-based approaches to understanding post-translational modifications (PTM) have been applied to bacteria and identified high levels of complexity, particularly based around serine/threonine/tyrosine phosphorylation[22], *N*- and *O*-linked glycosylation[23], acetylation[24] and protein cleavage. Furthermore, *in silico* analysis of a wide variety of bacterial genomes has highlighted many 'hypothetical' genes with sequence similarity to the kinases/phosphatases, acetyltransferases and glycosyltransferases that are likely to mediate these modifications. Eukaryotic-type phosphorylation events have now been described in a number of bacterial studies, including *B. subtilis* and *Escherichia coli*; however, the physiological significance of these events remain unknown. Phosphorylation appears to occur on only very minute fractions (for example, <0.1-1%) of well-characterised proteins and thus understanding their role may be troublesome. As yet, site-

Figure 2. Cell surface topology of lipoteichoic acid synthase from *S. aureus* and analysis using cell surface shaving and LC-MS/MS. Peptides highlighted in red were identified by trypsin shaving; green by proteinase-K; and blue by both.

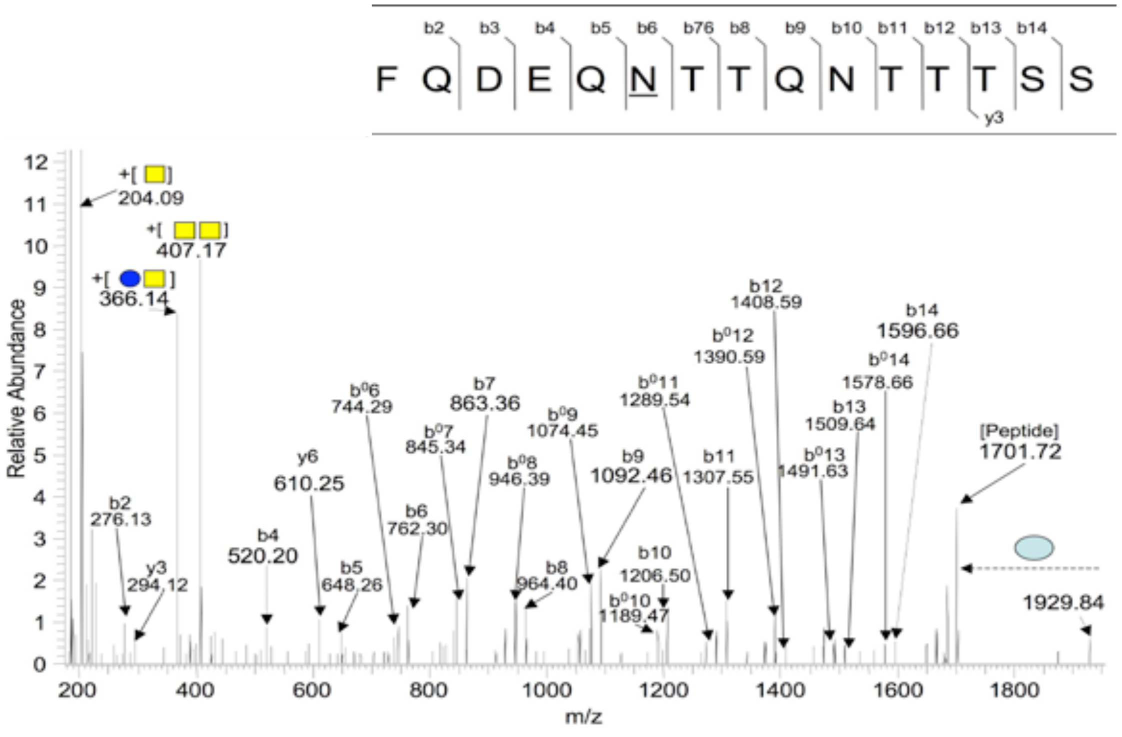


directed studies have not been undertaken. Conversely, the effect of bacterial adherence and invasion on host signalling is a burgeoning area, facilitated by the large amounts of quantitative data that can be generated by phosphoproteomics approaches[25]. *N*- and *O*-linked glycosylation is somewhat better understood than bacterial phosphorylation. *O*-linked glycans have been identified predominantly in Gram-negative bacteria and are found most frequently associated with pilin and flagellar proteins (for example, in *Neisseria* sp. and *C. jejuni*[23]). *N*-linked glycosylation is best studied in *C. jejuni* (Figure 3) and occurs on a substantial number of periplasmic and membrane-associated proteins[26]. Site-directed mutagenesis studies have, however, been unable thus far to determine the underlying function of the modification in virulence, although glycosylation deletion mutants are defective at adhering to host cells. It appears likely that a general role for glycosylation will involve protein stability and/or influence immunity. Finally, it is also pertinent to discuss protein degradation and cleavage in the context of PTM. Until recently, degraded proteins could only be identified by gel-based approaches, since peptide-based shotgun techniques rely on peptide analysis without an appreciation of the native form from which those peptides were derived (that is to say, without protein context). In recent times, terminal amine labelling of substrates (TAILS)[27,28] has provided an opportunity to utilise the technical advances of LC-MS/MS to specifically identify *N*- or *C*-termini from amongst a complex peptide mixture. Proteomics, particularly MS-based approaches, has greatly facilitated the definition of PTM in several bacterial species on a genome-wide scale.

Correlations between transcriptomics, proteomics and metabolomics

We view proteomics as a step in a systems biology pipeline that additionally requires genomics, transcriptomics and metabolomics. Although only a handful of studies have attempted to examine eukaryotic cells in a truly 'systems' approach, some pertinent points can be made about how these data are integrated. Our expectations of molecular biology are that elevated transcript should result in elevated protein expression, resulting in the possible decrease of that protein's substrate and an increase in the amount of the product catalysed by that reaction. Reality, however, even in supposedly simple bacterial cells suggests this is not always the case. In our own preliminary studies examining systems approaches in microbial pathogens, we have seen cases where an entire biochemical pathway is elevated under a change in physiological conditions at the protein and corresponding metabolite levels, but that no change can be detected at the transcript level. This is because systems analyses require temporal profiling to understand the dynamics of cell response. Importantly, however, this further

Figure 3. LC-MS/MS of a glycopeptide from *C. jejuni* using higher energy collision dissociation fragmentation. Light blue circle shows the presence of the asparagine-linked bacillosamine sugar; yellow squares indicate Gal*N*Ac; dark blue indicates a hexose sugar.



emphasises the need to examine all aspects of microbial '-omics' to understand biological function.

Conclusions

New technology is facilitating an era in which proteomics is capable of delivering on promises. We have now reached a point at which the complete microbial proteome is a reality. Relative and absolute quantification provide the opportunity to profile the microbial response to a change in environment at the protein level, in an ordered manner allowing clear elucidation of critical biochemical pathways, in the context of systems biology. While not the focus of this review, it is imperative to establish that these technologies can only be successfully implemented in a complementary manner by a devotion of resources to the bioinformatics needed to collate and interrogate the vast swathes of novel data in the hope of generating important new hypotheses for molecular and further '-omics' interrogation.

References

1. VanBogelen, R.A. *et al*. (1999) Diagnosis of cellular states of microbial organisms using proteomics. *Electrophoresis* 20, 2149–2159.
2. Link, A.J. *et al*. (1997) Comparing the predicted and observed properties of proteins encoded in the genome of *Escherichia coli* K-12. *Electrophoresis* 18,1259–1313.
3. Wilkins, M.R. *et al*. (1996) From proteins to proteomes: large-scale protein identification by two-dimensional electrophoresis and amino acid analysis. *Biotechnol.* 14,61–65.
4. Gygi, S.P. *et al*. (1999) Correlation between protein and mRNA abundance in yeast. *Mol. Cell. Biol.* 19, 1720–1730.
5. Griffin, T.J. *et al*. (2002) Complementary profiling of gene expression at the transcriptome and proteome levels in *Saccharomyces cerevisiae*. *Mol. Cell. Proteomics* 1, 323–333.
6. Yates, J.R. III *et al*. (2009) Proteomics by mass spectrometry: approaches, advances, and applications. *Annu. Revs. Biomed. Eng.* 11, 49–79.
7. Xie, F. *et al*. (2011) Liquid chromatography – mass spectrometry-based quantitative proteomics. *J. Biol. Chem.* 286, 25443–25449.
8. Malmström, L., Malmström, J. and Aebersold, R. (2011) Quantitative proteomics of microbes: principles and applications to virulence. *Proteomics* 11, 2947–2956.
9. Catrein, I. and Herrmann, R. (2011) The proteome of *Mycoplasma pneumoniae* – a supposedly 'simple' cell. *Proteomics* 11, 3614–3632.
10. Kühner, S. *et al*. [2009] Proteome organization in a genome-reduced bacterium. *Science* 326, 1235–1240.
11. Cordwell, S.J. *et al.* (2008) Identification of membrane-associated proteins from *Campylobacter jejuni* strains using complementary proteomics technologies. *Proteomics* 8, 122–139.
12. Becher, D. *et al.* (2011) From the genome sequence to the protein inventory of *Bacillus subtilis*. *Proteomics* 11, 2971–2980.
13. Otto, A. *et al.* (2010) Systems-wide temporal proteomic profiling in glucose-starved *Bacillus subtilis*. *Nature Commun.* 1, 137.
14. de Souza, G.A. and Wiker, H.G. (2011) A proteomic view of mycobacteria. *Proteomics* 11, 3118–3127.
15. Hempel, K. *et al.* (2011) Quantitative proteomic view on secreted, cell surface-associated, and cytoplasmic proteins of the methicillin-resistant human pathogen *Staphylococcus aureus* under iron-limited conditions. *J. Proteome Res.* 10, 1657–1666.
16. Gallien, F. *et al.* (2011) Selected reaction monitoring applied to proteomics. *J. Mass Spectrom.* 46, 298–312.
17. Schmidt, A. *et al.* (2011) Absolute quantification of microbial proteomes at different states by directed mass spectrometry. *Mol. Syst. Biol.* 7, 510.
18. Cox, J. and Mann, M. (2007) Is proteomics the new genomics? *Cell* 130, 395–398.
19. Malmström, J. *et al.* [2009] Proteome-wide cellular protein concentrations of the human pathogen *Leptospira interrogans*. *Nature* 460, 762–765.
20. Solis, N. and Cordwell, S.J. (2011) Current methodologies for proteomics of bacterial surface-exposed and cell envelope proteins. *Proteomics* 11, 3169–3189.
21. Rodriguez-Ortega, M. *et al.* [2006] Characterization and identification of vaccine candidate proteins through analysis of the group A *Streptococcus* surface proteome. *Nature Biotechnol.* 24, 191–197.
22. Macek, B. and Mijakovic, I. (2011) Site-specific analysis of bacterial phosphoproteomes. *Proteomics* 11, 3002–3011.
23. Nothaft, H. and Szymanski, C.M. (2010) Protein glycosylation in bacteria: sweeter than ever. *Nature Rev. Microbiol.* 8, 765–778.
24. Jones, J.D. and O'Connor, C.D. (2011) Protein acetylation in prokaryotes. *Proteomics* 11, 3012–3022.
25. Manes, N.P. *et al.* (2011) Discovery of mouse spleen signalling responses to anthrax using label-free quantitative phosphoproteomics via mass spectrometry. *Mol. Cell. Proteomics* 10, M110.000927.
26. Scott, N.E. *et al.* (2011) Simultaneous glycan-peptide characterization using hydrophilic interaction chromatography and parallel fragmentation by CID, higher energy collisional dissociation, and electron transfer dissociation MS applied to the *N*-linked glycoproteome of *Campylobacter jejuni*. *Mol. Cell. Proteomics* 10, M000031–MCP201.
27. Kleifeld, O. *et al*. (2010) Isotopic labelling of terminal amines in complex samples identifies protein *N*-termini and protease cleavage products. *Nature Biotechnol.* 28, 281–288.
28. Schilling, O. *et al.* (2010) Proteome-wide analysis of protein carboxy termini: *C*-terminomics. *Nature Meth.* 7, 508–511.

Biography

Stuart Cordwell is an Associate Professor in the School of Molecular Bioscience and the Discipline of Pathology, School of Medical Science at the University of Sydney. He is a co-Director of the Sydney University Proteome Research Unit (SUPRU). His research interests are in proteins, their post-translational modifications and their role in infections.