Shotgun proteomics of cyanobacteria—applications of experimental and data-mining techniques

Martin E. Barrios-Llerena, Poh Kuan Chong, Chee Sian Gan, Ambrosius P. L. Snijders, Kenneth F. Reardon and Phillip C. Wright

Advance Access publication date 10 May 2006

Abstract

Cyanobacteria are photosynthetic bacteria notable for their ability to produce hydrogen and a variety of interesting secondary metabolites. As a result of the growing number of completed cyanobacterial genome projects, the development of post-genomics analysis for this important group has been accelerating. DNA microarrays and classical two-dimensional gel electrophoresis (2DE) were the first technologies applied in such analyses. In many other systems, 'shotgun' proteomics employing multi-dimensional liquid chromatography and tandem mass spectrometry has proven to be a powerful tool. However, this approach has been relatively under-utilized in cyanobacteria. This study assesses progress in cyanobacterial shotgun proteomics to date, and adds a new perspective by developing a protocol for the shotgun proteomic analysis of the filamentous cyanobacterium *Anabaena variabilis* ATCC 29413, a model for N_2 fixation. Using approaches for enhanced protein extraction, 646 proteins were identified, which is more than double the previous results obtained using 2DE. Notably, the improved extraction method and shotgun approach resulted in a significantly higher representation of basic and hydrophobic proteins. The use of protein bioinformatics tools to further mine these shotgun data is illustrated through the application of PSORTb for localization, the grand average hydropathy (GRAVY) index for hydrophobicity, LipoP for lipoproteins and the exponentially modified protein abundance index (emPAI) for abundance. The results are compared with the most well-studied cyanobacterium, *Synechocystis* sp. PCC 6803. Some general issues in shotgun proteome identification and quantification are then addressed.

Keywords: *Shotgun proteomics; cyanobacteria; membrane; insoluble; abundance;* Anabaena variabilis

INTRODUCTION

Cyanobacteria (blue-green algae) are one of the largest groups of prokaryotes, and are able to carry out plant-like oxygenic photosynthesis using water as an electron donor. They occupy a wide range of illuminated niches in freshwater, marine, terrestrial and hypersaline environments. Cyanobacteria demonstrate morphological diversity ranging from single unicellular genera to complex multicellular filamentous strains, many of which have specialized differentiated cells such as the heterocyst in nitrogen-fixing *Anabaena* and *Nostoc* species [1].

Corresponding author. Philip C. Wright, Biological and Environmental Systems Group, Department of Chemical and Process Engineering, University of Sheffield, Mappin Street, Sheffield S1 3JD, UK. Tel: 44-114-222-7577, Fax: 44-114-222-7501, E-mail: P.C.Wright@Sheffield.ac.uk

Martin Barrios-Llerena is a graduate student at The University of Sheffield.

Poh Kuan Chong is a graduate student at The University of Sheffield.

Chee Sian Gan is a graduate student at The University of Sheffield.

Ambrosius 'Bram' Snijders is a graduate student at The University of Sheffield.

Kenneth F. Reardon is a professor of chemical and biological engineering at Colorado State University, USA. He and his research group are involved in projects that use proteomics as a tool to solve problems in bioprocessing, environmental biotechnology and biomedical engineering.

Phillip Wright is a professor of systems biology in the Department of Chemical and Process Engineering, The University of Sheffield, UK. His laboratory develops generic tools to aid the understanding of biological systems for chemical, environmental and medical applications.

The ability of these organisms to split water into hydrogen and oxygen makes them attractive candidates for hydrogen production [2–4], and the responsible enzymes (nitrogenases and hydrogenases) have been the focus of many genetic studies [5–7]. The adaptive character of these organisms to different, often adverse, environments also makes them interesting potential sources of therapeutic secondary metabolites [8]. Cyanobacterial genome sequences have been available since 1996 [9], and holistic, or systems biology, studies of these organisms are underway [10].

In this article, we present an assessment of the current state of the cyanobacterial proteomics field. In particular, we examine the main proteomics approaches and discuss the strides being made in shotgun proteomics in particular. This is achieved by comparing data in the literature from cyanobacterial shotgun proteomics with data generated in this study on the filamentous N_2-fixing cyanobacterium *Anabaena variabilis* ATCC 29413. After presenting the background context and some comments on the challenges of cyanobacterial proteomics, we present the results and discussion of our new data, and end with a future perspective. Although the focus here is on cyanobacteria, some more general shotgun proteomics issues are discussed.

TRADITIONAL CYANOBACTERIAL PROTEOMICS

The analysis of cyanobacterial proteins has been traditionally undertaken on fractions of the whole proteome [11, 12], primarily employing electrophoresis-based approaches. Techniques to obtain these fractions include aqueous polymer two-phase partitioning for plasma membrane isolation [11]. Studies have particularly focused on the inner membrane and the thylakoid [13, 14] subproteomes, where the photosystems are found together with the phycobilisome pigments involved in the channelling of solar energy.

The most well-studied cyanobacterium is *Synechocystis* sp. PCC 6803, on which a number of gel-based studies have been carried out, providing data on proteins in the periplasm (57 proteins identified) [13], plasma membrane (57 proteins identified) [14], thylakoid membrane (78 proteins identified) [15] and outer membrane (29 proteins identified) [16]. The Cyano2Dbase [17, 18] database for this organism lists 234 proteins from different fractions (soluble, insoluble, thylakoid membrane and secreted). The soluble subproteome of this organism was resolved using two-dimensional electrophoresis (2DE), with 74 unique proteins identified [19]. Finally, studies have been carried out of proteome changes under different salt concentrations in this strain [20] and under heterotrophic conditions [21].

SHOTGUN PROTEOMICS

Strategies for separation and identification

The shotgun proteomics approach was developed in an attempt to increase throughput and to access proteins that are difficult to resolve using gel electrophoresis [22–24]. Shotgun proteomics uses multi-dimensional separations [usually via high-performance liquid chromatography (HPLC)] of the complex mixture of peptides, generated from the treatment of a protein mixture derived from whole cell lysis with a protease such as trypsin. The rationale is that since the properties of peptides are more forgiving than proteins, standardized separation protocols can be brought to bear for proteome-wide measurements [25]. The complexity of the peptide mixture is typically simplified first using a strong cation exchange (SCX) resin that, once loaded with peptides, is sequentially eluted with salt solutions of increasing strength. Fractions of this eluent are then submitted for multiple reverse-phase liquid chromatography–tandem mass spectrometry (RP-LC–MS/MS) analysis [23]. Although shotgun proteomics was originally based on orthogonal chromatography, recently the field has expanded by also employing some combinations of gel-based, liquid-phase isoelectric focusing (IEF) and LC separation of proteins and peptides. This diversity of workflows has been demonstrated on proteins in human serum [26], in archaea [27] and in the cyanobacterium *Synechocystis* sp. PCC 6803 [28]. Thus, the definition of shotgun proteomics can be widened to include not only multi-dimensional LC separation, but also these alternative separations such as gel-based separation of peptides or protein separation in ion-exchange columns or IEF gels, all prior to loading the tryptic peptides into a RP-LC–MS/MS.

There is still much debate as to which might be the best method for shotgun proteomics. An attempt to address this issue was made in the cyanobacterium *Synechocystis* sp. PCC6803 [28], where different protein and peptide separation workflows were

examined. The best workflow combination for this particular organism employed 1D SDS-PAGE protein separation followed by gel-based IEF and RP-LC peptide fractionations. This result was confirmed more broadly in the proteomic study of the crenarchaeon *Sulfolobus solfataricus* P2 [27]. An important observation regarding all these shotgun approaches is that each workflow identified a set of unique proteins (peptides). Therefore, for comprehensive identification of any proteome, the best results would be achieved using different fractionation approaches (pre- and post-digestion) in parallel.

Quantification

The majority of shotgun proteomics studies reported to date have focused on coverage and identification. However, a successful proteomics experiment requires both protein identification and quantification [29], preferably on a global scale. In classical proteomics, 2DE is used for protein separation and quantification via staining intensity, with identification relying on mass spectrometry (MS). In shotgun proteomics workflows, gel-based separations are usually avoided altogether [23]. Instead, digested peptide mixtures are directly analysed by MS, often after some orthogonal pre-fractionation. Quantification in these workflows requires some type of labelling, with two fundamentally different approaches in use: *in vivo* and *in vitro* labelling. *In vitro* methods make use of chemical or enzymatic approaches to specifically derivatize functional groups. Examples of these methods include isotope-coded affinity tag (ICAT) [30] and ^{18}O labelling [31]. The mass shift observed in these approaches is usually more or less fixed. *In vivo* methods rely on the ability of the organism to biosynthetically incorporate isotope-labelled precursors [32]. In a popular approach called SILAC, labelled amino acid are supplied to cell cultures [33]. Supplying the cell culture with both labelled arginine and lysine ensures that all tryptic peptides are labelled and the mass shift is fixed. Complete labelling with ^{15}N or ^{13}C is a popular method for labelling of unicellular organisms, and has recently been extended to multicellular organisms [34–36].

In a quantitative shotgun proteomics experiment, the sample complexity is doubled with respect to MS since each peptide now appears twice in the mass spectrum, although the biological complexity remains the same. A disadvantage is that the intensity of each peptide is reduced because it appears now as two separate peaks. In addition, when MS/MS spectra are acquired for both isotopomers, the number of identifications is approximately halved. However, the increase in complexity is not always a disadvantage, as Snijders *et al*. [35] recently demonstrated. In that report, the increased complexity of MS spectra was exploited, and it was shown that MS/MS spectra obtained from uniform ^{15}N- and ^{13}C-labelled peptides contain additional information that can be used to reduce the number of false-negative peptide identifications [35].

An approach for quantitative protein analysis developed by Ross *et al*. [37], called iTRAQ, surmounts this problem of complexity by making use of isobaric mass labels at the N-termini and lysine side chains of peptides in a digest mixture. One of the main advantages is that these labels do not increase the complexity of the MS spectrum, and the peak intensity of each peptide is not reduced because peptides labelled with different isobaric tags have the same precursor mass. An additional advantage is that quantification occurs at the MS/MS stage, and the reporter ions are quantified more accurately, since lower levels of chemical and instrument generated noise are present in the MS/MS scan.

These recent advances, and ongoing efforts, will aid development of quantitative shotgun proteomics for the measurement of expression levels of proteins in cyanobacteria, and will lead to a greater understanding of cellular processes in these organisms.

EXPERIMENTAL AND DATA ANALYSIS CHALLENGES IN SHOTGUN PROTEOMICS

One of the main challenges for the success of a shotgun separation of cyanobacterial (and many other) proteins is their solubilization into a buffer compatible with downstream processing. In general, the solubilization of membrane proteins is a major concern in the proteomics community [38]. Recently, some alternative procedures have been implemented for membrane protein analysis using proteases with broad specificity such as proteinase K [39]. The use of non-ionic and zwitterionic detergents was investigated also in the solubilization of membrane proteins from red blood cells and *Arabidopsis thaliana* [40], using those extracts in conjunction with SDS-PAGE or 2DE. Other approaches used direct solubilization of membrane proteins into organic solvents coupled with a RP-LC

and MS analysis [41, 42]. This last procedure offers the opportunity for high throughput analysis.

As mentioned above, the use of zwitterionic and non-ionic detergents should improve the solubilization of membrane proteins from cyanobacteria. In this article, we report results from protein extraction with a mixture of two of these detergents: ASB-14, a general detergent used in the solubilization of membrane proteins, and dodecyl-β-d-maltoside (DM), used mainly in the solubilization of thylakoid membranes and cytoplasmic membrane proteins [43]. The use of these detergents, together with a shotgun approach, was assessed for the proteome analysis of the cyanobacterium *A. variabilis* ATCC 29413. The results are benchmarked against 2DE experiments of the same phenotype and against the shotgun proteome of another cyanobacterium, *Synechocystis* sp. PCC6803.

A number of challenges arise in the context of data analysis. In common with most experiments in proteomics, the application of shotgun proteomics presents issues with regard to identification and function. Despite the pace of genome sequencing efforts, the published sequences of cyanobacteria, like many other organisms, still have numerous hypothetical and/or poorly characterized proteins within their proteome.

Since there have been relatively few large-scale cyanobacterial shotgun data sets until recently, there has been little opportunity to examine in detail the types of proteins found. There is now such an opportunity, both to provide a deeper understanding of biological function as well as to guide researchers in designing better experiments to enrich lower-abundance proteins. This is true generally in shotgun proteomics outside a few key organisms.

MATERIALS AND METHODS

Strain and culture conditions

Axenic cultures of the filamentous N_2-fixing cyanobacterium *A. variabilis* ATCC 29413 were obtained from the American Type Culture Collection (Rockville, MD, USA), and grown in 250 ml Erlenmeyer flasks at 25°C in BG-11 medium under a 12:12 h light–dark cycle. Illumination was provided by fluorescent plant growth lamps (Gro-Lux lamps, Osram Sylvania, Danvers, MA, USA), with a photosynthetically active radiation of 50 μ einstein $m^{-2} s^{-1}$. Cells were harvested in the late-exponential phase by centrifugation at 10 000 *g* for 5 min at 4°C, and the pellets were collected for immediate protein extraction.

Protein extraction

Cell pellets were resuspended in a lysis buffer containing 7 M urea, 2 M thiourea, 4% CHAPS, 2% ASB-14, 1% DM, 20% glycerol, 200 mM KCl, 100 mM dibasic sodium phosphate pH 7.6 and 1 mM PMSF. The cells were disrupted by grinding in liquid nitrogen (three times). Cell debris and insoluble protein particulates were removed by centrifugation at 21 000 *g* for 30 min. The total protein concentration was measured using the Bio-Rad RC DC protein assay kit (Bio-Rad, Hercules, CA, USA) according to the manufacturer's protocol. Protein extracts were stored at −20°C until required.

Tryptic digestion and peptide extraction

Total protein of 2 mg (in ∼250 μl) were dissolved in an equal volume of digestion solution containing 8 M urea, 200 mM ammonium bicarbonate and 20 mM methylamine and mixed by vortexing for ca. 10 s. The proteins were then reduced with 20 mM tris(hydroxypropyl)phosphine (THP) at 50°C for 15 min and alkylated with 40 mM iodoacetamide in the dark at room temperature for 15 min. Endoprotease Lys-C (Roche Diagnostics Ltd., East Sussex, UK) was added to a final substrate-to-enzyme ratio of 100:1 (w/w) as previously described by Link *et al.* [44], and the reaction was incubated at 37°C for 15 h. The Lys-C digestion was diluted 4-fold with 100 mM ammonium bicarbonate, and modified trypsin (Roche Diagnostics Ltd, East Sussex, UK) was added to a final substrate-to-enzyme ratio of 50:1 (w/w). The trypsin digestion was incubated at 37°C for 15 h. A solid-phase extraction with SupelCleanC18 cartridges (Sigma-Aldrich, St Louis, MO, USA) was carried out on the complex peptide mixture according to the manufacturer's instructions. These peptide samples were lyophilized to dryness in a vacuum concentrator (Eppendorf Concentrator 5301, Eppendorf UK Ltd, Cambridge, UK) and then stored at −20°C until used.

Liquid chromatography

Peptides were re-dissolved in SCX buffer A (10 mM K_2HPO_4, 25% acetonitrile, pH 3), sonicated and centrifuged (5 min at 10 000 *g*). This solution was then loaded onto a 2.6 × 200 mm, 5 μm, 300 Å PolySulfoethyl A column (Poly LC Inc., Columbia,

MD, USA) connected to a BioLC HPLC system (Dionex-LC Packings, Hercules, CA, USA). Peptides were eluted with a linear gradient of 0–25% buffer B in 40 min and 25–100% buffer B in 20 min (10 mM K_2HPO_4, 25% acetonitrile, 500 mM KCl, pH 3) at a flow rate of 0.2 ml/min, with fractions collected at 1 min intervals. All collected fractions were lyophilized to dryness by vacuum centrifugation, and then redissolved in 3% acetonitrile solution (0.1% formic acid) prior to their analysis by MS/MS.

RP-HPLC-ESI-MS/MS analysis

Nano-ESI–MS/MS analyses were performed on a QSTAR XL mass spectrometer (Applied Biosystems-MDS Sciex, Foster City, CA, USA; MDS-Sciex, Concord, Ontario, Canada). Mass spectra were acquired by scanning a 300–1700 *m/z* range. Fractions were injected into the mass spectrometer using a Famos® auto-sampler (LC Packings, Dionex, Amsterdam, The Netherlands) and an Ultimate® pump (LC Packings) through a fused-silica capillary at 300 nl/min. Samples (20 µl) were applied at 30 µl/min to a 0.3 × 5 mm trap column (m-Precolumn Cartridge, PepMap C18, 5 µm, 100 Å; LC Packing, Amsterdam, The Netherlands) using a mobile phase A, containing 0.1% formic acid and 3% acetonitrile. After application and washing of the sample for 10 min, the trap column switched online to a 0.075 × 150 mm reversed-phase capillary column (C18 PepMap100, 3 µm, 100 Å, 75 µm i.d. × 15 cm). Peptides were eluted using a mobile phase B, containing 97% acetonitrile, and 0.1% formic acid. The linear gradient was 5–45% B over 76 min. Electrospray fused silica PicoTip™ needles were obtained from New Objective (Woburn, MA), and the spray voltage was set at 5.5 kV.

Data analysis and interpretation

Protein identification from the shotgun workflow was performed using ProID 1.1 (Applied Biosystems, MDS Sciex) against the NCBI protein non-redundant database (June, 2005) using a mass tolerance of 0.15 Da for the precursor mass and 0.1 Da for the fragment masses. Search parameters allowed for one missed cleavage of trypsin. Carbamidomethylation was selected as a fixed modification and oxidation of methionine was allowed to be variable. A list of proteins identified was generated using ProGroup Viewer v. 1.0.5 (Applied Biosystems, MDS Sciex). For single- and multi-peptide matches to a protein, a protein identification probability cut-off of 95% in ProGroup Viewer was used to filter matches.

Physical and chemical parameters, including the grand average hydropathy (GRAVY) index [45], were calculated using ProtParam [46] at http://us.expasy.org/tools/protparam.html. Species exhibiting positive GRAVY indices are considered hydrophobic, and those with negative indices are deemed hydrophilic. GRAVY values greater than +0.3 were used as an indication of hydrophobic proteins.

Prediction of protein localization was carried out by the program PSORTb v2.0 [47] (http://psort.org). PSORTb, in general, is an accurate and reliable program for predicting the location of proteins. This tool is able to categorize analysed proteins into one of the following localization sites: cytoplasm, cytoplasmic membrane (inner membrane), periplasm, outer membrane and extracellular. In some cases, the location is designated as unknown. Unfortunately, PSORTb does not detect lipoproteins, which may represent an essential class of membrane proteins. For the detection of these proteins, the program LipoP was used [48]. This program identifies lipoproteins by making use of the signal sequence cleaved by a lipoprotein signal peptidase (also called signal peptidase I and II). The GRAVY index [49] was used as an indicator of integral membrane proteins, together with the prediction by PSORTb.

The abundance of each protein was estimated by calculating the protein abundance index (PAI) [50] and the exponentially modified protein abundance index (emPAI) [51]. PAI was defined as the number of observable unique peptides per protein normalized by the number of peptides obtained via *in silico* digestion [50]. The emPAI scale was given by $10^{PAI} - 1$, an exponential relationship with PAI [51]. Proteogest software, which employs a Perl script, was used to generate lists of *in silico* digested peptides [52]. This bioinformatics tool was downloaded from the internet (http://www.utoronto.ca/emililab/program/proteogest.htm).

RESULTS AND DISCUSSION

Numbers and categories of proteins identified in the shotgun proteomics workflow

The whole-cell lysate from *A. variabilis* was digested with trypsin, and an off-line SCX separation of the

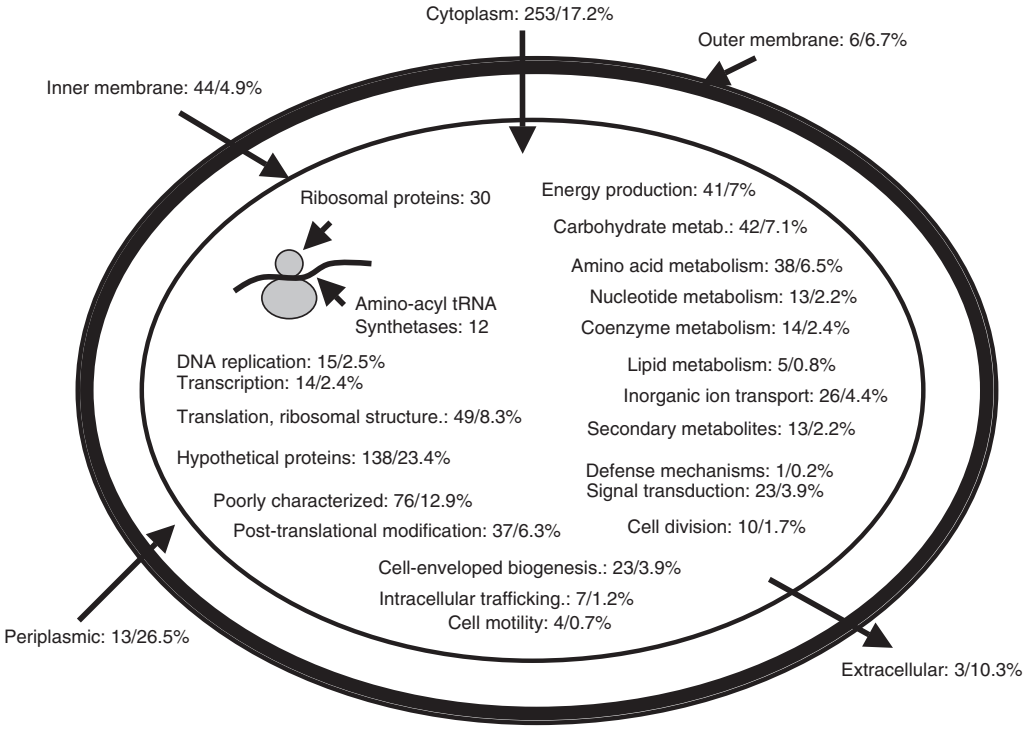

Figure 1: The *A. variabilis* proteome: a functional classification. A virtual cell in which the major functional processes are indicated. Numbers: *left side*, number of proteins identified; *right side*, percentage of the total number of proteins identified. The schematic also shows the protein location as predicted by PSORTb.

resulting peptides was performed, followed by online RP-LC ESI–MS/MS. The 66 SCX fractions generated 45 072 MS/MS spectra, with a total of 646 proteins confidently identified (see Table I, Supplementary Material). Of these 646 proteins, 341 proteins were identified by more than two peptides (with a confidence of 99%). Thus, the coverage of the *A. variabilis* proteome was 13%, with 48% of these proteins being identified by a single peptide. The fraction of proteins identified with two or more peptides can be improved through the use of longer MS running time (RP-HPLC gradients) and the amount of sample loaded.

A classification of these proteins according to major functional categories is represented in a virtual cell shown in Figure 1. Of these identified proteins, 33% are involved in cellular metabolism (energy production and conversion, carbohydrate, amino acid and nucleotide transport and metabolism, etc.), including the phycobilisome apparatus. The remaining proteins were distributed within cellular processes (19%, including cell division, cell enveloped biogenesis, post-translational modification), and information storage and processing (13%). Poorly characterized proteins (12%) and hypothetical proteins (23%) represent a significant fraction of the identified proteins, consistent with the incomplete status of the sequencing and annotation effort for *A. variabilis*. The classification of proteins into functional groups from proteins identified by two or more peptides shows no significant difference to the categorization obtained from proteins identified by a single peptide (data not shown). This observation is in line with that previously described by Aggarwal *et al*. [53].

Protein extraction and identification in cyanobacteria has proven to be a difficult process, as seen in the case of *Synechocystis* sp. PCC 6803 [28]. In that study, a typical single workflow shotgun approach led to the identification of 391 proteins from 40 SCX fractions (221 with more than two peptides) [28]. As can be seen in Table 1, an improvement in a single workflow has been made over the *Synechocystis* study by improving the process of protein digestion using zwitterionic and non-ionic detergents. Further details and an assessment by the protein's physical and chemical properties and abundance between these two cyanobacteria are presented in the following sub-sections.

Table 1: Comparison between *Anabaena variabilis* ATCC 29413 and *Synechocystis* sp. PCC 6803 of the number and percentage of proteins identified by location (against the theoretical number of proteins per location) using the shotgun workflow

Location	*Anabaena variabilis*			*Synechocystis* sp.		
	Observed	Theoretical[a]	%	Observed	Theoretical[a]	%
Cytoplasmic	253	1467	17.2	194	1014	19.1
Cytoplasmic Membrane	44	900	4.9	10	540	1.9
Extracellular	3	29	10.3	0	8	0.0
Outer Membrane	6	89	6.7	1	40	2.5
Periplasmic	13	49	26.5	11	44	25.0
Unknown	327	2505	13.0	175	1521	11.5
	646	5039	12.9	391	3167	12.3
GRAVY[b] $\geq +0.3$	23[c]	+1.104[d]		5[c]	+0.474[d]	
Lipoproteins[e]	29[f]	12[f]		8[f]	1[f]	
pI[g]	3.72	12.02		3.44	11.81	
MW (Da)[g]	4385.1	171314		7131.1	170087	

[a]Prediction of the protein location in the entire proteome was obtained from pre-computed results generated using PSORTb at http://psort.nibb.ac.jp/psortb/.
[b]Grand average hydropathy (GRAVY) values calculating using the ProtParam tool at http://us.expasy.org/tools/protparam.html.
[c]Number of proteins with GRAVY values higher than +0.3.
[d]The highest GRAVY values found in each strain.
[e]Lipoprotein signals were obtain using the LipoP tool at www.cbs.dtu.dk/services/LipoP/.
[f]Number of proteins with signal peptidase I or II, respectively.
[g]The pI and MW range in each strain.

Comparison of protein properties

Since a touted advantage of shotgun proteomics over 2DE-based proteomics is the ability to identify a broader range of proteins, we compared the results obtained here against a 2DE investigation of *A. variabilis* (Barrios-Llerena ME, Reardon KF, Wright PC. Proteomic analysis of *A. variabilis* ATCC 29413 and assignment of putative function through protein sequence homology, Submitted), with the results depicted in Figure 2. Both workflows used the same extraction protocol, and while the proteins in the 2DE study were acetone precipitated prior to the isoelectric focusing, the main difference between the two is the separation method. Table 2 shows the location of proteins identified by the two workflows as predicted by the PSORTb algorithm, along with an indication of the protein hydrophobicities determined from the GRAVY index. This comparison shows that the shotgun workflow yielded an improvement in the number of basic proteins and smaller proteins identified in comparison with the 2DE workflow. Thirty proteins have a pI value greater than 10 in the shotgun workflow, compared with only eight in the 2DE workflow. Taking as a reference a 40 kDa cut-off for the molecular weight, the number of proteins below this value is three times higher in the shotgun workflow compared with the 2DE workflow (497 and 159 proteins, respectively).

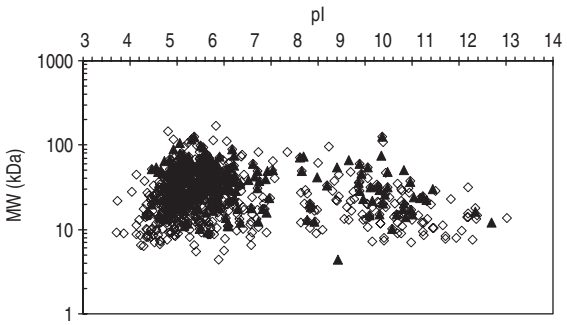

Figure 2: Comparison of the proteins identified in the shotgun (◇) and 2-DE (▲) workflows in *Anabaena variabilis* ATCC 29413. For the shotgun analysis, 2 mg of protein was digested with trypsin, and for the 2-DE analysis, 500 μg of protein was loaded into 3–10NL IPG strip and run in 15% SDS-PAGE.

Furthermore, the number of cytoplasmic membrane proteins increased from 9 to 44 proteins through the use of the shotgun workflow, and a similar result was obtained for the periplasmic proteins (13 proteins compared with 5).

Table 2: Comparison between shotgun and 2DE workflows of the number of proteins, and relative percentages, against the number of proteins identified and theoretically predicted in *Anabaena variabilis* ATCC29413

Location[a]	Theoretical[a]	Shotgun			2DE		
		Observed	% Obs[i]	% Theo[j]	Observed	% Obs[i]	% Theo[j]
Cytoplasmic	1467	253	39.2	17.2	116	45.7	7.9
Cytoplasmic membrane	900	44	6.8	4.9	9	3.5	1.0
Extracellular	29	3	0.5	10.3	1	0.4	3.4
Outer membrane	89	6	0.9	6.7	5	2.0	5.6
Periplasmic	49	13	2.0	26.5	5	2.0	10.2
Unknown	2505	327	50.6	13.2	118	46.4	4.7
	5039	**646**		**12.9**	**254**		**5.0**
GRAVY[b] $\geq +0.3$		23[c]		$+1.104$[d]	1[c]		$+0.441$[d]
Lipoproteins[e]		29[f]		12[f]	10[f]		3[f]
pI[g]		3.72		12.02	4.34		11.70
MW (Da)[h]		4385.1		171314	4432.0		127030

[a]Prediction of the protein location in the entire proteome was obtained from pre-computed results generated using PSORTb at http://psort.nibb.ac.jp/psortb/.
[b]Grand average hydropathy (GRAVY) values were calculated using the ProtParam tool at http://us.expasy.org/tools/protparam.html.
[c]Number of proteins with GRAVY values higher than +0.3.
[d]The highest GRAVY values found in each workflow.
[e]Lipoprotein signals were obtained using the LipoP tool at www.cbs.dtu.dk/services/LipoP/.
[f]Number of proteins with signal peptidases I and II, respectively.
[g]The pI and MW range in each strain.
[h]The pI and MW range in each workflow.
[i]Percentage observed—the percentage of the number of proteins identified in that predicted location against the total numbers identified.
[j]Percentage theoretical—the number of proteins identified in that location versus the number theoretically expected in that location.

Interestingly, comparison of the outer membrane proteins identified using these two techniques shows that only two proteins (out of five for 2DE and six for shotgun) were common between the two workflows. This low level of agreement may be due to the different separation methods used in the two workflows, and is also influenced by the low overall recovery of outer membrane (OM) proteins. The number of membrane proteins for the shotgun and 2DE workflows was further increased with the identification of 41 and 13 predicted lipoproteins using LipoP, respectively. The shotgun workflow yielded 29 proteins that present signal peptidase I, and 12 that include signal peptidase II (Table I, Supplementary Material), while the 2DE workflow resulted in 10 and three proteins with signal peptidases I and II, respectively (data not shown). All of these putative lipoproteins were designated as 'unknown' by PSORTb. This increased detection of hydrophobic proteins is reflected in the number of proteins with GRAVY values higher than +0.3 in the shotgun (23 proteins) compared with only one in the 2DE workflow.

It is of interest to compare the results from shotgun workflows for proteins from *A. variabilis* and *Synechocystis* sp. PCC6803 (Figure 3 and Table 1). In general, the distribution of proteins is similar for both strains (Figure 3), with the largest number of proteins having pI values between 4 and 7. The range of molecular weights for both strains is similar in the upper range (170 kDa). Ten proteins with molecular weights below 7.1 kDa were found in *A. variabilis* (down to 4 kDa); notably, three of these proteins were subunits of photosystems I and II (Psb-Y and -H and PSI-J).

The main distribution of the proteins found for these two strains using the shotgun workflow is shown in Table 1. The proportion of cytoplasmic proteins is similar in *A. variabilis* and *Synechocystis* (17 and 19%, respectively). However, a major difference in the number of cytoplasmic membrane proteins that can be observed (44 proteins in *A. variabilis* compared with 9 proteins in *Synechocystis*). As in the case of the shotgun workflow in *A. variabilis*, prediction of lipoproteins in *Synechocystis* resulted in nine predicted proteins, of which eight present with signal peptidase I and one with signal peptidase II. This prediction reveals a 4-fold increase in membrane proteins found in *A. variabilis* compared with *Synechocystis*, and may be attributed to the

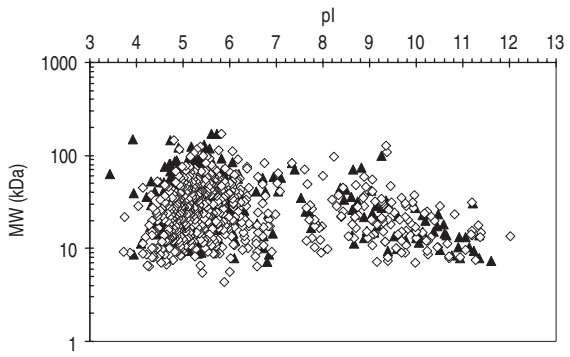

Figure 3: Distribution of proteins found identified in the shotgun workflow from *Anabaena variabilis* ATCC 29413 (◇) and *Synechocystis* sp. PCC 6803 (▲).

presence of ASB-14 and β-DM in the extraction buffer. The GRAVY indices reflect the same pattern towards hydrophobic proteins in *A. variabilis* (23 proteins with GRAVY values higher than +0.3) compared with five in *Synechocystis*.

Although relatively few outer membrane proteins were identified in this study, the detergents used in our extraction buffer have previously proven effective for hydrophobic proteins. For example, Norling's group used these detergents in the extraction of outer membrane proteins [16]. Protein extracts were concentrated and fractionated using a sucrose gradient and polymer two-phase separation. The resulting outer membrane protein fraction was then resolved with 2DE and 29 proteins were identified. The fractionation step employed by these researchers may be important in the detection of the relatively low-abundance outer membrane proteins.

Comparison of protein abundance

The protein abundance in these two strains was analysed using their emPAI [51]. Proteins with a high emPAI value are highly abundant and more likely to be identified using MS. Table 3 shows the emPAI values for *A. variabilis* (shotgun and 2DE workflow) and *Synechocystis* (shotgun workflow). A comparison of the shotgun and 2DE gel workflow emPAI values for *A. variabilis* reveals that the most abundant proteins are the phycobilisome proteins, with an emPAI value of 7.577 for the shotgun and 5.494 for the 2DE workflow. Both proteins are from the phycobilisome assembly.

Analysis of the shotgun data for the two cyanobacteria shows that 18 and 5 proteins have emPAI values higher than 3.0 in *A. variabilis* and *Synechocystis*, respectively. In both cases, the most abundant proteins were phycobilisome associated. However, in the case of *A. variabilis*, 10 other abundant proteins were observed, such as 'COG0783: DNA-binding ferritin-like protein' (inorganic ion transport protein), and 'COG0526: Thiol-disulfide isomerase and thioredoxins' (TCA cycle, energy metabolism). In *A. variabilis*, the protein with the highest emPAI value was an ATPase involved in DNA repair, with a value of 7.73, while the most abundant protein in *Synechocystis* was the allophycocyanin β-subunit with a value of 7.37.

From these results, it is apparent that abundance of phycobilisome proteins in cyanobacterial extracts is a major concern, as those proteins can interfere with the detection of the low-abundance proteins. In the shotgun workflow for *A. variabilis*, the percentage of abundant proteins (emPAI value higher than 3.0) is 2-fold higher than in *Synechocystis*, but this does not reduce the number of low-abundance proteins (emPAI value between 0 and 1.0) identified in this strain, as seen in Table 3.

CONCLUSION

Application of proteomics to cyanobacteria has mainly focussed on 2DE-based separations, with only one study in the cyanobacterium *Synechocystis* sp. PCC 6803 [28] showing an intensive use of the shotgun approach using multiple combinations of LC- and gel-based separation techniques. The application of an alternative shotgun workflow in *A. variabilis* ATCC 29413 in combination with non-ionic and zwitterionic detergents yielded an improvement in the identification of cytoplasmic membrane proteins, without compromising the identification of other proteins (cytoplasmic, periplasmic, etc.). This is the first time, to our knowledge, that a large-scale shotgun proteomics study has been performed within the genus *Anabaena*. Moreover, the use of four bioinformatic tools in this study showed them to be useful in data interpretation and in the gathering of properties and parameters (such as predicted protein location and abundance) inherent in the cells. At the same time, these tools validated the use of the alternative shotgun workflow in the improvement in identification of cytoplasmic membrane proteins. These tools are not perfect; for example, PSORTb was unable to classify some of these proteins and returned 'unknown' for the location.

Table 3: Number of proteins, and their overall relative percentage, detected for different emPAI values in *Anabaena variabilis* ATCC 29413 (shotgun and 2DE workflows) and *Synechocystis* sp. PCC 6803 (shotgun workflow)

emPAI[a]	Shotgun				2DE	
	A. variabilis	%[b]	Synechocystis	%[b]	A. variabilis	%[b]
0–1	548	84.8	405	87.7	194	76.4
1–2	61	9.4	35	9.0	47	18.5
2–3	19	2.9	8	2.0	5	2.0
≥3	18	2.8	5	1.3	8	3.1
	646		**391**		**254**	

[a]emPAI (exponentially modified protein abundance index) values calculated according to Ishihama *et al*. [51]
[b]percentage of the number of proteins with that emPAI value against the total number of proteins identified for that organism and workflow.

The use of the PAI confirms that the phycobilisome proteins are highly abundant proteins in both shotgun and 2DE workflows. Those proteins did not negatively impact the relative number of low-abundance proteins identified using our shotgun technique. However, in an absolute sense, there are still significant numbers of proteins not resolvable using any of the shotgun methods tested here or in the literature. In order to further improve the proteome coverage in cyanobacteria, a more exhaustive approach is required. For example, the analysis of the insoluble proteins (pellet) from the extract is likely to result in the identification of more membrane peptides. This could be accomplished by dissolving the pellet in organic solvents, as reported by Moslavac *et al*. [54] in *Anabaena* PCC 7120. The selective removal of the highly abundant proteins will be critical to access additional low-abundance proteins. Multiple injections of each sample into the MS can also further improve proteome coverage.

Another opportunity to increase proteome coverage lies in the observation of unmatched peptides, which occurs frequently, even with high-quality mass spectra. One cause of this may be poorly annotated genomes. An alternative strategy is to use the translation of the unannotated genome sequence into an amino acid sequence, encoding putative proteins in all six reading frames, and to identify unmatched peptides by MS/MS. This strategy has proven to result in more robust protein identifications, as demonstrated by Smith *et al*. [55] in the study of the unicellular organism *Tetrahymena thermophila*.

Finally, the use of a shotgun proteomics analysis in cyanobacterial strains should grow as the community becomes increasingly more confident. The versatility of multi-dimensional separations for different kinds of samples with extreme properties such as acidity and low aqueous solubility makes them attractive tools for a global proteome analysis. Moreover, with recent improvements in protein quantification techniques such as ICAT, iTRAQ and *in vivo* isotope enrichment, the use of quantitative shotgun approaches is extremely attractive. However, the issue of compatibility between separation techniques and quantification still needs to be addressed further.

Key Points

- The first large-scale shotgun proteomics in the genus *Anabaena* was performed successfully, resulting in the identification of 646 proteins.
- An alternative extraction method, coupled with the shotgun approach, was more effective for the detection of membrane proteins.
- Information regarding the properties and characteristics of the proteome of this organism was gained through the use of bioinformatic tools.
- The abundance of phycobilisome proteins was confirmed using an experimental PAI.
- This study provides the basis for the design of biologically focused experimental studies on this and closely related organisms.

Acknowledgements
We would like to thank the Overseas Research Student's Award Scheme and The University of Sheffield's PhD Scholarship Programme for provision of a PhD scholarship to M.B.L. This work was supported by the UK's Engineering and Physical Science Research Council (EPSRC) (Grants GR/S82800/01 and GR/S84347/01). P.C.W. thanks the EPSRC for an Advanced Research Fellowship (GR/A11311/02).

References

1. Whitton BA. Diversity, ecology, and taxonomy of the cyanobacteria. In: Mann NH, Carr NG (eds). *Photosynthetic Prokaryotes*. New York, Plenum Press, 1992, 1–51.

2. Cournac L, Guedeney G, Peltier G, *et al*. Sustained photoevolution of molecular hydrogen in a mutant of *Synechocystis* sp. strain PCC 6803 deficient in the type I NADPH-dehydrogenase complex. *J Bacteriol* 2004;**186**: 1737–46.
3. Borodin VB, Tsygankov AA, Rao KK, *et al*. Hydrogen production by *Anabaena variabilis* PK84 under simulated outdoor conditions. *Biotechnol Bioeng* 2000;**69**:478–85.
4. Kumazawa S. Photoproduction of hydrogen by the marine heterocystous cyanobacterium *Anabaena* species TU37–1 under a nitrogen atmosphere. *Mar Biotechnol (NY)* 2003;**5**: 222– 6.
5. Serebryakova LT, Medina M, Zorin NA, *et al*. Reversible hydrogenase of *Anabaena variabilis* ATCC 29413: catalytic properties and characterization of redox centres. *FEBS Lett* 1996;**383**:79–82.
6. Tamagnini P, Axelsson R, Lindberg P, *et al*. Hydrogenases and hydrogen metabolism of cyanobacteria. *Microbiol Mol Biol Rev* 2002;**66**:1–20.
7. Boison G, Bothe H, Schmitz O. Transcriptional analysis of hydrogenase genes in the Cyanobacteria *Anacystis nidulans* and *Anabaena variabilis* monitored by RT-PCR. *Curr Microbiol* 2000;**40**:315–21.
8. Burja AM, Banaigs B, Abou-Mansour E, *et al*. Marine cyanobacteria—a prolific source of natural products. *Tetrahedron* 2001;**57**:9347–77.
9. Kaneko T, Sato S, Kotani H, *et al*. Sequence analysis of the genome of the unicellular cyanobacterium s*ynechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res* 1996;**3**:109–36.
10. Burja AM, Dhamwichukorn S, Wright PC. Cyanobacterial postgenomic research and systems biology. *Trends Biotechnol* 2003;**21**:504–11.
11. Norling B, Mirzakhanian V, Nilsson F, *et al*. Subfractional analysis of cyanobacterial membranes and isolation of plasma membranes by aqueous polymer two-phase partitioning. *Anal Biochem* 1994;**218**:103–11.
12. Norling B, Zak E, Andersson B, *et al*. 2D-isolation of pure plasma and thylakoid membranes from the cyanobacterium *Synechocystis* sp. PCC 6803. *FEBS Lett* 1998;**436**:189–92.
13. Fulda S, Huang F, Nilsson F, *et al*. Proteomics of *Synechocystis* sp. strain PCC 6803. Identification of periplasmic proteins in cells grown at low and high salt concentrations. *Eur J Biochem* 2000;**267**:5900–7.
14. Huang F, Parmryd I, Nilsson F, *et al*. Proteomics of *Synechocystis* sp. strain PCC 6803: identification of plasma membrane proteins. *Mol Cell Proteomics* 2002;**1**: 956–66.
15. Wang Y, Sun J, Chitnis PR. Proteomic study of the peripheral proteins from thylakoid membranes of the cyanobacterium *Synechocystis* sp. PCC 6803. *Electrophoresis* 2000;**21**:1746–54.
16. Huang F, Hedman E, Funk C, *et al*. Isolation of outer membrane of *Synechocystis* sp. PCC 6803 and its proteomic characterization. *Mol Cell Proteomics* 2004;**3**:586–95.
17. Sazuka T, Ohara O. Towards a proteome project of cyanobacterium *Synechocystis* sp. strain PCC6803: linking 130 protein spots with their respective genes. *Electrophoresis* 1997;**18**:1252–8.
18. Sazuka T, Yamaguchi M, Ohara O. Cyano2Dbase updated: linkage of 234 protein spots to corresponding genes through N-terminal microsequencing. *Electrophoresis* 1999;**20**: 2160–71.
19. Simon WJ, Hall JJ, Suzuki I, *et al*. Proteomic study of the soluble proteins from the unicellular cyanobacterium *Synechocystis* sp. PCC6803 using automated matrix-assisted laser desorption/ionization-time of flight peptide mass fingerprinting. *Proteomics* 2002;**2**:1735–42.
20. Huang F, Fulda S, Hagemann M, *et al*. Proteomic screening of salt-stress-induced changes in plasma membranes of *Synechocystis* sp. strain PCC 6803. *Proteomics* 2006;**6**:910–20.
21. Kurian D, Jansen T, Maenpaa P. Proteomic analysis of heterotrophy in *Synechocystis* sp. PCC 6803. *Proteomics* 2006;DOI 10.1002/pmic.200500413.
22. Wolters DA, Washburn MP, Yates JR, 3rd. An automated multidimensional protein identification technology for shotgun proteomics. *Anal Chem* 2001;**73**:5683–90.
23. Washburn MP, Wolters D, Yates JR. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat Biotechnol* 2001;**19**:242–7.
24. Peng J, Elias JE, Thoreen CC, *et al*. Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. *J Proteome Res* 2003;**2**:43–50.
25. Whitelegge JP. Mass spectrometry for high throughput quantitative proteomics in plant research: lessons from thylakoid membranes. *Plant Physiol Biochem* 2004;**42**: 919–27.
26. Barnea E, Sorkin R, Ziv T, *et al*. Evaluation of prefractionation methods as a preparatory step for multidimensional based chromatography of serum proteins. *Proteomics* 2005;**5**:3367–75.
27. Chong PK, Wright PC. Identification and characterization of the *Sulfolobus solfataricus* P2 proteome. *J Proteome Res* 2005; **4**:1789–98.
28. Gan CS, Reardon KF, Wright PC. Comparison of protein and peptide prefractionation methods for the shotgun proteomic analysis of *Synechocystis* sp. PCC 6803. *Proteomics* 2005;**5**:2468–78.
29. Aebersold R, Mann M. Mass spectrometry-based proteomics. *Nature* 2003;**422**:198–207.
30. Gygi SP, Rist B, Gerber SA, *et al*. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotechnol* 1999;**17**:994–9.
31. Yao X, Freas A, Ramirez J, *et al*. Proteolytic ^{18}O labeling for comparative proteomics: model studies with two serotypes of adenovirus. *Anal Chem* 2001;**73**:2836–42.
32. Beynon RJ, Pratt JM. Metabolic labeling of proteins for proteomics. *Mol Cell Proteomics* 2005;**4**:857–72.
33. Ong SE, Blagoev B, Kratchmarova I, *et al*. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol Cell Proteomics* 2002;**1**:376–86.
34. Krijgsveld J, Ketting RF, Mahmoudi T, *et al*. Metabolic labeling of *C. elegans* and *D. melanogaster* for quantitative proteomics. *Nat Biotechnol* 2003;**21**:927–31.
35. Snijders AP, de Vos MG, Wright PC. Novel approach for peptide quantitation and sequencing based on ^{15}N and ^{13}C metabolic labeling. *J Proteome Res.* 2005;**4**:578– 85.
36. Oda Y, Huang K, Cross FR, *et al*. Accurate quantitation of protein expression and site-specific phosphorylation. *Proc Natl Acad Sci USA* 1999;**96**:6591–6.

37. Ross PL, Huang YN, Marchese JN, *et al*. Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents. *Mol Cell Proteomics* 2004;**3**: 1154–69.

38. Santoni V, Molloy M, Rabilloud T. *Membrane proteins and proteomics: un amour impossible? Electrophoresis* 2000;**21**:1054–70.

39. Rabilloud T. Membrane proteins ride shotgun. *Nat Biotechnol* 2003;**21**:508–10.

40. Arie JP, Sassoon N, Betton JM. Chaperone function of FkpA, a heat shock prolyl isomerase, in the periplasm of *Escherichia coli*. *Mol Microbiol* 2001;**39**:199–210.

41. Blonder J, Goshe MB, Moore RJ, *et al*. Enrichment of integral membrane proteins for proteomic analysis using liquid chromatography-tandem mass spectrometry. *J Proteome Res* 2002;**1**:351–60.

42. Blonder J, Goshe MB, Xiao W, *et al*. Global analysis of the membrane subproteome of *Pseudomonas aeruginosa* using liquid chromatography-tandem mass spectrometry. *J Proteome Res* 2004;**3**:434–44.

43. Walcher W, Timperio AM, Zolla L, *et al*. Characterization of a variant of the spinach PSII type I light-harvesting protein using kinetically controlled digestion and RP-HPLC-ESI-MS. *Anal Chem* 2003;**75**:6775–80.

44. Link AJ, Eng J, Schieltz DM, *et al*. Direct analysis of protein complexes using mass spectrometry. *Nat Biotechnol* 1999;**17**: 676–82.

45. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. *J Mol Biol* 1982;**157**:105–32.

46. Gasteiger E, Hoogland C, Gattiker A. Protein identification and analysis tools on the ExPASy server. In: Walker JM (ed). *The Proteomics Protocols Handbook*. Hatfield, Herts, UK, Humana Press, 2005, 571–607.

47. Gardy JL, Laird MR, Chen F, *et al*. PSORTb v.2.0: Expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. *Bioinformatics* 2005;**21**:617–23.

48. Juncker AS, Willenbrock H, von Heijne G, *et al*. Prediction of lipoprotein signal peptides in Gram-negative bacteria. *Protein Sci* 2003;**12**:1652–62.

49. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. *J Mol Biol* 1982;**157**: 105–32.

50. Rappsilber J, Ryder U, Lamond AI, *et al*. Large-scale proteomic analysis of the human spliceosome. *Genome Res* 2002;**12**:1231–45.

51. Ishihama Y, Oda Y, Tabata T, *et al*. Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. *Mol Cell Proteomics* 2005;**4**: 1265–72.

52. Cagney G, Amiri S, Premawaradena T, *et al*. *In silico* proteome analysis to facilitate proteomics experiments using mass spectrometry. *Proteome Sci* 2003;**1**:5.

53. Aggarwal K, Choe LH, Lee KH. Quantitative analysis of protein expression using amine-specific isobaric tags in *Escherichia coli* cells expressing rhsA elements. *Proteomics* 2005;**5**:2297–308.

54. Moslavac S, Bredemeier R, Mirus O, *et al*. Proteomic analysis of the outer membrane of *Anabaena* sp. strain PCC 7120. *J Proteome Res* 2005;**4**:1330–8.

55. Smith JC, Northey JGB, Garg J, *et al*. Robust method for proteome analysis by MS/MS using an entire translated genome: demonstration on the ciliome of *Tetrahymena hermophila*. *J Proteome Res* 2005;**4**:909–19.