Proteomic applications in microbial identification

John P Bowman

University of Tasmania
Tasmanian Institute of Agricultural Research
School of Agricultural Science
College Road, Sandy Bay, Hobart
TAS 7005
Tel (03) 6226 6380
Fax (03) 6226 2642
Email john.bowman@utas.edu.au

Proteomics-based approaches have been used in microbial taxonomy for the last several decades. Recent improvements in instruments and software have led to the appearance of mass spectrometric fingerprinting and peptide survey approaches allowing for highly rapid and accurate taxonomic diagnoses suitable for high-throughput laboratories as well as means to deeply analyse entire proteomes.

Taxonomy has an important role to play in microbiology being mainly for: 1) identification of pathogens performed in clinical diagnostic laboratories; 2) epidemiological strain typing; and 3) for diversity and bioprospecting studies in which bacterial names are used to encapsulate information. In the case of clinical diagnosis, a key aspect is rapidity and accuracy; for epidemiology, discrimination of strains in a reliable and meaningful way is required; while broader ecology-oriented applications of taxonomy rely on a robust classification system well served by ancillary data that is phenotypic, biochemical, chemical and genomic. Methodologies providing these aspects are continuingly improving. Typical of the current zeitgeist, we want the best of worlds – speed and data richness. That is, approaches that might be obtainable in real time; allow discrimination at fine scales and that are highly information-rich. A challenging prospect, since some of these desired aspects are to some extent mutually exclusive. However, specific developments in proteome-based fingerprinting using matrix-assisted laser desorption ionisation-time of flight spectrometry (MALDI-TOF); and multidimensional protein information technology (MudPIT) have the potential to meet some if not all of these challenges.

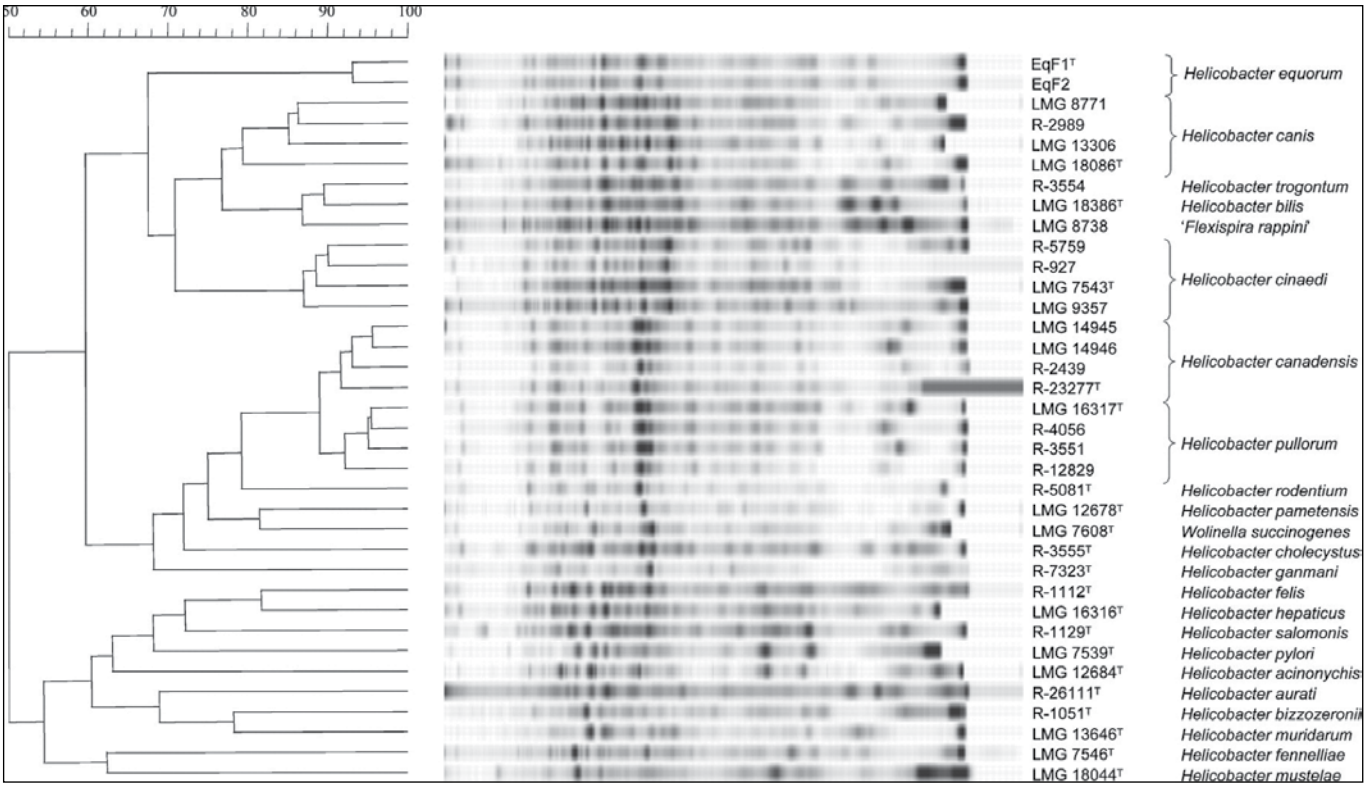


Figure 1. One-dimensional SDS-PAGE analysis of proteins extracted from *Helicobacter* species with profiles analysed by densitometric and numerical analysis using GelCompar v. 4.2 (Applied Maths). From Moyaert *et al.*[16], with permission from the Society for General Microbiology.

Proteomics-based analyses up until the last five or so years have generally involved one-dimensional (1D) or two-dimensional (2D) gel-based technologies. The patterns of proteins resolved on the basis of molecular weight with polyacrylamide gels in 1D from extracts have been used in taxonomic comparisons of strains since the 1970s[1]. These methods have been rather useful in resolving related species or strains (example in Figure 1). Since most descriptions are now based on single isolates, these methods are rarely used for direct characterisation work, and are being increasingly replaced by PCR-based fingerprinting methods such as REP, ERIC, AFLP and so on[2]. Advances in instrument technology have provided new ways to do proteomics and, as described below, these have been or can be adapted for microbial taxonomic applications.

The principal of MALDI-TOF as it can be applied to taxonomy involves comparison of a series of protein fragment masses with the assumption that particular bacterial species or strains possess diagnostic patterns of these features, deriving from strain- or species-dependent proteins. The company Bruker Daltonics has developed a MALDI-TOF instrument (the Biotyper) and database that includes a large range of clinically relevant bacterial species. Several laboratories are expanding this database for specific applications[3]. The MALDI Biotyper streamlines the technical aspects of MALDI-TOF, basically adapting it for use in general diagnostic laboratories and such instruments started appearing in 2009. Independent studies[4,5] showed the Biotyper approach is more reliable than biochemical approaches and is more accurate at identifying genera (0.1% error level). Species identification error was 3% for *Enterobacteriaceae*, 6% for staphylococci, 8% for non-fermentative Gram-negatives, 15% for streptococci, 16% for *Pasteurellaceae* and *Neisseraceae* and 15% for yeasts[4]. This compares with error rates two to three times higher for biochemical strip-based assays. Misidentifications by MALDI-TOF MS are likely due to database problems in which strains were misidentified in the first place[6]. In terms of specificity, the approach could be useful for delineating intraspecific features of species, but this depends on dedicated development of databases in line with other approaches of strain subtyping such as multilocus sequence typing. As an example, subtypes of various *Yersinia* species, such as *Y. pseudoenterocolitica*, can be potentially delineated but other *Yersinia* spp. cannot[7], simply due to the nature of the current taxonomic framework. For example, the species *Y. pestis*, *Y. pseudotuberculosis* and *Y. similis* all have rather similar MALDI-TOF fragment profiles, reflecting their close genetic allegiances[7].

In the case of non-clinical-based MALDI-TOF applications (for example, food microbiology[8]) the strength of the methodology is also dependent on database development and whether this

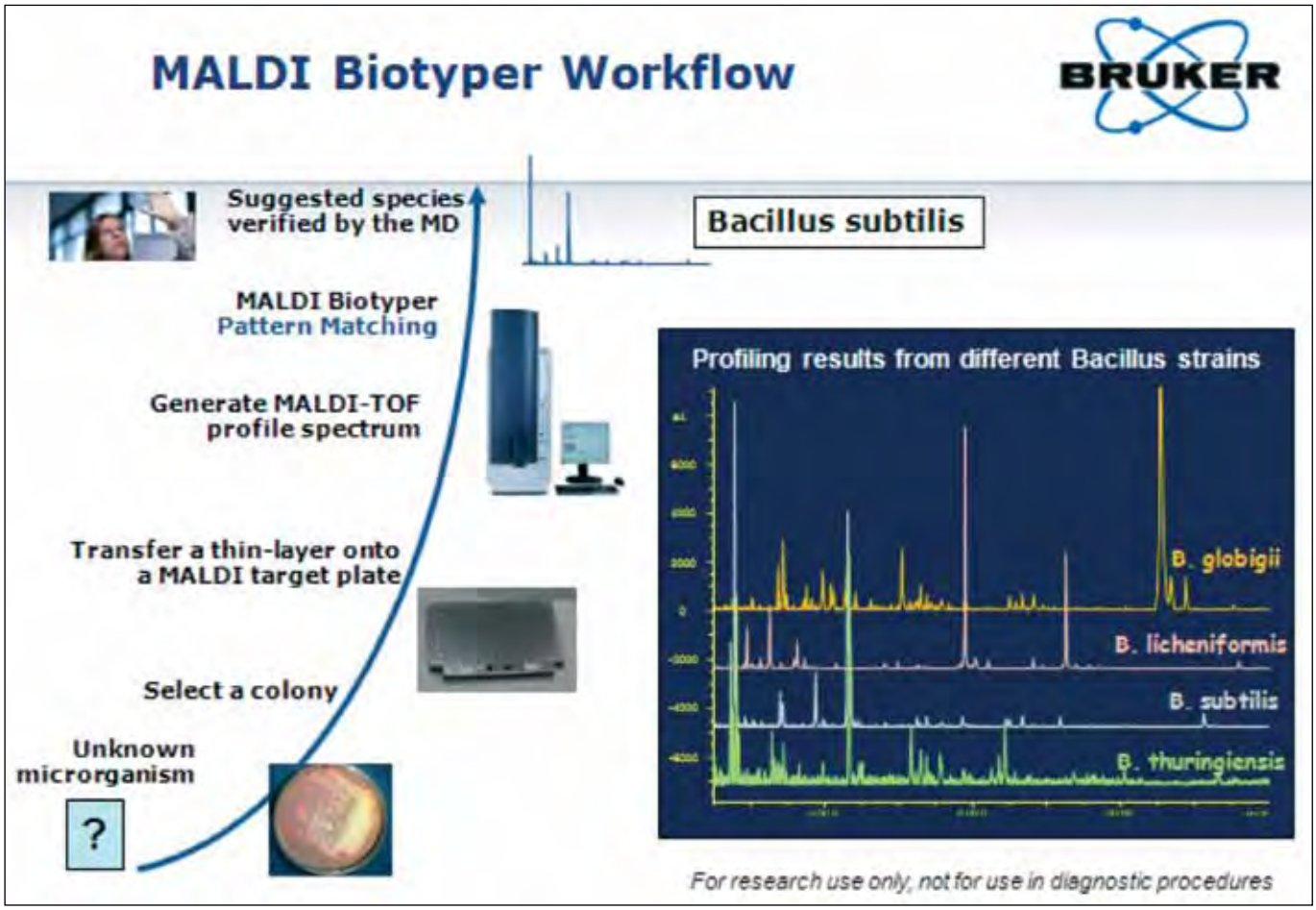
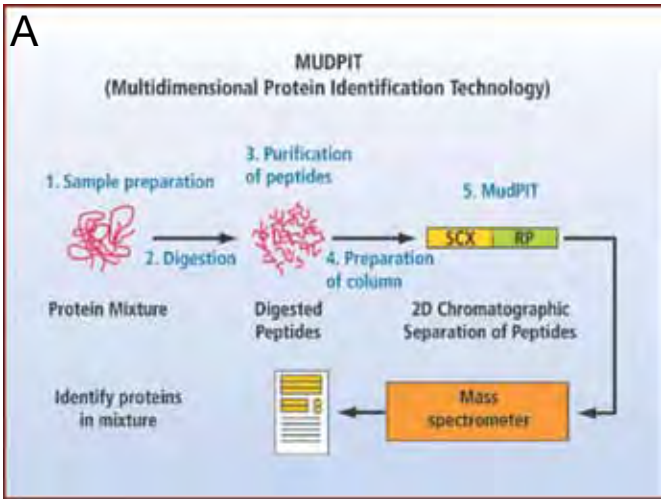
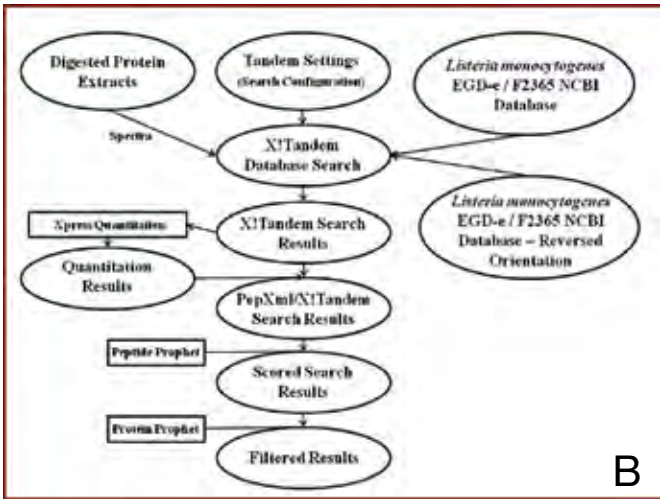


Figure 2. Schematic showing the sample preparation method for analysis and identification of bacteria using MALDI-TOF analysis as applied in the MALDI Biotyper (Bruker Daltonics). From Wolfgang Pusch, with permission.

happens likely depends on its general take-up and utilisation. Methodologically sample preparation is very simple and rapid (Figure 2), which for clinical laboratories is ideal as it streamlines processes and thus can generate potential cost benefits. Samples are mixed with a matrix chemical that absorbs laser energy, encouraging ionisation of protein fragments that have masses of 2,000–20,000 daltons. The samples are then spotted onto ground steel sample targets. A series of mass spectra are then obtained and software used to generate a list of protein fragments differentiated by molecular mass. Software is used to compare different peak lists and thus a database can be built up. Distinct fragments found can be potentially matched to genome data based on distinct size or analysed further by liquid chromatography/mass spectrometry (LC/MS) to obtain peptide sequence data. There is strong potential for improvements in MS technology and the databases to allow for higher reliability and greater versatility of MALDI-TOF technology in microbial identification and could eventually allow rapid determination of specific proteins that differentiate strains.

MudPIT provides an alternative means for evaluation of Bacteria and Archaea (and any other form of biological material) and relies on existing accumulated proteomes inferred from genome data (for example, Uniprot, SWISSPROT, EXPASY, NCBI databases). Unlike MALDI-TOF, MudPIT analysis is far more comprehensive and provides information depth rather than rapid identification. The method essentially profiles as much as possible the entirety of a given proteome by a random survey of peptides derived from a digested protein extract[9]. Proteins can be labelled or unlabelled, with the latter option becoming more popular and has been shown to be sufficiently reliable enough to allow quantification of protein abundance[10]. The process, outlined in Figure 3A, relies on high-resolution separation of peptides by 2D LC and subsequent analysis of the peptides relying on





C

Run	Peptide	Charge	Protein	Total Peptide	Max Peptide Prophet	Avg Retention Time
data_April4148	R.GVIALTYFGIGNSR.A	2	gi|85819718|gb|EAQ40875.1|	25	1.0000	22957.80008
data_April4148	F.SGTYKSNAFLPGLGVTYKL.-	2	gi|91216691|ref|ZP_01253656.1|	22	1.0000	26534.00906
data_April4148	R.DVTWTEFFSSLEADLK.F	2	gi|91216970|ref|ZP_01253933.1|	21	1.0000	12436.28997
data_April4148	L.GATVGADVDINPLIIGVGIGMRF.-	2	gi|91215125|ref|ZP_01252097.1|	17	1.0000	22180.32933
data_April4148	K.GAYVAESVYVGDEYLDALVDIK.S	2	gi|91215206|ref|ZP_01252178.1|	16	1.0000	14514.49365
data_April4148	R.RSELALVSYGLF.A	2	gi|91216916|ref|ZP_01253880.1|	16	0.9999	19541.8064
data_April4148	R.SLVFNMEGVTK.G	2	gi|91216895|ref|ZP_01253859.1|	16	1.0000	11087.56256
data_April4148	-.MYAIVEIAGQQFK.V	2	gi|91217999|ref|ZP_01254951.1|	15	1.0000	9513.90599
data_April4148	R.IGIAAQALGIAAGAYELAK.D	2	gi|120437230|ref|YP_862924.1|	15	1.0000	17041.75352
data_April4148	R.LILEEAWPFGNVATDITYK.I	2	gi|91215942|ref|ZP_01252911.1|	15	1.0000	12450.05339
data_April4148	K.IGELIAEAFGK.V	2	gi|91216876|ref|ZP_01253840.1|	14	1.0000	8814.78711
data_April4148	A.DVDINPLIIGVGIGMRF.-	2	gi|91215125|ref|ZP_01252097.1|	13	1.0000	10476.84604
data_April4148	A.LVESDGDFDLAIENLR.K	2	gi|150024556|ref|YP_001295382.1|	13	0.9981	9568.94779
data_April4148	K.AMLEDIAILTGGTVISEER.G	2	gi|91216876|ref|ZP_01253840.1|	13	1.0000	12859.68382
data_April4148	K.LTFTPIFMEAVAR.T	2	gi|91216054|ref|ZP_01253023.1|	13	0.9999	10162.85006
data_April4148	R.VITTGVPVLLIAADAR.S	2	gi|91215148|ref|ZP_01252120.1|	13	1.0000	9062.7192
data_April4148	K.APIGTIFTDSVFTPIK.N	2	gi|91216884|ref|ZP_01253848.1|	12	1.0000	16125.6416
data_April4148	K.ESFFDVASVDELIFR.G	2	gi|91216916|ref|ZP_01253880.1|	12	1.0000	12108.44661
data_April4148	K.LTYFLEPIEDIYLYSER.G	2	gi|91218179|ref|ZP_01256127.1|	12	1.0000	10265.51082
data_April4148	K.NLPDFILADVTPLK.I	2	gi|91217230|ref|ZP_01254191.1|	12	1.0000	9215.92497
data_April4148	R.NTVGEVYGFIVDDLTDAAR.D	2	gi|91216917|ref|ZP_01253881.1|	12	1.0000	17210.72502
data_April4148	R.VLLSSLEGFAITSVR.I	2	gi|91216884|ref|ZP_01253848.1|	12	1.0000	10965.79248
data_April4148	R.YIAVQNELVAAYNVLR.D	2	gi|91215091|ref|ZP_01252063.1|	12	1.0000	9113.20158

Figure 3. Schematic showing A) the analytical analysis of complex cell lysates by the MudPIT procedure; B) an example of an automated bioinformatics pipeline for protein identification for *Listeria monocytogenes* (based on the Computational Protein Analysis System, www.LabKey.org); C) a partial extract of spectral count data output from a MudPIT analysis of *Psychroflexus torquis* ATCC 700755^T (using a LTQ-Orbitrap MS system, ThermoScientific) showing peptide sequence, protein identification (matched to the NCBI database), total spectral counts for the given peptide, probability of correct match (ProteinProphet Score, a value of 1.0 being practically an error-free match), and retention time (in seconds) during the 2D liquid chromatography separation phase.

highly accurate mass estimation, high responsiveness, and sensitivity to dynamic ranges in ion intensity. A good instrument touted for this application is the Linear Trap Quadropole (LTQ)-Orbitrap, a modified ion-trap mass spectrometer that offers about 1: 2 million error ratio in mass calculations[9]. Mass spectra are surveyed over an eight to nine hour period with spectra accumulated as a spectral count, which can serve as a means of quantification of protein abundance[11]. The basic assumption is that the more abundant the protein the more likely it will be observed, relative to its molecular weight. To identify proteins, bioinformatic pipelines are necessary since a single run can yield >10,000 mass spectra. An example of such a pipeline is shown in Figure 3B and relies on two levels of verification to estimate the possibility of a false hit. This is critical since false-positive identification is a problem in proteomics and rigorousness in post-evaluation of mass spectra data is now an expectation. The approach is being used for example to develop a tissue-level Human Proteome[12]. In the case of bacteria potentially as much as 50–60% of the proteome can be revealed by MudPIT surveys, taking into account extraction limitations. Membrane-associated and secreted proteins will be underrepresented, if proteins are extracted using standard methods; a number of approaches can be used to obtain more complete protein coverage[13]. Due to this level of comprehensiveness MudPIT can generate a list of proteins, for example, Figure 3C, which could be potentially used as an adjunct to phenotypic and physiological characterisation of organisms. At this stage the application has not been directed specifically towards taxonomic characterisation; however, the potential is plain in its use as a diagnostic tool. For example, MudPIT was recently directed towards surveillance of *Bacillus anthracis* in mixed populations[14]. MudPIT has been also used to compare differences between strains, for example, strains representing different genetic lineages of *Listeria monocytogenes*[15] revealing considerable potential variation in functional aspects. MudPIT can be advocated as a forthcoming approach that could be used to assess taxonomic differences and also determine important cellular properties such as metabolic pathways, virulence proteins, stress tolerance proteins and specific metabolic activities.

Overall, applications for proteomics are relevant to all of biology and provide a means to utilise genomic data in a more effective way. In taxonomy it is possible proteome profiling can be used as both a rapid characterisation tool as well as an identification tool. Software is increasingly becoming available that allows one to map proteomic data onto genome information that is, the omics viewer within the BioCyc database (www.biocyc.org). The main limitation is the state of the bioinformatic knowledge related to protein functionality; however, this should improve given the large emphasis in microbiological sciences towards molecular and genomics-oriented studies in the last decade.

References

1. Cash, P. (2009) Proteomics in the study of the molecular taxonomy and epidemiology of bacterial pathogens. *Electrophoresis* 30, S133–S141.
2. Rademaker, J.L.W. *et al.* (2000) Comparison of AFLP and rep-PCR genomic fingerprinting with DNA–DNA homology studies: *Xanthomonas* as a model system. *Int. J. Syst. Evol. Microbiol.* 50, 665–677.
3. Ferreira, L. *et al.* (2010) Identification of *Brucella* by MALDI-TOF mass spectrometry, fast and reliable identification from agar plates and blood cultures. *PLoS ONE* 5, e14235.
4. van Veen, S. *et al.* (2010) High-throughput identification of bacteria and yeast by matrix-assisted laser desorption ionization-time of flight mass spectrometry in conventional medical microbiology laboratories. *J. Clin. Microbiol.* 48, 900–907.
5. Saffert, R.T. *et al.* (2011) Comparison of Bruker Biotyper MALDI-TOF mass spectrometer to BD Phoenix Automated Microbiology System for identification of Gram negative bacilli. *J. Clin. Microbiol.* DOI:10.1128/JCM.01890–10.
6. De Bel, A. *et al.* (2010) Correct implementation of matrix-assisted laser desorption ionization-time of flight mass spectrometry in routine clinical microbiology. *J. Clin. Microbiol.* 48, 1991–1992.
7. Lasch, P. *et al.* (2010) Characterization of *Yersinia* using MALDI-TOF mass spectrometry and chemometrics. *Anal. Chem.* 82, 8464–8475.
8. Böhme, K. *et al*. (2010) Species differentiation of seafood spoilage and pathogenic Gram-negative bacteria by MALDI-TOF mass fingerprinting. *J. Proteome Res.* 9, 3169–3183.
9. Yates, J.R. *et al.* (2009) Proteomics by mass spectrometry: approaches, advances, and applications. *Annu. Rev. Biomed. Engineer.* 11, 49–79.
10. Cooper, B. *et al.* (2010) Relative, label-free protein quantitation: spectral counting error statistics from nine replicate MudPIT samples. *J. Amer. Soc. Mass Spectrom.* 21, 1534–1546.
11. Lundgren, D.H. *et al.* (2010) Role of spectral counting in quantitative proteomics. *Expert. Rev. Proteomics* 7, 39–53.
12. Orchard, S. *et al.* (2010) Tackling quantitation: a report on the annual spring workshop of the HUPO-PSI. *Proteomics* 10, 3062–3066.
13. Otto, A. *et al.* (2010) Systems-wide temporal proteomic profiling in glucose-starved *Bacillus subtilis*. *Nature Comm* 1, article 137.
14. Jabbour, R.E. *et al.* (2010) Double-blind characterization of non-genome-sequenced bacteria by mass spectrometry-based proteomics. *Appl. Environ. Microbiol.* 76, 3637–3644.
15. Donaldson, J.R. *et al.* (2009) Comparative proteomic analysis of *Listeria monocytogenes* strains F2365 and EGD. *Appl. Environ. Microbiol.* 75, 366–373.
16. Moyaert, H. *et al.* (2007) *Helicobacter equorum* sp. nov., a urease-negative *Helicobacter* species isolated from horse faeces. *Int. J. Syst. Evol. Microbiol.* 57, 213–218.

Biography

John Bowman is Associate Professor of Microbiology at the University of Tasmania. His research interests primarily are focused on understanding stress physiology and functional genomics of food-borne pathogens, Antarctic and aquaculture-related microbiology, and bioremediation-related systems.