Metaproteomics for Analysis of Microbial Function in the Environment

Philip L. Bond[1]* and Margaret Wexler[2]

[1]Advanced Wastewater Management Centre, The University of Queensland, Brisbane, Queensland 4072, Australia

[2]School of Biological Sciences, University of East Anglia, Norwich NR4 7TJ, UK

*For correspondence. E-mail phil.bond@awmc.uq.edu.au;

Tel. (+61) 7 33467841; Fax (+61) 7 33654726

Metaproteomics for Analysis of Microbial Function in the Environment

Philip Bond and Margaret Wexler

This report briefly describes the approach of using proteomic analyses to examine protein expression directly from environmental samples (termed metaproteomics). This approach has potential for solving one of the major challenges facing microbial ecologists, by providing insight of microbial function directly within samples.

The emerging opportunity

There is increasing emphasis to study microbial communities directly in their environments. Recent molecular analysis of environmental samples has vastly increased understanding of microbial diversity. However, the next big challenge is to understand details of function in these environments, particularly to link the phylogenetic and functional information.

Recent metagenomic sequencing projects, that analyse genomic DNA directly from environmental samples, are providing opportunities to make the above link. These studies vastly expand our knowledge of the genetic diversity and the physiological and metabolic potential within selected environments that include seawater samples [1, 2], an acid mine biofilm [3], and activated sludge [4]. The presently escalating sequence data (genomic and metagenomic) provides increasing potential for application of high throughput functional approaches. Transcriptomics and proteomics can be applied directly within

mixed culture to detect expression profiles and provide functional insight of microbial environments.

The procedure for metaproteomic analysis is basically that utilized for proteomic study of pure culture. This involves (i) sample preparation, (ii) protein extraction, (iii) separation of the proteins or peptides, usually in two-dimensions, and (iv) mass spectrometry (MS) analysis for identification of the proteins (figure 1). Detecting protein expression from environmental samples is not new [5, 6], however, interest has recently grown. Sequence data, improved protein separation techniques, and the rapidly improving protein identification by mass spectrometry provide new opportunity to apply large-scale proteomics and protein identification to environmental samples.

An ongoing challenge for proteomics is the identification of proteins. This can be achieved from peptide mass data if metagenomic data is available. For example, following separation of proteins or peptides, MS or tandem MS (MS/MS) can be used to generate peptide mass fingerprints (PMF) [7]. The PMF patterns are then compared to the metagenomic database for protein identification. Another approach for identification is to estimate the *de novo* protein sequence from the MS/MS data, and then search for homologous sequences. The latter requires extra effort compared to the PMF approach, however, *de novo* sequencing is especially useful for protein identification when corresponding metagenomic data in unavailable. This approach was recently used to identify more than 100 proteins that were differentially expressed following exposure of bacterial communities to cadmium [8].

Examples of metaproteomics studies

So far only a handful of studies in the literature examines the proteome of mixed culture samples. These studies include detection of proteins in high abundance during biological phosphorus removal in activated sludge wastewater treatment [9, 10]. Proteins associated with dissolved organic matter in soil and water have been analysed to detect the presence of broad taxonomic groups of microorganisms [11]. Expression profiles that have been examined include: an estuary transect [12], infant fecal samples [13], and freshwater samples following exposure to heavy metals [14]. In a landmark study, high-throughput proteomic analyses have recently been performed on acid mine biofilms (see more below) [15].

Metaproteomics was first applied to a laboratory-scale activated sludge reactor [9]. In that study, comparisons of proteome profiles are made, to determine metabolic details of a wastewater treatment process known as enhanced biological phosphorus removal (EBPR). Large-scale protein separation was performed by 2DE, and initially proteins were identified by MS/MS *de novo* sequencing of peptides [9]. However, metagenomic sequences of EBPR performing sludges recently became available [4], thus facilitating protein identification by analysis of PMF patterns. By this means, over 30% of highly expressed proteins chosen from 2DE gels, could be matched to the metagenome database (unpublished data). These results of comparative expression offer insight into EBPR biochemistry and enable refinement of the EBPR metabolic model.

The most extensive metaproteomic analysis to date was performed on an acid mine biofilm of low diversity [15]. The mine is characterised by low pH (~0.8) and microbially mediated iron oxidation that contributes to the acid mine drainage production. Here, liquid chromatography (LC) was used to separate the protein mixture following protein extraction and trypsin digestion. This LC-MS/MS approach utilises two dimensional chromatographic separation (typically strong cation exchange with reversed phase), that is often coupled with tandem MS for analysis of complex peptide mixtures [7, 16]. Following chromatographic separation, the peptides are further fragmented and analysed by MS/MS for peptide mass pattern matching to a database. Thus, corresponding sequence data is required for protein identification. In this case the metagenomic data set was of a similar biofilm from another part of the mine [3]. From the proteins identified (~2000) a high coverage (48%) of the predicted proteins for the dominating microorganism (*Leptospirillum* sp.) was obtained [15]. One highly abundant protein, annotated as a hypothetical, was further investigated and found to be an iron oxidising cytochrome, a key component of the energy generation in these biofilms. Here the proteomic results were instrumental in guiding the ensuing biochemical investigations.

Future directions

Proteomic analysis of mixed communities is challenging, especially in complex samples such as soil, as a typical analysis may only resolve <1% of the metaproteome [18]. Nevertheless, mixed community studies are exciting and timely given the improved techniques and capabilities of proteomics and

environmental genome analyses. The approach holds great promise for comparative analysis to examine response to a range of environmental perturbations such as stress and redox, and for monitoring metabolic and physiological activities.

There are a number of potential metaproteomic applications suited to the different protein separation techniques (2DE and LC). 2DE is labour intensive, however, presently it is preferred for quantification of expression and comparative studies. There are also some useful in-gel aspects of 2DE, for example, for incorporation of radiolabel to detect newly synthesized proteins. The LC-MS/MS approach, together with advanced *de novo* sequencing [19], hold much promise for high throughput metaproteomics. Additionally, quantitative analysis of LC-MS data has recently been achieved in pure culture studies [17], and likely metaproteomic studies will follow.

References

1. DeLong EF, Preston CM, Mincer T, et al. Community genomics among stratified microbial assemblages in the ocean's interior. *Science*, 2006. **311**(5760):496-503.
2. Venter JC, Remington K, Heidelberg JF, et al. Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, 2004. **304**(5667):66-74.
3. Tyson GW, Chapman J, Hugenholtz P, et al. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, 2004. **428**(6978):37-43.

4. Martin HG, Ivanova N, Kunin V, et al. Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. *Nature Biotechnology*, 2006. **24**(10):1263-1269.

5. Ehlers MM and Cloete TE. Protein profiles of phosphorus- and nitrate-removing activated sludge systems. *Water SA*, 1999. **25**(3):351-356.

6. Ogunseitan OA. Protein profile variation in cultivated and native freshwater microorganisms exposed to chemical environmental pollutants. *Microbial Ecology*, 1996. **31**(3):291-304.

7. Aebersold R and Mann M. Mass spectrometry-based proteomics. *Nature*, 2003. **422**(6928):198-207.

8. Lacerda CMR, Choe LH, and Reardon KF. Metaproteomic analysis of a bacterial community response to cadmium exposure. *Journal of Proteome Research*, 2007. **6**(3):1145-1152.

9. Wilmes P and Bond PL. The application of two-dimensional polyacrylamide gel electrophoresis and downstream analyses to a mixed community of prokaryotic microorganisms. *Environmental Microbiology*, 2004. **6**(9):911-920.

10. Wilmes P and Bond PL. Towards exposure of elusive metabolic mixed-culture processes: the application of metaproteomic analyses to activated sludge. *Water Science and Technology*, 2006. **54**(1):217-226.

11. Schulze WX, Gleixner G, Kaiser K, Guggenberger G, Mann M, and Schulze ED. A proteomic fingerprint of dissolved organic carbon and of soil particles. *Oecologia*, 2005. **142**(3):335-343.

12. Kan J, Hanson TE, Ginter JM, Wang K, and Chen F. Metaproteomic analysis of Chesapeake Bay microbial communities. *Saline Systems.*, 2005. **1**:7-19.

13. Klaassens ES, de Vos WM, and Vaughan EE. Metaproteomics approach to study the functionality of the microbiota in the human infant gastrointestinal tract. *Applied and Environmental Microbiology*, 2007. **73**(4):1388-1392.

14. Maron PA, Ranjard L, Mougel C, and Lemanceau P. Metaproteomics: a new approach for studying functional microbial ecology. *Microb Ecol.*, 2007. **53**(3):486-93. Epub 2007 Mar 13.

15. Ram RJ, VerBerkmoes NC, Thelen MP, et al. Community proteomics of a natural microbial biofilm. *Science*, 2005. **308**(5730):1915-1920.

16. Link AJ, Eng J, Schieltz DM, et al. Direct analysis of protein complexes using mass spectrometry. *Nature Biotechnology*, 1999. **17**(7):676-682.

17. DeSouza L, Diehl G, Rodrigues MJ, et al. Search for cancer markers from endometrial tissues using differentially labeled tags iTRAQ and clCAT with multidimensional liquid chromatography and tandem mass spectrometry. *Journal of Proteome Research*, 2005. **4**(2):377-386.

18. Wilmes P and Bond PL. Metaproteomics: studying functional gene expression in microbial ecosystems. *Trends in Microbiology*, 2006. **14**(2):92-97.

19. Ma B, Zhang KZ, Hendrie C, et al. PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Communications in Mass Spectrometry*, 2003. **17**(20):2337-2342.

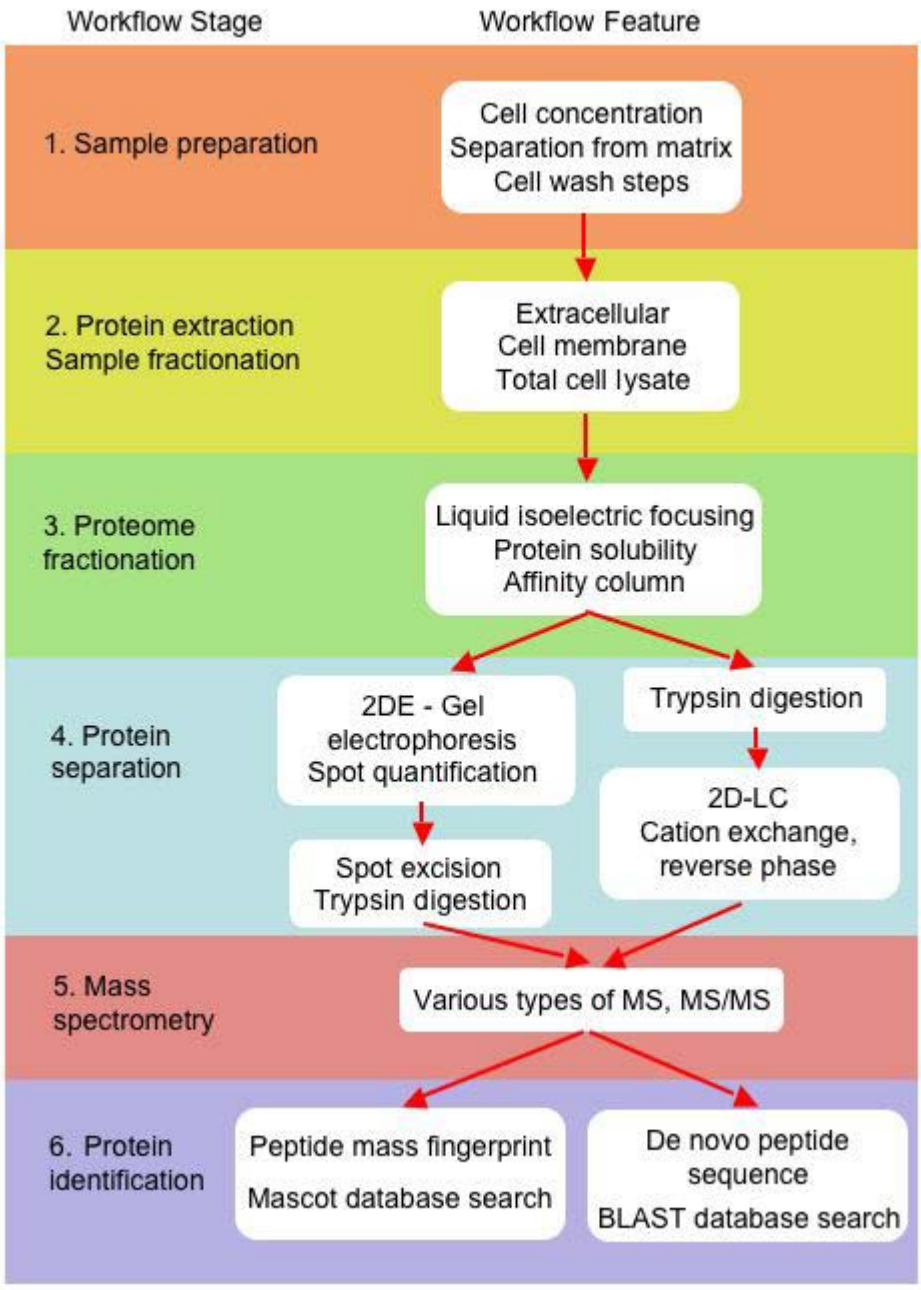

Figure 1. Workflow for a metaproteomic analysis may consist of six stages. Sample preparation may be required (stage 1). e.g. cells may need to be concentrated or purified away from interfering substances, such as humic acids in soil. Protein extraction is performed (stage 2) and

fractions of interest may be targeted, e.g. extracellular, cell membranes, or whole cell fractions. The procedures in these stages must have minimal effect on the protein expression itself and sufficiently preserve those extracted. To assist latter 2D separations, the extracted metaproteome may be subdivided (stage 3). For example this could based on solubility. The 2D separations may be performed by either 2-dimensional polyacrylamide gel electrophoresis (2DE), or by use of liquid chromatographic (LC) methods. Following 2DE, gel images are analysed and spots quantified. Chosen spots are then excised and trypsin digested for mass spectrometry analysis. For LC, the protein mixture is trypsin digested prior to separation, the separated peptides then flow directly into the mass spectrometer. In stages 5 and 6, peptide mass fingerprints can be generated by MS. Further mass analysis may be performed by MS/MS following fragmentation of the peptides. Additionally, de novo protein sequence data can be determined from the MS/MS data. Algorithms such as Mascot (http://www.matrixscience.com) enable the MS data to be searched against sequence databases for protein identification.