

Nanoflow LC–Q-TOF MS for *De Novo* Peptide Sequencing in Microbial Proteomics

Bart Devreese and Jozef Van Beeumen, Laboratory of Protein Biochemistry and Protein Engineering, Ghent University, Belgium.

The publication of several genome sequences has dramatically changed the scope of biochemical sciences. The availability of the libraries of genes that an organism can use to develop and to adapt to external stress has shifted biochemical research to the analysis of the final gene products; that is, proteins, their post-translational modifications and their interactions with other biomolecules. The first step in this ‘proteomics’¹ approach is the identification of proteins involved in a particular biological process. This work is generally based on so-called *differential expression analysis*, which generates information about genes that are switched on or off in the presence of a particular external signal.

Some of the methods (e.g., microarray techniques) for differential analysis focus on messenger ribonucleic acid (mRNA) expression, and take advantage of the array of molecular biology tools available (e.g., polymerase chain reaction technology to amplify the transcripts of genes expressed at low levels). However, critics state that quantitative analysis of mRNA levels does not always reflect the presence of a functional protein. Indeed, many proteins are post-translationally modified, or need to be translocated to specific cell organelles before they can exhibit their function. Therefore, it is widely accepted that analysis of gene expression at the protein level is necessary to study the functional role of individual genes in depth.

Several techniques are available to perform differential analysis of protein expression. Novel approaches describe the use of protein arrays² or multidimensional separation techniques³ of which some use differential isotopic labelling methods.⁴ Despite these new developments,

two-dimensional gel electrophoresis is still the most widespread method for the routine separation of proteins expressed by individual cells. Proteins are separated first on the basis of charge by isoelectrofocusing, using immobilized pH-gradient polyacrylamide strips now commercially available in various lengths and pH gradients. In the second dimension proteins are separated by mass using the well-known technique of sodium dodecylsulfate-polyacrylamide gel electrophoresis (SDS-PAGE). The protein spots are routinely visualized using Coomassie blue or silver staining techniques, although the use of fluorescent dyes is emerging. In addition, several manufacturers have developed software tools for the quantitative analysis of these gel profiles, allowing the identification of up- or down-regulation of gene expression when comparing protein extracts from cells grown under different conditions. The final task of this proteomics approach is the identification of protein spots for which differential expression has been demonstrated. Mass spectrometry (MS), using either electrospray ionization (ESI) or matrix-assisted laser desorption/ionization (MALDI), is the key technology for this purpose. Generally, individual protein spots are digested using trypsin and, after extraction, the peptide mixture is analysed via MS. Different approaches for protein identification from the MS data have been developed. The most widely used is peptide mass fingerprinting (PMF).⁵ In this technique the search is simply based on an algorithm that looks in a database for a simulated tryptic digest of a protein that best matches with the experimental one: masses of the obtained tryptic peptides are compared with those from peptides

theoretically expected from a cleavage after lysine and arginine residues of all the proteins present in the database. MALDI-time of flight (TOF) MS is the most appropriate method to perform PMF because of the low demands on sample preparation, and particularly because of the low complexity of the spectra; that is, the one peptide–one peak result in contrast to the multiple charging events in ESI. While PMF is certainly useful for high-throughput approaches, because it is easily automated, its scoring rate is often limited. It accounts poorly for post-translational modifications, handles poorly with mixtures, and certain groups of proteins are difficult to identify. For example, the scoring algorithms are mainly based on counting the number of hits (the more peptides matching, the higher the score). This way, one easily overlooks proteins with limited numbers of tryptic peptides, such as small proteins (<10 kDa) and large hydrophobic (membrane) proteins.

It has been shown that by including some sequence information, be it only a few residues, in the database search increases the reliability and the scoring rates by one order of magnitude. Although there are new developments coupling TOF-TOF and Q-TOF technology with MALDI,^{6,7} most of the MS–MS approaches in proteomics are still performed using ESI on ion trap or Q-TOF instruments. The disadvantage of ESI is its low tolerance to buffer salts, and sample clean-up is often a necessity. Some manual tools using small spinning columns or micropipette tips with chromatographic media have been developed, but none are useful in high-throughput environments. Therefore, on-line chromatographic separations prefacing mass spectrometric analyses are

the best choice. We have implemented a nanoflow liquid chromatography (LC) approach to combine high throughput with the sensitivity necessary to allow protein identification from 2D-PAGE separations.

The Nano LC System

An Ultimate micro LC system (LC Packings — A Dionex Company, Amsterdam, The Netherlands) is used (Figure 1). A Famos autosampler (LC Packings) is used for the injection of samples. The samples are first injected onto a micro precolumn, a procedure that has two advantages. First, it allows a significant reduction in injection times (nanoflow LC is performed at flow-rates of 100–150 nL/min and so it would take at least 30 min to inject a 3 μ L sample). The sample is injected at a flow-rate of 10 μ L/min using an external pump. The second advantage is the preliminary desalting: samples are cleaned up prior to injecting onto the separating column, thereby extending the lifetime of the nano LC column. After the injection/desalting step, the valve is switched to make a connection with the nano LC column. The peptides stored on the precolumn are eluted to the separating column and then separated using a standard reversed-phase (RP) gradient. The endcapping of the columns (PEPMAP, Dionex) allows the use of formic acid rather than trifluoroacetic acid (TFA) as the counter ion in the RP-LC separation, without loss of resolution. This has an important impact on the sensitivity in the mass spectrometer because TFA dramatically reduces signal intensities in ESI. In our system, the outlet of the column is connected to a homemade nano-ESI interface for which fused-silica capillary needles are used (New Objective, Woburn, Massachusetts, USA). Mass spectral data are acquired on a Micromass Q-TOF instrument (Manchester, UK). The instrument is run in the data-dependent MS to MS–MS switching mode, which means that any peptide eluting from the column with a mass peak above a certain threshold is automatically selected for MS–MS analysis.

De Novo Sequence Analysis and Microbial Proteomics

The Q-TOF MS–MS data quality not only allows the use of MS–MS data for supporting PMF in database searching for identifying proteins from organisms of which the genome has been sequenced, but also allows partial sequence analysis (*de novo* sequencing). The sequence data

Although there are new developments coupling TOF-TOF and Q-TOF technology with MALDI, most of the MS–MS approaches in proteomics are still performed using ESI on ion trap or Q-TOF instruments.

Figure 1: Schematic of the nanoflow LC system: (a) flow scheme during sample loading on the micro precolumn and (b) flow scheme during gradient elution of the peptides.

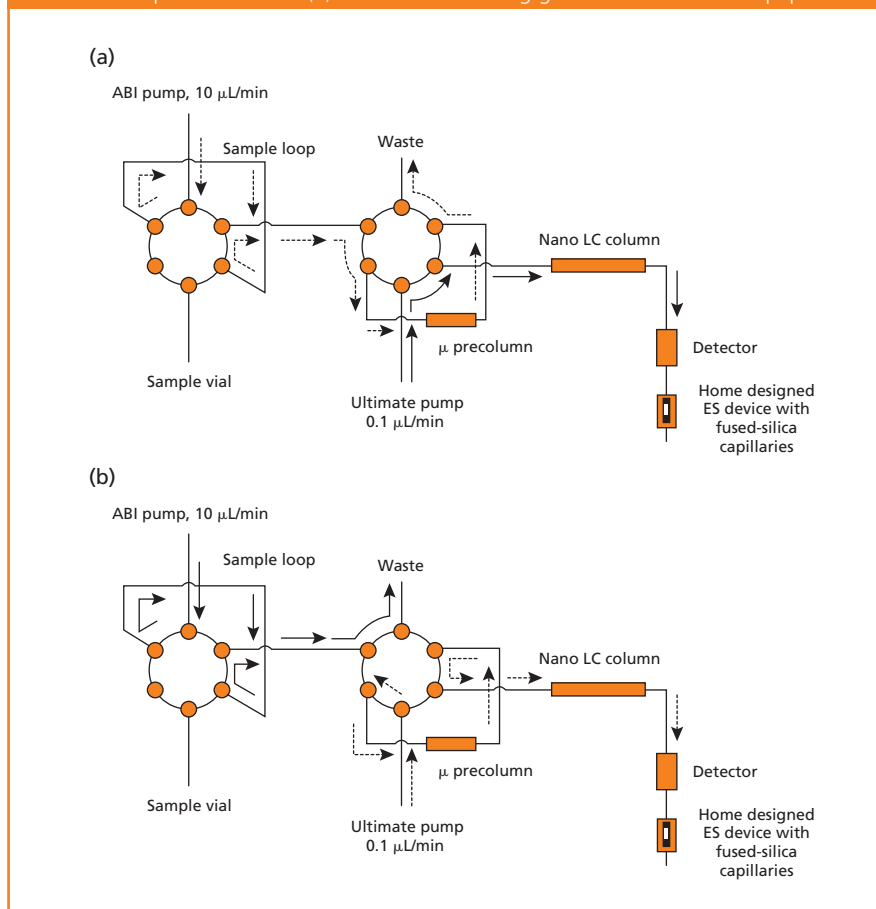


Figure 2: Details of the 2D-PAGE of a total lysate of *Shewanella hanedai* grown at (a) 12 °C and (b) 4 °C. Arrows highlight the two analysed proteins.

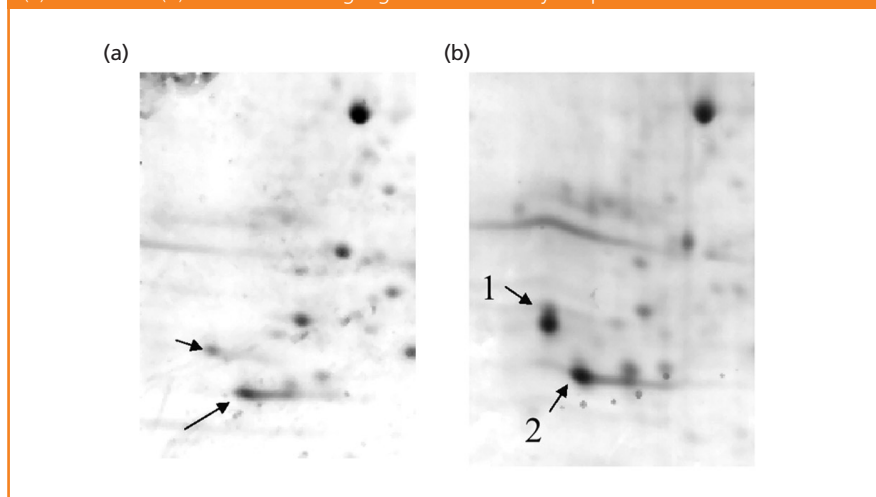


Figure 3: Total ion current chromatogram for the MS–MS scan of the tryptic digest of spot 1 shown in Figure 2. The numbers on top of the peaks refer to the mass of the molecular ion selected for fragmentation.

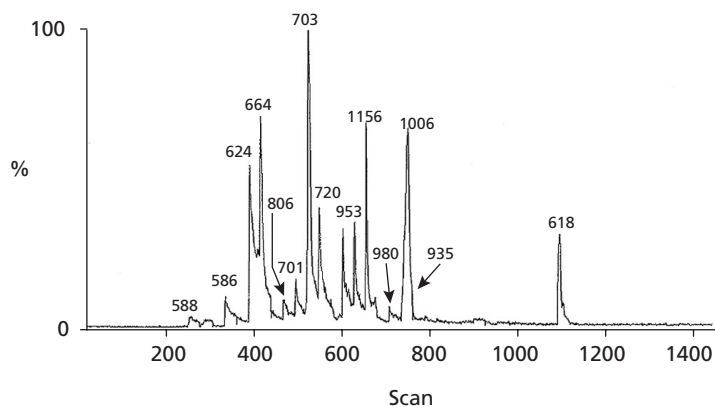
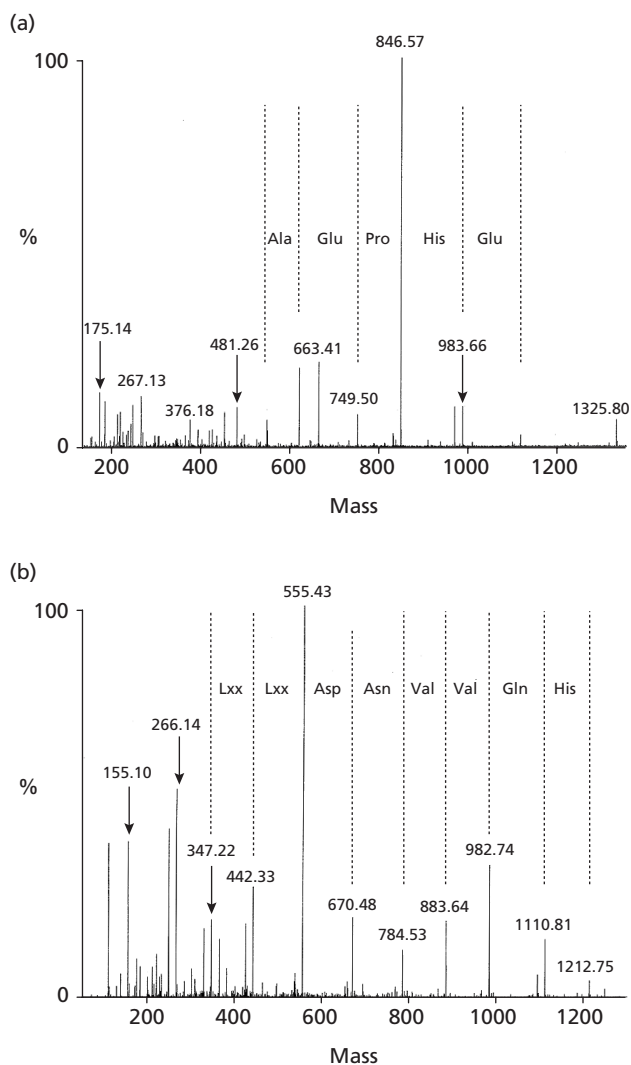


Figure 4: MS–MS spectra of two peptides with molecular ions of (a) m/z 664 and (b) 624 automatically selected for fragmentation as seen in Figure 3.



can then be used in similarity searches to identify proteins from model organisms of which no genome sequencing efforts have been undertaken. Today, according to the database maintained by the Institute of Genomic Research, the genomes of some 60 bacterial strains have been completely sequenced, and more than 100 other sequencing efforts are underway. Nevertheless, many model organisms of biomedical or environmental importance remain of which no genome sequencing effort has been initiated. Our mass spectrometric approach for *de novo* sequence analysis allows the initiation of proteomic efforts without needing genomic information on the bacterium of interest. The increasing number of protein sequences in the databases, covering a wide range of bacterial subfamilies, provides the data necessary for efficient similarity searching approaches for protein identification.

Example — *De Novo* Sequencing of Cold-Induced Proteins

One of our study targets is a psychrophilic bacterium, *Shewanella hanedai*. It is a marine Antarctic bacterium and a potential source of poly-unsaturated fatty acids, widely used as food supplements.⁸ We have initiated a study of the cold adaptation of this organism, because the synthesis of these fatty acids seems to stem from a mechanism to increase membrane flexibility, which is an essential way to maintain motility and to ensure transmembrane transport of nutrients at near-zero degrees Celsius. The protein and DNA sequence data available on this organism are very limited, and we must rely on similarity searching for database searching. Figure 2 shows the 2D-PAGE gels of a total lysate of *S. hanedai* grown at different temperatures. Clearly, two protein spots are higher in expression when the organism is grown at lower temperatures. The spots were cut from the gel, digested, and prepared for nano LC–MS–MS analysis. For a detailed description of the sample preparation we refer to a forthcoming paper.⁹ The mass spectrometer was used in automated MS to MS–MS switching mode, in which it first generates a survey scan. Each time a peptide elutes from the column and gives an ESI signal above a certain threshold it is selected for collision-induced dissociation analysis. Figure 3 shows the total ion current chromatogram display for this MS–MS analysis. After the peptide analysis is finished (either because the signal drops

The Q-TOF MS–MS data quality not only allows the use of MS–MS data for supporting PMF in database searching for identifying proteins from organisms of which the genome has been sequenced, but also allows partial sequence analysis (*de novo* sequencing).

below a threshold level, or because of a set time limit) the instrument switches back to a survey scan. Selection of an individual peak from the TIC chromatogram allows the analysis of the MS–MS spectrum of the corresponding peptide, as shown in Figure 4. The quality of Q-TOF MS–MS is generally sufficient to read at least half of the sequence of a 15–20 amino acid residue peptide, delivering a portion of the sequence that allows a significant database searching based on similarity. We have mainly used the Blast-search routine provided by the National Center for Biotechnology Information (www.ncbi.nlm.nih.gov), freely accessible for the academic world. We thus take advantage of the growing number of proteins available in this database, as

provided by the genome project efforts on several micro-organisms. Figure 5 provides the identification of proteins based on the stretches of sequence determined from the MS–MS spectra. The first protein, alkylhydroperoxidase C, belongs to a family of stress response proteins that is expressed by many micro-organisms during exposure to external stress. The second protein is similar to a member of the Fe-superoxide dismutase family. This clearly shows that *S. hanedai* deals with stress responses in a very similar way to mesophilic organisms, because similar proteins are expressed when providing cold shock to organisms such as *E. coli*. The difference is that the onset of the stress response occurs at a much lower temperature in the psychrophilic organism.

Conclusions

The use of nano LC, incorporating an on-line microconcentration/desalting step, has proven to be an excellent tool for sample introduction in ESI Q-TOF technology. It provides both the sample clean-up (important for prolonging the lifetime of the column as well as the stability of the electrospray) and the sensitivity necessary to analyse tryptic digests from single spots of 2D-PAGE gels, generally providing femtomole amounts of peptides. The Q-TOF MS–MS spectra are of sufficient quality to allow (partial) *de novo* sequence analysis of the peptides, with sequence stretches of sufficient length for identification of peptides based on similarity searching.

References

1. M.R. Wilkins et al., *Bio/Technology*, **14**, 61 (1996).
2. A. Mirzabekow and A. Kolchinsky, *Current Opin. Chem. Biol.*, **6**, 70 (2002).
3. G.J. Opitck et al., *Anal. Biochem.*, **258**, 349 (1998).
4. S.P. Gigy et al., *Nature Biotechnol.*, **17**, 994 (1999).
5. D.J.C. Pappin et al., *Curr. Biol.*, **3**, 327 (1993).

Figure 5: Identification of the two proteins highlighted in Figure 2 based on similarity searching. The sequences shown in red are those obtained from mass spectral analysis and were used in the Blast search. The peptides were found to possess the highest similarities to proteins from *Vibrio cholerae*, which is evolutionary related to *Shewanella hanedai*.

Spot 1: Alkyl hydroperoxidase C					
MLRSKKMVLV	GRKAPDFTAA	AVLGNGEIVD	NFNFAEFTKG	KKAVVFFYPL	DFTFVCPSEL
IAFDNRYEDF	LAKGVEVIGV	SIDSQFSHNA	WRNTAVENGG	IGQVKYPLVA	DVKHEICKAY AY
DVEHPEAGVA DVEHPEAGVA	FRGSFLIDEE	GMVRHQVVND HQVVND	LPLGRNIDEM LLPGR	LRMVDALNFH	QTHGEVCPAQ
WEAGKAGMEA	SPKGVA AFLS	QYSADLK			
Spot 2: Fe-superoxidase dismutase					
MAFELPALPY	AKDALEPHIS	AETLDFHHGK	HHNTYVVKLN LD	GLIPGTEFEN GLVEGTELAE	KSLEEI KTS K
TGGIFNNAAQ	VWNHTFYWHC	LSPNGGGEPT SPDG	GAVAEAINAA DDA	FGSFADF KAK FGSFA	FTDSAINNFG TDSAVNNFG
SSWTWL VKKA SAWTWL VK	DGTLAITNTS	NAATPLTEEG	VTPLLTVDLW	EHAYYIDYRN	VRPDYMNFGW
ALVNWDFVAQ	NLAK				

6. K.F. Medzhiradzky et al., *Anal. Chem.*, **72**, 552 (2000).
7. P. Verhaert et al., *Proteomics*, **1**, 118 (2001).
8. J.P. Bowman et al., *Int. J. Sys. Bacteriol.*, **47**, 1040 (1997).
9. B. Devreese et al., *J. Chromatogr. A*, in press (2002).

Bart Devreese obtained his PhD degree in 1997 in the laboratory of Professor Van Beeumen, having used the novel soft ionization mass spectrometric techniques in the elucidation of the primary structures of redox proteins from sulfate reducing bacteria. In 1999 he obtained a Postdoctoral Fellowship from the Fund for Scientific Research-Flanders, Belgium and initiated proteomic analysis of psychrophilic and other extremophilic micro-organisms.

Jozef Van Beeumen is head of the laboratory of Protein Biochemistry and Protein Engineering at Ghent University, Belgium. His scientific career is centered around protein structure analysis, for the major part on bacterial redox systems, antibiotic resistance and extremophilic organisms. He was the first to use the techniques of ESI-MS and MALDI-MS in the field in Belgium in the early 1990s. His laboratory is also equipped with X-ray crystallographic equipment to study the 3-dimensional structure of proteins. He recently became president of the Department of Biochemistry, Physiology and Microbiology at Ghent University.