Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Biotechnology

Next-generation proteomics faces new challenges in environmental biotechnology

Jean Armengaud

Environmental biotechnology relies on the exploration of novel biological systems and a thorough understanding of the underlying molecular mechanisms. Next-generation proteomics based on the latest generation of mass analyzers currently allows the recording of complete proteomes from any microorganism. Interpreting these data can be straightforward if the genome of the organism is established, or relatively easy to perform through proteogenomics approaches if a draft sequence can be obtained. However, next-generation proteomics faces new, interesting challenges when the organism is distantly related to previously characterized organisms or when mixtures of organisms have to be analyzed. New mass spectrometers and innovative bioinformatics tools are reshaping the possibilities of homology-based proteomics, proteogenomics, and metaproteomics for the characterization of biological systems. Novel time- and cost-effective screening strategies are also possible with this methodology, as exemplified by whole proteome thermal profiling and subpopulation proteomics. The complexity of environmental samples allows for unique developments of approaches and concepts.

Address
CEA, DRF, IBiTec-S, SPI, Li2D, Laboratory ''Innovative Technologies for Detection and Diagnostics'', Bagnols-sur-Cèze F-30200, France

Corresponding author: Armengaud, Jean (jean.armengaud@cea.fr)

Current Opinion in Biotechnology 2016, **38**:174–182

This review comes from a themed issue on **Environmental biotechnology**

Edited by **Benardo Gonzalez Ojeda** and **Regina Wittich**

http://dx.doi.org/10.1016/j.copbio.2016.02.025

Introduction

Understanding environmental biological processes in their most intimate detail is crucial in order to exploit environmentally friendly technologies, to remediate contaminated environments, and to preserve biodiversity. Exploration of the diversity of life on Earth has pushed scientists to probe life in the most extreme environments, with exquisite successes [1,2] leading to the discovery of very diverse molecular mechanisms and catalytic enzymes. Indeed, the existence of organisms able to withstand low water activity on Earth and the recent discovery of the presence of water on Mars open up the most incredible exo-environmental biotechnological perspectives [3,4]. The exploration of life on Earth is currently being fueled by, first, high-throughput meta-approaches based on the shotgun principle, which consists of sequencing as many fragments as possible and reconstructing the whole picture with the separate pieces of the puzzle, second, multi-omics technologies, and third, powerful bioinformatics. Driven by never-ending technological leaps, this exploration is now reaching an important crossroads: which direction to take for real innovation in terms of biology? While the main basic biological processes have been discovered and described, how can the most deeply hidden processes be unearthed? What efforts should be made in terms of technology, given that collecting vast quantities of data has ultimately become rather easy? Which breakthrough technological developments will soon be accessible to environmental biotechnology? Methods based on advanced mass spectrometry have evolved rapidly over the last decade, particularly for the analysis of large biomolecules. Fast scanning of complex proteomes is becoming routine in most proteomics platforms equipped with the latest generation of mass analyzers. These new tools open up a new era where complete proteomes can be quickly established for a given biological model and dozens of physiological conditions [5,6]. Today, the complexity of unknown biological samples or mixtures of organisms, that is, microbiomes, can be addressed categorically [7]. I present here the more recently published work relating to the new challenges that next-generation proteomics is facing in terms of environmental biotechnology, and the perspectives of this methodology.

Impact of mass spectrometry in environmental biotechnology

Mass spectrometry of biomolecules should combine soft ionization, to allow intact ions to enter into the mass spectrometer, a high-resolution fast scan of the mass/charge (m/z) ratio of these ions, and high-sensitivity detection. In common practice, three ionizations are currently used to analyze biomolecules, the first two being well adapted for proteins or peptides: first, matrix assisted laser desorption ionization (MALDI), where the analytes are mixed with a chemical matrix, dried, and then shot with a laser to form ions that are then analyzed and quantified with a detector; second, electrospray ionization (ESI), which allows the formation of ions from a liquid vein subjected to high electrical potential and the dissipation of the liquid sample in a homogeneous form, being perfectly adapted to the

generation of ions from the outlet of a reverse-phase chromatography column; and third, desorption electrospray ionization (DESI), which is based on a nebulized electrospray of highly charged microdroplets picking up the analytes present at the surface of the sample [8]. Mass spectrometry is now commonly used in analytical science to determine the exact molecules present in a given sample and to quantify them. As a result, this biophysical approach has impacted widely on environmental biotechnology in assessing the presence of contaminants or substrates, monitoring their variations, and characterizing biological processes of biotechnological interest and how organisms function. Astounding changes in terms of performance and coverage have been noted for high-throughput identification and quantitation of small molecules, that is, metabolomics [10•]. This has also occurred for protein measurements. For example, based on the molecular weight profile of the most abundant, basic, small proteins present in bacteria, determined by whole-cell MALDI time-of-flight mass spectrometry, it is possible to determine the taxonomy of any bacterium within a few minutes if the profile has been documented in the database. This approach has drastically changed microbiological diagnostics [9]. Proteomics, which is devoted to the high-throughput identification and quantitation of proteins and their numerous proteoforms, mostly relies on tandem mass spectrometry. First, the global molecular weight of the analyte is measured, then the analyte is fragmented, and then the molecular weights of the fragments are determined, giving insights into the analyte's intimate structure. Two possible strategies can be implemented: bottom-up proteomics where proteins are first proteolyzed into small peptides before tandem mass spectrometry for the identification of the peptide sequences, or top-down proteomics where native proteins are directly analyzed and then fragmented in the mass spectrometer for partial sequence determination [10•]. Proteomics is a cornerstone approach for identifying key players of biological processes, quantifying their abundances, and characterizing them in terms of structure and post-translational modifications. As shown in Figure 1, next-generation proteomics is based on protein sequence database searches fueled by genomics or RNAseq massive sequencing. Protein sequences can be obtained either after genome annotation or translation in all possible frames of the unannotated nucleic acid molecule sequences. Proteomics data can be interpreted even with only a draft genome to hand, as exemplified by two studies carried out on the *Tistlia consotensis* bacterium to decipher its salt tolerance mechanisms [11,12]. Such a strategy can be adapted in a straightforward manner for any nonmodel organism, but the use of RNAseq is recommended to create a protein database when eukaryotes are being investigated [13•]. Alternatively, *de novo* sequencing by high-resolution tandem mass spectrometry, especially through top-down approaches [14], and novel database search strategies [15] could be proposed for organisms that are phylogenetically distant from relatives that have been genome sequenced. Figure 2 shows how mass spectrometry-certified peptides are assigned to proteins in shotgun proteomics. When a mixture of very closely related organisms is analyzed, specific peptide variants may allow the distinction of subpopulations. Together with common peptides, these can be used to define protein groups. Figure 3 shows the main principle of metaproteomics when complex mixtures of organisms are analyzed as a whole. Metaproteins may be described for a general picture of taxa and the molecular functions present in the sample. Depending on the depth of the analysis and complexity of the sample, strain-resolved results may be obtained.

Next-generation proteomics

A drastic change in proteomics throughput was introduced with the commercialization, from 2005 onwards, of mass spectrometry hybrid instruments based on the Orbitrap analyzer and ion pulsed injection with a C-trap. The Orbitrap analyzer consists of an outer barrel-like electrode and a central spindle-like electrode along the axis, where ions oscillate axially across the trap at a frequency inversely proportional to the square root of their m/z ratio. Other instruments such as Q-TOF have also been improved to reach similar performances in terms of resolving power, mass accuracy (below 1 ppm for peptides), sensitivity, scan speed, and dynamic range. A new generation of Orbitrap has recently been released, with higher resolving power or higher scan speed for the same value of resolving power: high-field and ultra-high-field Orbitraps allow unprecedented resolving power on ionizable molecules: up to 500,000 full width at half maximum at m/z 200 [16•]. Further increases in this resolving power and scan speed are expected in the coming years, opening up wide new perspectives for proteomics. Currently, several tens of thousands of high-resolution MS/MS spectra can be recorded within an hour for a cellular protein extract. Kelstrup *et al.* [17••] exemplified the quick identification of 28,010 unique peptides corresponding to 4461 distinct proteins from one microgram of HeLa digest based on a dataset of 63,480 MS/MS spectra acquired in 60 min mass spectrometry time with a Q-exactive HF instrument (Thermo) incorporating a high-field Orbitrap analyzer. Some additional effort in terms of reverse phase fractionation and the analysis of replicates allowed the identification of 9515 proteins from a HeLa digest [18] with an Impact II Q-TOF (Brucker Daltonics). The age of complete proteomes has, thus, arrived [19], at least for proteomes from generalist bacteria in a specific physiological condition, for which less than two thousand distinct proteins are expected to be observed, even if the genome encodes more than 4000 protein sequences [20]. Absolute quantitation of a thousand proteins can be achieved by targeted mass spectrometry approaches [21,22], providing interesting perspectives for metabolic engineering, although in most cases relative quantitation is sufficient to obtain precious biological insights.