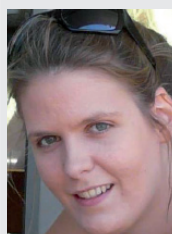


Shotgun proteomics: concept, key points and data mining

Expert Rev. Proteomics 7(1), 5–7 (2010)



Sabine Matallana-Surget

Department of
Proteomics and
Microbiology, University
of Mons, 6, Avenue du
Champs de Mars
(Pentagone), B-7000
Mons, Belgium
Tel.: +32 6537 3317
Fax: +32 6537 3320
[sabine.matallana@
obs-banyuls.fr](mailto:sabine.matallana@obs-banyuls.fr)



Baptiste Leroy

Department of
Proteomics and
Microbiology, University
of Mons, 6, Avenue du
Champs de Mars
(Pentagone), B-7000
Mons, Belgium
Tel.: +32 6537 3311
Fax: +32 6537 3320
[baptiste.leroy@
umons.ac.be](mailto:baptiste.leroy@umons.ac.be)



Ruddy Wattiez

Author for correspondence
Department of
Proteomics and
Microbiology, University
of Mons, 6, Avenue du
Champs de Mars
(Pentagone), B-7000
Mons, Belgium
Tel.: +32 6537 3312
Fax: +32 6537 3320
[ruddy.wattiez@
umons.ac.be](mailto:ruddy.wattiez@umons.ac.be)

“The goals of shotgun proteomics are as diverse as possible, focusing on only one cell type or one single organism and also being applied to whole microbial communities.”

Concept

Mass spectrometry (MS)-based approaches are used increasingly to address diverse questions in proteomics research, enabling one to comprehensively analyze complex samples of proteins on a global level. Owing to all the effort devoted to explore entire genomes, we now have access to a multitude of completed genome sequences, which has recently opened up a new era of ‘shotgun proteome sequencing’. The concept of shotgun proteomics basically consists of the multidimensional separation (MudPIT) of a complex mixture of peptides, generated from the treatment of a specific protease (e.g., trypsin), analyzed using MS/MS and then submitted to automated database searching.

Nowadays, the shotgun concept has expanded and does not exclusively refer to a gel-free methodology consisting of a direct loading into an online strong cation exchange (SCX)/reverse phase (RP)-liquid chromatography (LC) coupled to MS/MS (2D LC-MS/MS); note that the field has widened and can also include alternative fractionations, such as gel-based SDS-PAGE, isoelectric focusing and other LC separations of proteins and/or peptides. This diversity of workflows has already been demonstrated through different studies [1–5]. However, although a shotgun strategy is remarkably powerful and has clearly increased the overall data throughput compared with 2D gels, both approaches still remain complementary [6–8]. Today, the shotgun approach is increasingly used, not only for the identification of proteins but also for quantification. The goals of shotgun proteomics are as diverse as possible, focusing on only one cell type or one single organism and also

being applied to whole microbial communities. Although spectacular advances in MS have been achieved, we still have to face some other restricting steps, which we will discuss here, regarding both sample preparation/prefractionation and data validation.

Sample preparation & prefractionation: a challenging task

One of the most important issues in proteomics has always been, and still remains, the difficulty to access minor or under-represented proteins in a complex sample. Despite the undeniable improvement in mass spectrometer sensitivity, we have to face the remaining challenging task that is the dynamic range of protein concentrations. For example, the proteomic analysis of plasma and serum samples represents a great challenge owing to the wide dynamic range in protein concentrations ranging from a few picograms per milliliters for cytokines to milligrams per milliliters for a few highly abundant proteins, such as albumin and immunoglobulin. As a consequence, a dynamic range of protein concentrations and the presence of a few highly abundant proteins need to be considered before shotgun proteomic studies are undertaken. Hopefully, diverse strategies based on sequential protein isolation, traditional prefractionation or efficient depletion/enrichment methodologies have been developed in order to address this drawback, thus enabling identification of low-abundance proteins that may present a potential biological interest.

An appropriate protocol of protein isolation may constitute the initial step of prefractionation schemes and, for example,

can rely on selective protein isolation by using different buffers with increasing denaturation forces, thus enabling the separation of proteins according to their solubility. We could also mention the subcellular fractionation, an easy-to-use methodology that still remains particularly efficient and can be performed easily using differential centrifugation in order to target a particular subproteome or the overall set of fractions to increase the comprehensiveness of the proteome analysis. Membrane proteins, even if analyzed efficiently and detected with shotgun methodologies, are generally under-represented in crude protein extracts. In response to this challenge, a simple procedure based on differential centrifugations can efficiently resolve a complex protein extract by separating membrane fractions from the soluble one, thus allowing the analysis of a particular pool of proteins.

Owing to their complexity, protein extracts frequently need further fractionation to isolate low-abundance proteins from the high-abundance ones, such as additional online or offline LC separations, prior to loading into the mass spectrometer (3D LC-MS/MS). For instance, the immobilized metal affinity chromatography procedure that we are currently developing in our team, aims at performing an online separation of a complex proteome in a reduced number of fractions, ideally only two, based on the protein affinity for different metals, thus, also allowing the examination of the binding behavior of several proteins. In this case, the goal is to clearly increase the resolution of the chromatographic system by adding a third step as orthogonal as possible to the two others. In some cases, the prefractionation can also refer to an enrichment step, taking as an example the analysis of phosphorylations or other under-represented post-translational modifications. In this case, it is even possible in shotgun proteome analysis to directly include an enrichment step onto a fully automated standard 2D LC procedure, thus increasing automation and being perfectly suitable for the limited amount of biological material available. This latter procedure would then enable one to significantly increase the information content and would constitute by far one area in which shotgun proteomics shows particular promise.

“...it is even possible in shotgun proteome analysis to directly include an enrichment step onto a fully automated standard 2D LC procedure, thus increasing automation...”

Finally, we would like to emphasize another key point to increase the proteome coverage of some microorganisms that consists of simply multiplying the overall set of biological conditions, such as playing on media composition, growth and/or stress conditions. This latter observation remains obvious but is still, unfortunately, too often ignored. Indeed, by varying experimental conditions, several additional proteins may become detectable, and can be related to a particular stress or growth condition. This could represent a ‘biological fractionation’ performed by the organism itself, with all the proteins not being expressed all the time or facing all the conditions. For this reason, an experimental design has to be considered in detail before undertaking a comprehensive proteome description.

Data mining dataset validation: a crucial step

This section discusses the issues in regards to the data mining and validation of complex protein datasets. Results of large-scale proteomic experiments are often presented as a list of protein identifications (or a table with ratio values, for quantification) and most of the time criteria of their validation are not described. Thus, because of the lack of a common nomenclature to assign (or quantify) peptides, significant differences exist in the way different research groups validate their data.

“...a simple procedure based on differential centrifugations can efficiently resolve a complex protein extract by separating membrane fractions from the soluble one, thus allowing the analysis of a particular pool of proteins.”

The task of validating the peptides sequences is far from easy, requiring particular attention and being time consuming. When considering huge proteins searches where there are hundreds or thousands of low-scoring peptide matches, some problems may appear if the data are not inspected and validated manually. For instance, a protein with a score of 50 (using MASCOT), could provide poor MS/MS ion spectra, and most people would not consider this to be sufficient evidence for the existence of this protein. On the other hand, others may systematically cut off the list of identification at a score of 50 but some good spectra would then be lost. Clearly, some problems may arise if we want to assign a cut-off point for proteins and it currently seems to be quite a hot topic. Similarly, with quantitative shotgun, depending on the technology used, manual examination often reveals that software makes mistakes in identification and/or calculating ratios. This can result in the assignment of incorrect peptide sequences to MS/MS spectra or a bad estimation of quantitative ratios, which, in turn, can result in incorrect protein identification and lead to wrong or incomplete biological interpretation of the data. Thus, in response to this challenge, we definitively claim that a systematic identification of the overall set of proteins has to be validated by manual inspection of the MS/MS ion spectra, ensuring that a series of consecutive sequence-specific b- and y-type ions is well observed (for identification) and, if necessary, calculating the ratio again (for quantification).

Another major challenge often encountered in shotgun proteomics is related to the protein inference problem. Although a single-matched peptide is generally assumed to be sufficient to conclude that a protein is present in the sample, the analysis might turn out to be more difficult when working with eukaryotic organisms. Indeed, some sequences of identified peptides can be shared by several proteins in the database, making the discrimination very difficult between different proteins that are either isoforms or that share extensive homology [9]. New algorithms have been developed recently to address this challenge [9–2]. Furthermore, some amino acids have identical masses (Ile/Leu) or can be difficult to resolve (Asp/Asn and Glu/Gln/Lys), making it difficult to identify a correct peptide sequence. Hopefully, those latter ambiguities can be solved by using new-generation mass spectrometers providing a high mass accuracy [13].

Last but not least, it seems to us noteworthy to remind that the abundance of proteins can change, not only as a result of gene expression but also by increasing protein stability and turnover. In other words, after validation, biological interpretations and conclusions always have to be carefully thought about by the authors, before being brought.

Conclusion

In conclusion, significant improvements at both sample preparation/fractionation and MS sensitivity levels have revolutionized comprehensive shotgun proteome analysis. It appears that a single technology or method alone cannot address issues associated

with the dynamic range of protein concentrations in a given proteome. We have emphasized the needs of both appropriate experimental designs as well as correct interpretations of large protein datasets.

Financial & competing interests disclosure

The authors have no relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

No writing assistance was utilized in the production of this manuscript.

References

- 1 Barnea E, Sorkin R, Ziv T *et al.* Evaluation of prefractionation methods as a preparatory step for multidimensional based chromatography of serum proteins. *Proteomics* 5, 3367–3375 (2005).
- 2 Chong PK, Wright PC. Identification and characterization of the *Sulfolobus solfataricus* P2 proteome. *J. Proteome Res.* 4, 1789–1798 (2005).
- 3 Gan CS, Reardon KF, Wright PC. Comparison of protein and peptide prefractionation methods for the shotgun proteomic analysis of *Synechocystis* sp. PCC 6803. *Proteomics* 5, 2468–2478 (2005).
- 4 Mastroleo F, Leroy B, Van Houdt R *et al.* Shotgun proteome analysis of *Rhodospirillum rubrum* S1H: integrating data from gel-free and gel-based peptides fractionation methods. *J. Proteome Res.* 8(5), 2530–2541 (2009).
- 5 Vaezzadeh AR, Deshusses JM, Waridel P *et al.* Accelerated digestion for high-throughput proteomics analysis of whole bacterial proteomes. *J. Microbiol. Methods* 80(1), 56–62 (2010).
- 6 Hu S, Xie Y, Ramachandran P *et al.* Large-scale identification of proteins in human salivary proteome by liquid chromatography/mass spectrometry and two-dimensional gel electrophoresis-mass spectrometry. *Proteomics* 5(6), 1714–1728 (2005).
- 7 Wolff S, Otto A, Albrecht D *et al.* Gel-free and gel-based proteomics in *Bacillus subtilis*, a comparative study. *Mol. Cell. Proteomics* 5, 1183–1192 (2006).
- 8 Keller M, Hettich R. Environmental proteomics: a paradigm shift in characterizing microbial activities at the molecular level. *Microbiol. Mol. Biol. Rev.* 73(1), 62–70 (2009).
- 9 Nesvizhskii AI, Aebersold R. Interpretation of shotgun proteomic data: the protein inference problem. *Mol. Cell. Proteomics* 4, 1419–1440 (2005).
- 10 Alves P, Arnold RJ, Novotny MV *et al.* Advancements in protein identification from shotgun proteomics using predicted peptide detectability. *Pac. Symp. Biocomput.* 12, 409–420 (2007).
- 11 Zhang B, Chambers MC, Tabb DL. Proteomic parsimony through bipartite graph analysis improves accuracy and transparency. *J. Proteome Res.* 6, 3549–3557 (2007).
- 12 Li YF, Arnold RJ, Li Y *et al.* A Bayesian approach to protein inference problem in shotgun proteomics. *J. Comput. Biol.* 16(8), 1183–1193 (2009).
- 13 Cox J, Mann M. Computational principles of determining and improving mass precision and accuracy for proteome measurements in an Orbitrap. *Am. Soc. Mass Spectrom.* 20(8), 1477–1485 (2009).