

A mini review: proteomic analysis, a post-genomic approach

Jean-François CHICH*

Unité de Recherche en Biochimie et Structure des Protéines, INRA, Domaine de Vilvert,
Bâtiment 526, 78352 Jouy-en-Josas Cedex, France

Abstract — We present here the proteome tool and its current state. Genomics and transcriptomics are briefly described and their limitations are discussed. The various aspects of proteomic approach are then presented and discussed. Proteomics is defined as the expressed complement of a genome. Technical aspects such as two dimensional gel electrophoresis, mass spectrometry and searches in generalist and EST databases changed the protein identification process. Proteomics can be divided into a systematic approach, or cell map proteomic and a pragmatic approach, or study of changes in protein expression.

proteomic / database / mass spectrometry

Résumé — **L'analyse protéomique, une approche post-génomique.** Nous présentons ici l'outil protéomique et « l'état de l'art ». Génomique et transcriptomique sont brièvement décrites ainsi que leurs limites. Les différents aspects de l'approche protéomique sont ensuite présentés et discutés. La protéomique est définie comme le complément protéique d'un génome. Les aspects techniques comme l'électrophorèse bidimensionnelle, la spectrométrie de masse dans les banques généralistes ou les banques d'EST ont changé le processus d'identification des protéines. La protéomique peut être subdivisée en une approche systématique, ou protéomique de cartographie et une approche pragmatique, ou étude de l'expression des protéines.

protéomique / banque de données / spectrométrie de masse

1. INTRODUCTION

Genomics is characterized by systematic acquisition of structure and function data, using high throughput technics. The aim of

this approach is to transform sequence data into biological data. However, a genome is a static reflection of a biological system. Knowing a gene sequence does not automatically signify knowledge of the function

* Tel.: (33) 1 34 65 21 48; fax: (33) 1 34 65 21 63; e-mail: chich@jouy.inra.fr

of encoded protein and regulations of the gene.

The goal of techniques known as “transcriptome” is the study of mRNA of a cell, a tissue or an organism [21]. Transcriptome reflects the qualitative and quantitative properties of mRNA and allows the monitoring of the gene expression in given physiological conditions. Approaches using DNA chips are very powerful and allow the study of the expression of several thousands of genes. However, no strong correlation exists between the amount of a given protein and its transcripts [5].

Proteome was defined for the first time in 1995 as the “total protein complement of a genome” [22] or, in the case of a pluricellular organism, as the protein complement of a tissue. In other words, proteomics is the study of the properties of proteins (expression level, post-translational modifications, interactions etc.) on a large scale to obtain an integrated view of normal or pathological cellular processes or interactions at the protein level. Now some genomes are fully known, the question is how cellular products interact. The first fully sequenced genome is from *Haemophilus influenzae* [4]. About twenty genomes have been fully sequenced and at least 30 will become available within 1 or 2 years [19]. It quickly became clear that numerous proteins encoded by numerous genes did not have defined functions and for most of them, functions were attributed according to homologies. For example, on the 1 743 identified genes from *Saccharomyces cerevisiae*, 31% are homologous to those from other species and 43% encode for unknown function proteins [3].

In short, DNA sequence by itself is not able to predict: (1) if and when mRNA are translated; (2) relative concentration of the gene products; (3) the nature of post-translational modifications; (4) the effects of a K.O. or an overexpressed gene; (5) the phenotype of various multigenic phenomena like stress, aging, drug administration, etc.

[10]. A systematic post-genomic study is necessary and it is just what proteomics offers.

Currently, applications of this “new science” are in the clinical and pharmaceutical fields; it is possible to correlate rapidly protein modification levels with a drug effect [2], hormone effect or a stress effect [6]. However proteomics is in its childhood and is likely to develop in the near future.

2. PROTEOMICS

The technique allowing the separation on an acrylamide gel of proteins according to their pI and molecular weight was first described by O’Farrell [13]. This technique is the most powerful one to separate a complex mixture of proteins. Up to 2 000 proteins can be separated in a reproducible manner. This revelation method is used to visualize and quantify proteins. But visualization is not an identification and up till recently this last one was performed, after a western blott, by Edman sequencing and search in databanks. However this slow and expensive method does not allow a systematic identification of proteins.

This obstacle was recently overtaken thanks to the development of very sensitive mass spectrometry techniques and to the increase of entries in protein databanks. In the near future, protein identification by mass spectrometry will become a routine and high throughput job.

Proteome research encompasses nothing novel but is an extension of techniques which have been used for some two decades [10]. In fact, analysis of the “proteic complement of a genome” became possible thanks to the improvement of ancient techniques: (1) reproducibility of separation techniques by the apparition of immobilized pH gradients; (2) powerful image analysis software allowing the construction of polypeptidic maps and the comparison between them; (3) sensitivity increase and

easier use of mass spectrometer to analyze peptides and proteins; (4) taking account of the necessity to obtain information on proteins in addition to genomics.

2.1. Dual approach

The definition of proteomics cannot help to state its goals and potential applications. It seems possible to describe two approaches [1, 7]. The first is a systematical one and consists of studying global proteomic expression. It is a systematic identification of given cell proteins in a reference physiological state. Such information could be useful as a databank. The second is a pragmatic one and consists of establishing bidimensional maps of the cellular expression in given conditions and allows the study of cellular pathways and their modification by a drug, a biological stimulus etc.; proteic targets will be identified. However both approaches are complementary and provide a global vision of cellular physiology.

2.2. The technology

2.2.1. Protein separation

The first step is the preparation of proteins. It is impossible to give general rules because any preparation has an influence on the results. However the quantity of salts in the sample should be kept very low to limit artefacts. To eliminate them, some methods with their advantages and disadvantages are available. Thus, several techniques should be tested prior to choosing one of them.

Low amount proteins (10–1 000 copies per cell) can be concealed by constitutive proteins (>10 000 copies per cell). In this case, one should envisage a selective enrichment of these proteins by organites purification steps, chromatographies... IEF gels can be selective by themselves and special care should be taken for highly basic, hydrophobic and low soluble proteins [15].

The chosen pH gradient could also influence the final result. The gradient should be chosen according to the pH area where proteins precipitate. However, first analysis should be made on a wide pH gradient (3 to 10).

Usually, proteins are visualized by radiolabelling or by staining with Coomassie brilliant blue or silver. Recently, fluorescent stains [18] have been used. They are very sensitive and have a wide dynamic range. Ideally, the staining intensity should be independent of the nature of the protein and should correlate with the amount of protein. Subsequent steps like mass spectrometry require that stains do not interfere. For example if gels are silver stained, cross-linking agents should be avoided [16].

Commercial software allows theoretically an automated spots detection but in fact, it is necessary to work from 4 to 8 h on a gel and is therefore incompatible with high throughput analysis. This problem should be resolved very quickly.

Ultimately, it should be possible to have available robots, which cut out spots without manual intervention, to increase analysis speed and especially to avoid a contamination by keratin.

2.2.2. Protein identification

Sequencing according to Edman is no longer used in proteomics. However, it remains the reference method to characterize a N-terminus sequence of a protein. Disadvantages are numerous, for instance it necessitates a supplementary step of blotting on a PVDF membrane after the second dimension.

Currently, the most used method for protein identification and characterization of post-translational modifications is mass spectrometry.

The MALDI-TOF technique (Matrix-Assisted-Laser-Desorption-Ionization-Time-Of-Flight) is mainly used for the PMF (Peptide Mass Fingerprinting [8, 23]). Briefly,

a proteolytic digest (usually using trypsin) is performed on an acrylamide piece corresponding to a spot. Peptides mass is then measured and compared to a list of in silico generated peptides, using the same proteolytic enzyme in a protein databank. The measured mass is so accurate (10 ppm or 0.001%) that no further information is necessary to identify a protein. If sequence information is required (missing data in the bank, incompletely sequenced genome, post-translational modifications, etc.), two techniques, CID (Collision Induced Dissociation [11]) or PSD (Post-Source Dissociation [17]) can be used on this kind of mass spectrometer. They further fragment the peptides, smaller ions are obtained and the molecular weight difference between these ions gives a N-ter or C-ter sequence. Other more efficient mass spectrometers can be used in the MS/MS configuration.

Without sequence data, search in databanks can be more difficult and even not successful. A peptide sequence tag can be used to search in proteic and EST (Expressed Sequence Tag) databanks. A sequence tag can be obtained by additional fragmentation of peptides. It is necessary to obtain several tags for a single protein to limit errors. If the genome is fully sequenced, 1 to 3 tags by protein are sufficient for identification.

If several other mass spectrometry techniques can be used [24], they allow only a limited number of analyses per day. A great advantage of the MALDI-TOF besides its precision is that it can be used for a high throughput (>1 000 samples per d), using PMF.

Mass spectrometry techniques allow the detection of very low amounts of peptides, of the order of about ten femtomoles of protein. To analyse post-translational modifications, quantities of the order of picomole are enough [12]. Sensitivity of these apparatus suggest the possibility of, in the near future, the analysis of low copy number proteins [1]. Actually, sensitivity is probably

not limiting but the signal/noise ratio becomes more important. Thus, in the future, it should be possible to improve the conditions of sample preparation, thanks to an increase in automation.

2.3. Proteome use

We have seen at the beginning of this paper that proteomics can be considered from two complementary angles of view.

2.3.1. Systematic approach

The aim of this approach is the identification of all proteins of a cell or a tissue or a species and to add annotations to each of them (see for example [9, 14, 20]). This approach is complicated by the fact that it is difficult to define which proteome should be selected as reference for the databank. As previously seen, if a genome is static, a proteome is far from static. Events like activation, disease etc. can change deeply the proteomic phenotype.

Thus a proteomic databank should contain all the information on the nature and status of biological material and also parameters used for its identification in a database. Its post-translational modifications, subcellular localization, molecules interacting with it and evolution of all this data should also be indicated.

Currently, to our knowledge, no fully complete database is available. However some are under construction (<http://www.expasy.ch/ch2d/>, <http://pcsf.brcf.med.umich.edu/eco2dbase/> or <http://www.bio-mol.unisi.it/2d/2d.html>, or <http://microbio2.biologie.uni-greifswald.de:8880/sub2d/pub/sub2d.ci>, etc.).

2.3.2. Pragmatic approach

Here, the aim is to compare two physiological states. In a classical biological test, a single variable is measured, for example, the activity of an enzyme. If proteomics is

used as a biological test, one can measure several variables (expression level, evolution of post-translational modification, etc.) on all the proteins of the sample at the same time. This approach allows the study of biological systems and metabolic pathways with a maximum of controlled variables [2, 6]. An additional interest is that target proteins can be easily identified by mass spectrometry.

This approach was applied successfully in toxicological or medical fields. On the other hand, one limit is that all proteins are not inevitably detected for reasons inherent to experimental context.

3. CONCLUSION

Proteomics has already shown a good capability to complement technologies used to study molecular mechanisms of the cell. Molecular cellular biology needs that information on DNA and RNA to be completed by information at the transcription level: it is what proteomics offers. It is at its beginning and new, fast and more efficient strategies will soon be available.

It seems necessary to place proteomics in a larger context. It is more a conceptual change than a technical one. It is not one more isolated biological discipline; it is integrated in the context of the integrative biology which is looking to find a non reductionist approach, a more global one from gene to physiology.

REFERENCES

- [1] Blackstock W.P., Weir M.P., Proteomics: quantitative and physical mapping of cellular proteins, *Trends Biotech.* 17 (1999) 121–127.
- [2] Cash P., Argo E., Ford L., Lawrie L., McKenzie H., A proteomic analysis of erythromycin resistance in *Streptococcus pneumoniae*, *Electrophoresis* 20 (1999) 2259–2268.
- [3] Fey S.J., Nawrocki A., Larsen M., Görg A., Roepstorff P., Skews G.N., Williams R., Larsen M.P., Proteome analysis of *Saccharomyces cerevisiae*: a methodological outline, *Electrophoresis* 18 (1997) 1361–1372.
- [4] Fleischmann R.D., Adams M.D., White O., Clayton R.A., Kirkness E.F., Kerlavage A.R., Bult C.J., Tomb J.F., Dougherty B.A., Merrick J.M., McKenney K., Sutton G., Fitz-Hugh W., Fields C., Gocayne J.D., Scott J., Shirley R., Liu L., Glodex A., Kelley J.M., Weidmen J.F., Phillips C.A., Spriggs T., Hedblom E., Cotton M.D., Utterback T.R., Hanna M.C., Nguyen D.T., Saudek D.M., Brandon R.C., Fritchmann J.L., Fuhrmann J., Geoghagen N.S.M., Gnehm C.L., McDonald L.A., Small K.V., Fraser C.M., Smith H.O., Venter J.C., Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd, *Science* 269 (1995) 496–512.
- [5] Gygi S.P., Rochon Y., Franza B.R., Aebersold R., Correlation between protein and mRNA abundance in yeast, *Mol. Cell. Biol.* 19 (1999) 1720–1730.
- [6] Hartke A., Bouché S., Giard J.-C., Benachour A., Boutibonnes P., Auffray Y., The lactic acid stress response of *Lactococcus lactis* subsp. *lactis*, *Curr. Microbiol.* 33 (1996) 194–199.
- [7] Haynes P.A., Gygi S.P., Figeys D., Aebersold R., Proteome analysis: Biological assay or data archive?, *Electrophoresis* 19 (1998) 1862–1871.
- [8] Henzel W.J., Billeci T.M., Stults J.T., Wong S.C., Grimley C., Watanabe C., Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases, *Proc. Natl. Acad. Sci. USA* 90 (1993) 5011–5015.
- [9] Hermann T., Wersh G., Uhlemann E.M., Schmid R., Burdovski A., Mapping and identification of *Corynebacterium glutamicum* proteins by two-dimensional gel electrophoresis and microsequencing, *Electrophoresis* 19 (1998) 3217–3221.
- [10] Humphery-Smith I., Cordwell S.J., Blackstock W.P., Proteome research: complementarity and limitations with respect to the RNA and DNA worlds, *Electrophoresis* 18 (1997) 1217–1242.
- [11] Hunt D.F., Buko A.M., Ballard J.M., Shabanowitz J., Giordani A.B., Sequence analysis of polypeptides by collision activated dissociation on a triple quadrupole mass spectrometer, *Biomed. Mass Spectrom.* 8 (1981) 397–408.
- [12] Neubauer G., Mann M., Mapping of phosphorylation sites of gel-isolated proteins by nano-electrospray tandem mass spectrometry: potentials and limitations, *Anal. Chem.* 71 (1999) 235–242.
- [13] O'Farrell P.H., High resolution two-dimensional electrophoresis of proteins, *J. Biol. Chem.* 250 (1975) 4007–4021.
- [14] Quadroni M., James P., Dainese-Hatt P., Kertesz M.A., Analysis of global responses by protein and peptide fingerprinting of proteins isolated by two-dimensional gel electrophoresis. Application to the sulfate-starvation response of *Escherichia coli*, *Eur. J. Biochem.* 266 (1999) 986–996.

- [15] Rabilloud T., Adessi C., Giraudel A., Lunardi J., Improvement of the solubilization of proteins in two-dimensional electrophoresis with immobilized pH gradients, *Electrophoresis* 18 (1997) 307–316.
- [16] Shevchenko A., Wilm M., Vorm O., Mann M., Mass spectrometric sequencing of proteins from silver-stained polyacrylamide gels, *Anal. Chem.* 68 (1996) 850–858.
- [17] Spengler B., Kirsch D., Kaufmann R., Jaeger E., Peptide sequencing by matrix-assisted laser-desorption mass spectrometry, *Rapid Commun. Mass Spectrom.* 6 (1992) 105–108.
- [18] Steinberg T.H., Haugland R.P., Singer V.L., Applications of SYPRO orange and SYPRO red protein gel stains, *Anal. Biochem.* 239 (1996) 238–245.
- [19] Tomb J.F., A panoramic view of bacterial transcription, *Nat. Biotechnol.* 16 (1998) 23.
- [20] van Bogelen R.A., Schiller E.E., Thomas J.D., Neidhardt F.C., Diagnosis of cellular states of microbial organisms using proteomics, *Electrophoresis* 20 (1999) 2149–2159.
- [21] Velculescu V.E., Zhang L., Zhou W., Vogelstein J., Basrai M.A., Bassett D.E. Jr., Hieter P., Vogelstein B., Kinzler K.W., Characterization of the yeast transcriptome, *Cell* 24 (1997) 243–251.
- [22] Wasinger V.C., Cordwell S.J., Cerpa-Poljak A., Yan J.X., Gooley A.A., Wilkins M.R., Duncan M.W., Harris R., Williams K.L., Humphery-Smith I., Progress with gene-product of the Mollicutes: *Mycoplasma genitalium*, *Electrophoresis* 16 (1995) 1090–1094.
- [23] Wise M.J., Littlejohn T.G., Humphery-Smith I., Peptide-mass fingerprinting and the ideal covering set for protein characterisation, *Electrophoresis* 18 (1997) 1399–1409.
- [24] Yates III J.R., Mass spectrometry and the age of the proteome, *J. Mass Spectrom.* 33 (1998) 1–19.