Pure Appl. Chem., Vol. 76, No. 4, pp. 829–837, 2004.

Proteomics principles and challenges*

György Marko-Varga‡

Department of Analytical Chemistry, Lund University, Box 124, SE-221 00 Lund, Sweden

Abstract: Proteomics studies rely on high-resolving separation techniques whereby the biological complexity can be unraveled. Both gel-based and liquid-phase separation principles are needed in order to mine deeper into the low abundant area of the proteins expressed in human cell samples. The expression levels that can be reached today with proteomics technology are still only reaching part of all proteins present in complex biological samples. The challenge to be met is not an easy task and will, in many aspects, be determined by the speed and success whereby the analytical field can make milestone progress. Examples will be given from our research group on proteomics studies, as well as principles and challenges within this research area.

INTRODUCTION

The complete human genome project (HUGO) was officially completed in the first quarter of 2003. With about 99 % of the sequencing done, actually almost three years ahead of the planned deadline and way below the expected budget given for the project initiative, the HUGO project was completed. For most of the whole genome, the function of a large part of the proteins remains unknown. Mapping of the human genome, applying advanced mathematical algorithms and mining the genomic databases linked to the fast progress within the mass spectrometry field, has provided important contributions to the principle foundations of protein expression analysis research.

Within the post-genomic area, attention is given to the proteomics research area, where the human proteome initiative (HUPO) is the organization that actually drives the program to map protein expression studies within various disease areas [1]. Currently, there are 49 worldwide research groups that are involved in the generation of the first profiling study project, establishing profile data from human blood samples [2]. The human plasma proteome is the most complex proteome due to the inter-, and intra-channeling into the tissues of the various organs that serve as the subsets. The wide dynamic range, close to nine orders of magnitude ranging from the high abundant region with albumin and immunoglobulins as examples to be compared with low copy number proteins in this biofluid such as cytokines, transcription factors, etc. The tissue leakage that is entering into the cardiovascular system is mostly proteins that have a functional role within the cell. Release into the blood stream is a mechanism whereby the cell can operate its cell cycles and necessary metabolism whereby cell damages and cell death events become measurable in the human blood compartment [3].

Proteomics is the research area revealing the temporal dynamics of proteins expressed in a given biological compartment at a given time. The definition has until recently been covering proteins as gene products. Lately, there has been an alteration of the proteomics definition to include not only gene products, but also structural alteration of these gene products occurring in cellular metabolisms and turnover (i.e., post-translational modifications) [4]. The identity annotation of human proteins is increasing rap-

*Plenary lecture presented at the Southern and Eastern Africa Network of Analytical Chemists (SEANAC), Gaborone, Botswana, 7–10 July 2003. Other presentations are published in this issue, pp. 697–888.

‡E-mail: gyorgy.marko-varga@analykem.lu.se

idly, it can be followed on the Internet within public databases. Still, there is a large need for protein function annotations. Predicting putative functions of proteins is made by bioinformatic tools [5,6]. Recently, an interesting work was published by Kanemaki et al. [7], who made functional proteomic identification of DNA replication proteins by induced proteolysis in vivo.

The qualitative dynamic range within biofluids, cell supernatants, and tissues is a challenging analytical task to take on. The low-molecular-weight region (typically, the area below 20 kDa) is not very well identified. It has to do with the fact that smaller proteins and poly-peptides have fewer peptides to use for sequence identity using mass spectrometry [8]. In earlier work, we developed sample preparation techniques dedicated to complex protein samples where we efficiently removed high-molecular-weight proteins—typically, bio-macro molecules >25 kDa [9]. This allows the specific low-molecular-weight fraction to be analyzed in an efficient way [9].

Two-dimensional gel electrophoresis is the most powerful analytical separation technique for proteins, already introduced in the early 1970s [10,11]. The large-size globular structures of proteins do not allow themselves for chromatographic high-resolution separations due to the unfavorable diffusion coefficients in the separation mechanism. However, by making protein digests out of the protein samples, the task becomes easier since now we are looking at the separation of peptide mixtures. On the other hand, this is highly advantageous using chromatographic separation mechanisms. The price that needs to be paid is that the analyte number will increase by about a factor of ten to several hundreds. That means that the size of the actual sample complexity we are looking at can be in hundreds of thousands of peptides in a biological sample [12,13]. Proteomics today is playing a significant role in pharmaceutical R&D and within the clinic that stretches in the drug development process from early discovery to clinical testing.

BASIC CONCEPTS AND ANALYTICAL PROCEDURES

There seem to be a difference in-between m-RNA regulation and the actual synthesis of proteins [14,15]. Figure 1 illustrates the link in-between the DNA, RNA, and protein. While the genome is identical in all bodies, the protein levels will vary in both cell type and time. The kinetics within proteomics studies is of vital importance and might be the missing link to why the correlative regulation factors of m-RNA and protein are found to be different. Modification of the protein structure is a part in the metabolism of the cell's natural life cycle. A large number of post-translational modifications (PTMs) do occur, such as phosphorylation events. These intracellular mechanisms are of vital importance as they

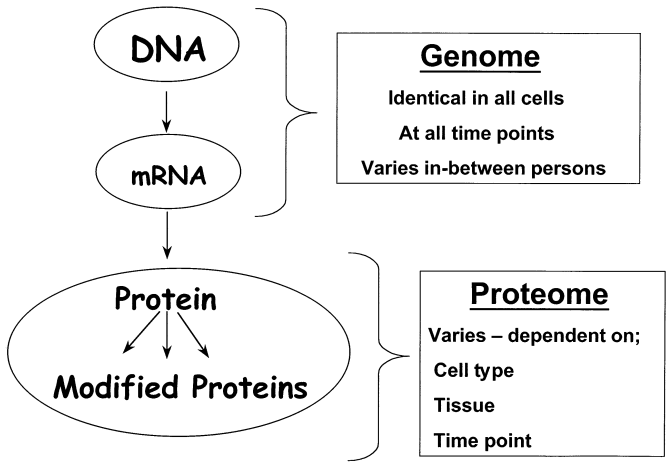

Fig. 1 Schematics of the DNA, RNA, and protein events in the cell.

have a direct link to the biological activity that takes place within the cell. However, most of these PTM activities will not impose a biological effect, such as, for instance, a signaling event.

Diseased-linked alterations are most commonly observed in the tissue where morphological changes occur. The degree of severity of the disease will in many cases be linked to, for instance, the decline in organ function and capacity. Figure 2 shows the sample handling steps that often are undertaken in clinical studies from which the proteomics expression profiling is performed. The tissue as such can be analyzed directly by solubilization where a total expression read-out will be generated from the tissue sample as shown in Fig. 1A. There is also a possibility of isolating cells from the tissue by using favorable cultivation conditions and thereby incubating these target cells under proliferative conditions (see Fig. 2B). The human cells will increase in number, and this makes the protein study easier in terms of sample availability (see Fig. 2C). It is of mandatory importance that viability controls of the cells are made, as well as control assays, which ensures the phenotype. Primary human cells as cell lines is another alternative to making in vitro studies with subsequent expression analysis (see Fig. 2D).

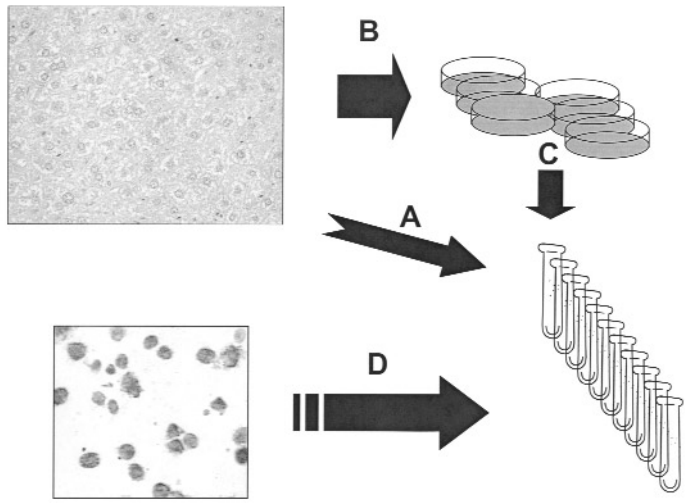

Fig. 2 Illustration of tissue and cell handling principles and methodologies in proteomic studies including: (A) tissue isolation and subsequent sample preparation and protein analysis; (B) cell isoaltion and cultivation from tissue; (C) proteomics study from cultivated cells; and (D) isolated cell oragenelle proteic studies.

The enormous challenge that the proteomics field holds, and where the "holy grail" still needs to be found and defined, can be summarized as simple as not sufficient sensitivity. Figure 3 illustrates the daily struggle and challenge that the research area is working on and developing, where the major limitation is the lack of a protein PCR technology. This has certainly put much pressure on both academic and industrial groups to find new analytical technologies and methodologies that can circumvent this limitation. The large dynamic range, which reaches up to nine orders of magnitude, is also not a trivial challenge to meet since it holds a high degree of sample composition variation.

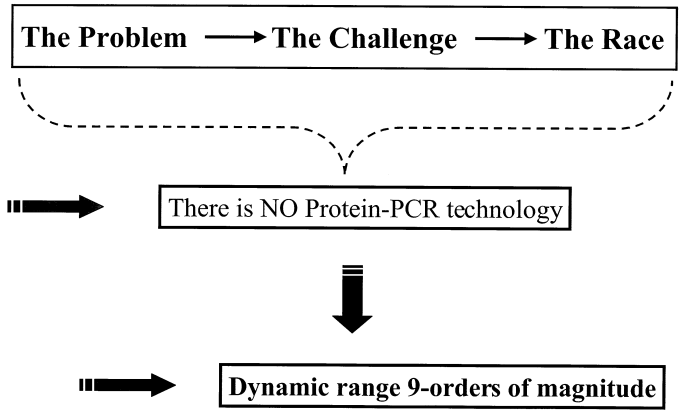

Fig. 3 Schematic overview of the proteomics challenges.

PROTEIN EXPRESSION SEPARATION

Gel-based separations

One widespread proteomics strategy involves the use of 2D gel electrophoresis followed by enzymatic degradation of isolated protein spots and identification using mass spectrometry.

Gel-based separations are and have been extremely efficient techniques to isolate and separate proteins utilizing both 1D and 2D polyacrylamide gel electrophoresis (PAGE) [16,17].

One major reason that this separation technology has been so successful is the resolving power with a fast profiling image fingerprint generation. The development and introduction of laser desorption ionization MS; MALDI-TOF was extremely timely. Peptide mass fingerprinting is a very powerful means to identify proteins, lately, the ability to sequence by MALDI has added on to the abilities to sequence peptides, and also to analyses of structural alterations [18]. We have developed cell models and protocols to analyze cell samples taken at various time points [19].

Figure 3 illustrates the protein expression found in a human cell sample where two staining techniques are being compared. The left-hand image shows the resulting gel where silver staining was applied, whereas the right-hand gel was stained by Sypro Ruby, a fluorescent stain. The fluorescent staining was a post gel staining method. It is clear upon comparison of the two gels that to a large extent they resemble each other. However, there are also clear differences in-between the two images, especially on the higher-molecular-weight regions where the specificity of silver and Sypro Ruby is obvious. Surely this difference in methodology holds a limitation upon comparison of these expression data made from the same sample. On the other hand, it can be made useful by reaching complementary and extended information of the protein composition of the sample. The improved separation resolution of these gels that were run in-between PI 3–10, the high PI area around pH 8 and above shows improved resolution using the fluorescent staining. This effect is to a major extent in most cases determined not by the staining technique as such, but rather on the running conditions of the SDS-PAGE gel. Another advantage that Sypro Ruby staining offers is improved MS analysis compatibility.

Something that is striking when working with 2D gels is the large number of protein spots that can be easily be separated. In many cases, thousands of proteins are resolved and easily annotated on the gel. Considering the fact that, for instance, a cell sample will hold not only these, but certainly a factor of 10–100 higher is a challenge that is not easy to address. Figure 4 shows a calculation taken as an example where a 10 fmol protein level is considered for a successful MS identification. It is clear then that if a protein is expressed at high abundance, 106, thousand of cells are sufficient for a positive identification. On the other hand, if the protein is low in abundance (e.g., 10 copies per cell), in these situ-

ations hundreds of millions of cells are required in the proteomics experiment to make a positive identification. The in-between abundance regions are shown in Fig. 5.

Closing in on the high-molecular-weight region, where connective tissue proteins are annotated in these cell samples [20,21], also shows several examples of the isomeric resolution that 2D PAGE offers, where the protein spot trains are clearly seen within the magnified part of the gel image (gel below). These vertical series of proteins closely positioned to one another on the gel is examples of post-translational modifications of proteins. In many cases, the protein structure modification is due to different phosphorylation states that are being mapped, where one identified protein can have a large number of isoform states, up to the 10s and 20s.

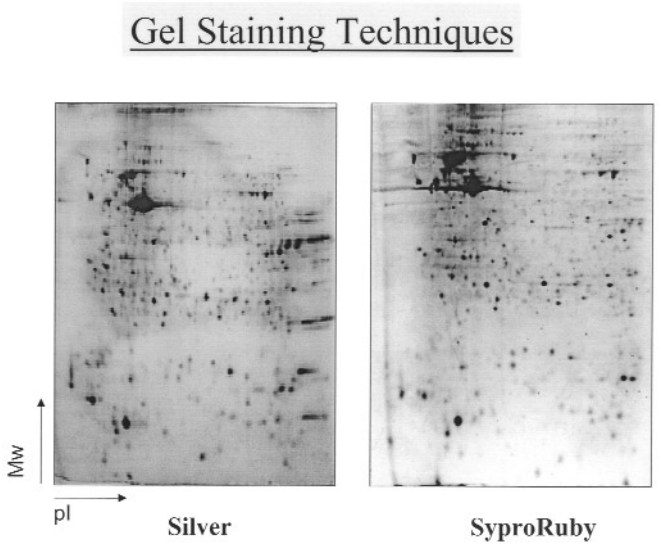

Fig. 4 2D gel electrophoresis separation image generated from human cell lysate using a PI range of 3–10, and with a molecular-weight window of 7–145 kDa. Post-gel image staining using fluorescent stain, Sypro Ruby, and silver staining.

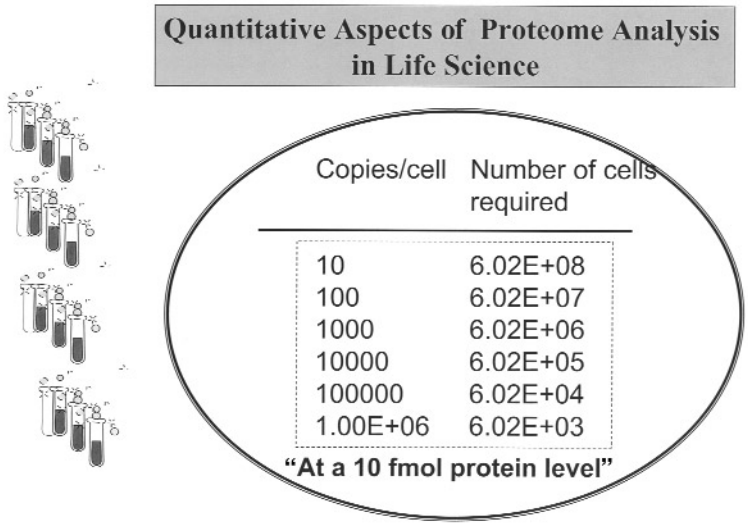

Fig. 5 Overview of quantitative aspects of protein identification and cell numbers needed for a given expression level.

LIQUID CHROMATOGRAPHY SEPARATIONS

Proteomics studies are also being made using liquid chromatography. Just as 2D separation generates the high separation efficiency in gels, several dimensions are required using LC. In most cases, at least two dimensions are being applied to complex biological samples where most often electrostatic separation mechanisms are being combined with hydrophobic separations. The great advantage using LC is that the interface in-between the first and the second dimension can be made online, as presented in Fig. 6. In this approach, a continuous real-time protein analysis can be made. While one sample is being analyzed in the set-up presented in Fig. 6, the second column (1^{st} dimension) is conditioned. By the time the separation is completed in the first column, the next fraction is introduced into column 2, while column 1 is equilibrated, made ready for the next sample to be introduced. Figure 7 shows the analysis process whereby the separations are made using coupled column LC described in Fig. 6 and the resulting MS spectra generated using both electrospray and MALDI ionization principles [22,23]. The database registration and query path of the annotation procedure is an important part of the bioinformatic work within proteomics. The peptide sequence identities are made where a scoring of the protein identity is made, reflecting the statistical significance whereby the identity is determined and the likelihood that it is the correct protein identity.

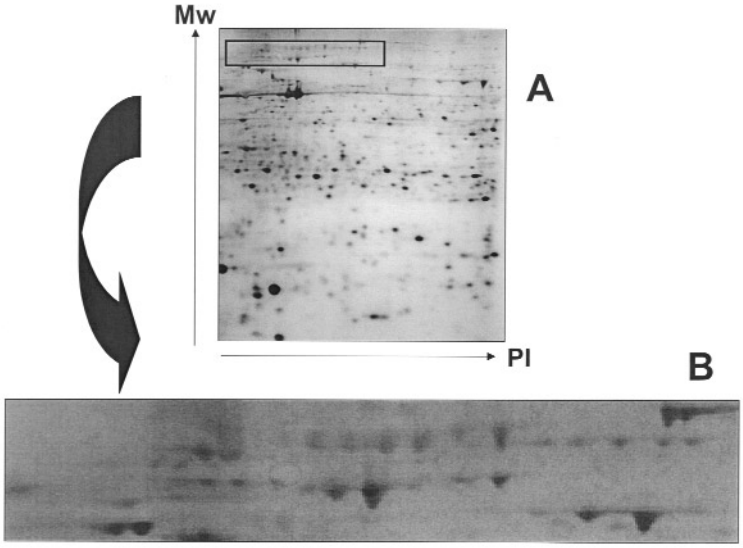

Fig. 6 (A) Resulting gel image fluorescent post-gel staining of a human cell sample with the same gel running conditions as in Fig. 4. (B) The magnified part of the gel image capture demonstrates the high-density and high-quality expression map from part of the proteome expressed in these cells.

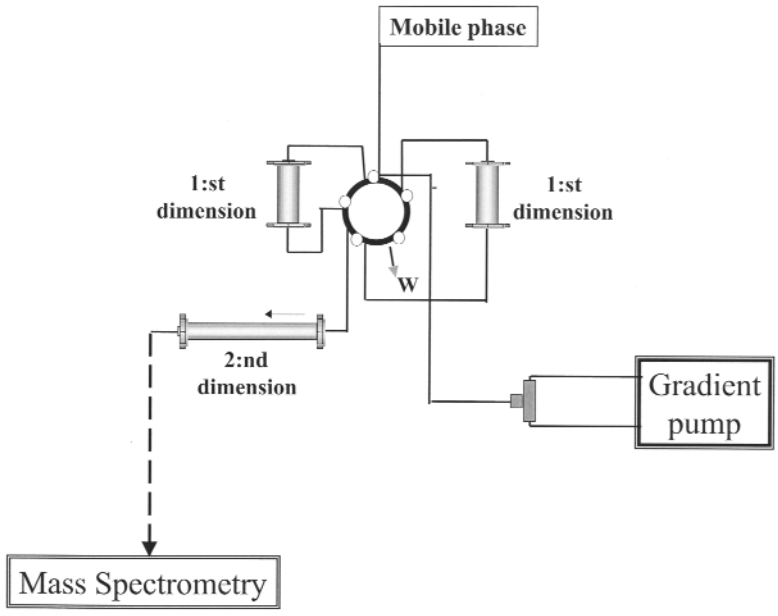

Fig. 7 Chromatographic online configurations and set-up of 2D LC system for proteomics studies.

CONCLUSIONS

Performing 2D PAGE separations allows the proteome analysis to be made in a protein form where the size as well as the charge of each respective protein can be determined. It is a great advantage when it comes to the mass spectrometric identification since the database query using publicly available databases (e.g., SWISSPROT, NCBI, or EMBLE) will make the identification more probable. The statistical significance of an identity hit will be higher in cases where the molecular size of the proteins from the 2D gel being identified is accurate. The PI value is additionally a useful physical property to be used, however, this is in many cases altered due to protein modifications such as phosphorylation or sulfatation. The resolution of 2D PAGE-separated proteins is in the order of 1500–5000 as shown in this work, although the latter high number of protein spots are generally only obtained when metabolic labeling is made. The really big challenge and disadvantage of the 2D PAGE approach is that the gel image annotations made can be performed highly accurately with the high resolving power, but cannot be made ID-approved by mass spectrometry. The main reason for failure of identification is simply due to the fact that the protein spot that is excised and digested has an unfavorable digestion kinetics. The Michaelis–Menten kinetics works against you in situations where the substrate levels are low (i.e., the protein spots are faint). It is also a difference in resolution running straightforward 2D PAGE studies and post-gel stainings in comparison to metabolically labeled cells that are separated by 2D PAGE [19]. It should also be stressed that its not really the same experiment that is being performed since the methionine actually was added to the cell culture during the incubation and incorporated the amino acid. Newly synthesized proteins can be analyzed in this way, while the post-gel staining technique does not distinguish in-between newly synthesized and proteins already expressed with a long half-life. The combination of the two techniques is really useful in cell studies where detailed protein mechanisms and processes are being investigated.

Figure 8 captures the most important decision-making steps within the proteomics work process that is applicable in most studies. It is important to make realistic starting points for an expression experiment. These should be realistic and, above all, clearly defined. One should also have an understanding of the abundance level on which a given set of proteins is to be analyzed. The experimental ef-

fort will be directly linked to the abundance level. It is also a good rule of thumb to generate preliminary test-analysis experiments to get a feeling for the quality of information that can be obtained with that particular biological material. The continuity and success of the study is, in many cases, determined at this point. The large separation effort either by 2D gels and/or multidimensional LC needs appropriate analysis and subsequent annotations and identities. The differential regulation of proteins in such studies then needs to be verified in functional assays that can be either in vitro or in vivo assays.

The mining of genomic data to identify function is a great challenge to the life science field where in silico developments are crucial. Complementary areas such as transgenic models, gene silencing, and bioinformatics are currently expanding the field of functional genomics. The genomics link to the functional proteomics field will be the future direction that will drive the functional aspects of proteomics [24].

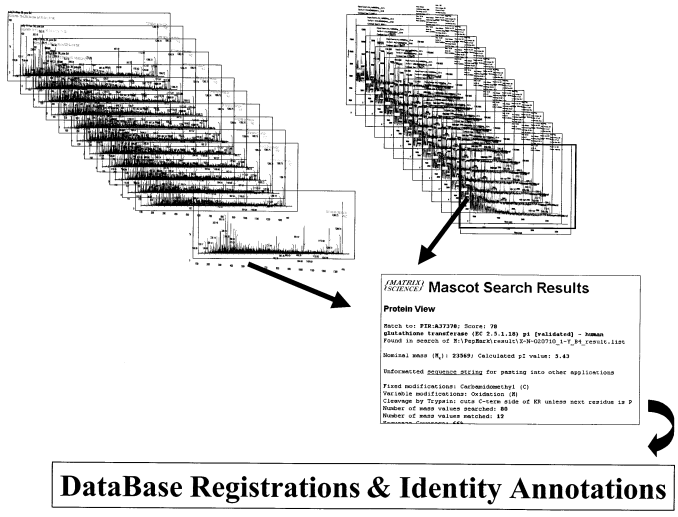

Fig. 8 Strategy for MALDI and electrospray MS identification of protein fractions isolated by multidimensional chromatography where both MS platforms are aligned to a software search engine (MASCOT) from where databases are queried and built.

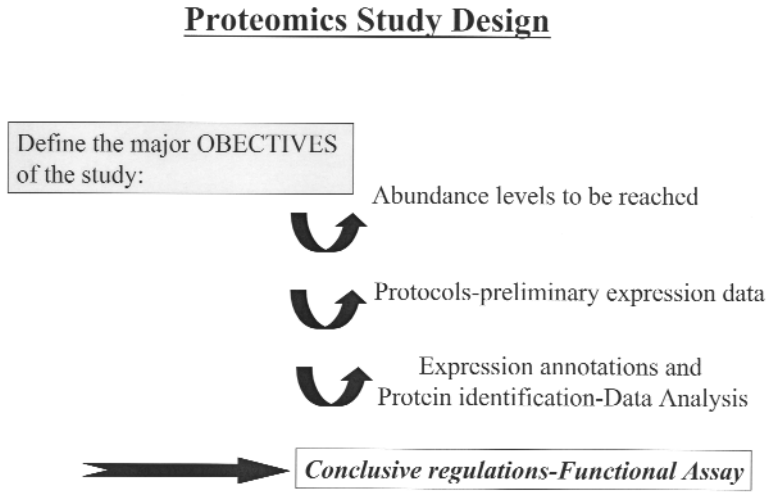

Fig. 9 Summary of the proteomics study design linked to the proteomics process.

REFERENCES

1. <http//www.HUPO.org>
2. S. Hanash. Oral presentation at BIO-2003 Symposia, Washington DC, 28 June 2003.
3. N. L. Anderson and N. G. Anderson. *Mol. Cell. Proteomics* **1**, 845 (2002).
4. R. Aebersold and M. Mann. *Nature* May **1**, 222 (2003).
5. M. Riley. *Microbiol. Rev*. **57**, 862 (1991).
6. L. J. Jensen, R. Guipta, H. H. Staerfeldt, S. Brunak. *Bioinformatics* **19**, 635 (2003).
7. M. Kanemaki, A. Sanchez-Diaz, A. Gambus, K. Labib *Nature* **433**, 720 (2003).
8. G. Marko-Varga and T. Laurell. *Proteomics* **4**, 6 (2002).
9. G. Marko-Varga. *Talanta.* Submitted for publication.
10. P. H. O'Farrell. *J. Biol. Chem.* **250**, 4007 (1975).
11. J. Klose. *Humangenetik*. **26**, 231 (1975).
12. S. P. Gygi, B. Rist, S. A. Gerber, F. Turacek, M. H. Gelb, R. Aebersold. *Nat. Biotechnol*. **17**, 884 (1990).
13. S. P. Gygi, B. Rist, T. J. Griffin. *J. Eng. Aebersold. Proteome. Res*. **1**, 47 (2002).
14. N. G. Anderson, A. Matheson, N. L. Anderson. *Proteomics* **1**, 3 (2001).
15. G. L. G. Miklos and R. Maleszka. *Proteomics* **1**, 169 (2001).
16. F. Lottspeich. *Angew. Chem., Int. Ed.* **38**, 2476 (1999).
17. T. Rabillou. *Proteomics* **2**, 3 (2002).
18. M. Vestal, S. Martin, P. Juhasz. *Cell. Mol. Proteom.* **3**, 34 (2002).
19. C. Bratt, C. Lindberg, G. Marko-Varga. *J. Chromatogr. A* **909**, 279 (2001).
20. G. Westergren-Thorson, Johan Malmström, G. Marko-Varga. *Electrophoresis* **22**, 1776 (2001).
21. G. Westergren-Thorson, Johan Malmström, G. Marko-Varga. *J. Pharm. Biomed. Anal.* **24**, 815 (2001).
22. K. Wagner, K. Racaityte, K. K. Unger, T. Miliotis, L.-E. Edholm, R. Bischoff, G. Marko-Varga. *J. Chromatogr. A* **893**, 293 (2000).
23. J. Malmström, K. Larsen, L. Malmström, E. Tufvesson, K. Parker, J. Marchese, B. Williamsson, D. Patterson, S. Martin, G. Westergren-Thorson, P. Juhasz, G. Marko-Varga. *Electrophoresis* **24**, 3806 (2003).
24. C. Corty. *Genomics Proteomics* **5**, 32 (2003).