One genome—different proteomes

Not only the genome, but in particular the proteins present – the proteome – determine the appearance and state of a biological organism.

Proteome Analysis: A Pathway to the Functional Analysis of Proteins

Friedrich Lottspeich*

Genomics, the sequencing of whole genomes, is progressing at a constantly increasing rate. Many projects have started, 20 of which have already been concluded, and within a few years the human genome will have also been completely sequenced. Just recently a new milestone was achieved with the decoding of the genome of the helminth *Caenorhabditis elegans*, with 97 million base pairs the largest yet. Its analysis offers promise of information on the approximately 19000 genes that have been found. However, a "decoding" of these data in the true sense of the word is still far removed. What can we learn from genome data? What is the value of these data? Without doubt, genome data and their analysis have for the first time revealed the immense complexity of nature. In place of the dogma "one gene, one protein, one function", an understanding of interwoven regulatory networks has developed. We have learned how to handle large amounts of data. Data bank structures have been constructed, and highly specialized programs for data mining have been or are being developed. Classical protein chemistry has also profited from the genome projects. No longer must every individual amino acid of a large protein be analyzed—in many cases a hopeless task—rather it is sufficient to determine small regions of a protein. However, not everything can be derived from the DNA sequence. What is the function of the gene product, what do the active proteins look like? The answers to these questions are often not set out in the primary sequence. What is the situation at the mRNA level? The most recent developments with cDNA chips have given rise to much hope that changes in mRNA can be rapidly and cost effectively analyzed. Nonetheless, can the abundance of mRNA also characterize the complex relationships which constitute a certain metabolic situation or a pathological status? In part, yes; in a few cases it has been possible to correlate an observation with an altered mRNA pattern. But whenever posttranslational modifications, interactions, or degradation and transport phenomena determine the function of a protein, mRNA can no longer provide any information. Here the only recourse is to turn to the level of the proteins and to investigate directly their type, modifications, and above all their abundance. Proteomics, the quantitative analysis of proteins present in an organism at a certain time and under certain conditions, is a key to functional analysis. In the near future proteomics will determine target search and target selection for both basic research, for example in the unraveling of reaction and regulation networks, as well as for applied research, as in the development of medicines.

Keywords: analytical methods · electrophoresis · mass spectrometry · proteins · proteomes

1. Introduction

The genome projects that are underway worldwide have as their target the sequencing and subsequent analysis of complete genomes. The sequence analyses of many genomes of simple organisms, such as bacteria (*Escherichia coli*, *Bacillus subtilis*, *Helicobacter pylori*, *Haemophilus influenzae*, etc.) and the yeast *Saccharomyces cerevisiae* as a representative of eukaryots, have already been determined.[1] Large genomes, such as those of the human being or of the plant *Arabidopsis thaliana* are also part of a concerted worldwide action; the human genome project will probably be finished within a few years. However, the complete genome of an organism gives only a relatively static overview of the functional potential of an organism and does not describe the immense dynamic process which occurs in a living organism. For example, every somatic cell of the butterfly illustrated in Figure 1 and its caterpillar contains identical genetic information. The conversion of this genetic information—that is, the expression of the different genes into proteins—takes place, however, during the different development stages of an organism as well as in different cell types

[*] Dr. F. Lottspeich
Max-Planck-Insitut für Biochemie
D-82152 Martinsried (Germany)
Fax: (+49) 89-8578-2802
E-mail: lottspei@biochem.mpg.de

  1433-7851/99/3817-2477 $ 17.50+.50/0

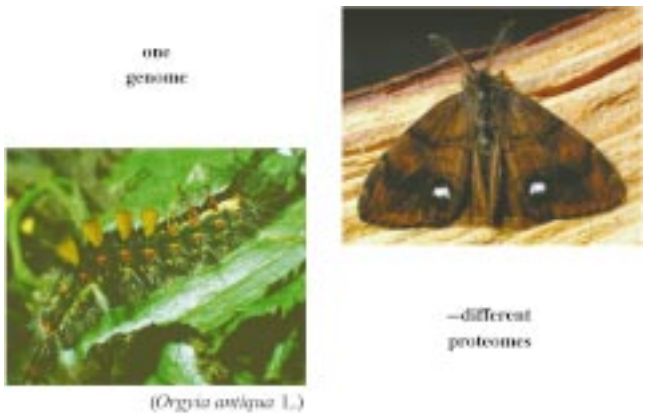

Figure 1. Caterpillar and butterfly of *Orgyia antiqua L.* (reprinted with permission of TOPLAB GmbH, Martinsried, Germany).

and under different environmental conditions. This leads to an enormous individual phenotypic diversity in nature.

During the conversion of genetic information into proteins, the tools and raw materials of a cell, there are regulatory mechanisms that adjust the relative amounts of individual proteins with the utmost precision. The smallest of disturbances of this finely tuned protein expression can lead to considerable biological consequences. Nature has therefore developed a complex (and very robust) regulation network which at all levels of information conversion—from transcription, through translation, to the level of the protein—is equipped with numerous feedback and reference points. Thus, the expression of proteins is regulated by transcription and translation factors, which are themselves proteins and therefore subject to the same regulation mechanisms of synthesis and protein degradation. Moreover, the regulation network becomes more complex since proteins can be structurally altered by posttranslational modifications (e.g. phosphorylation, glycosylation, processing), and hence their biological activity and function can be modulated. Knowledge of posttranslational modifications of a protein is therefore an important factor in the understanding of its mode of action and is thus essential for a functional analysis. These endogenous molecular relationships are themselves affected to a considerable extent by exogenous parameters such as temperature, culture conditions, and stress.

All these complex, interconnected processes are under precise spatio-temporal regulation. It has been demonstrated that even the most lavish molecular biological methods for the functional analysis of individual genes ("gene disruption", "knock-out techniques", etc.) frequently yield results that are either ambiguous or are of no practical value at the molecular level. Furthermore, in general no conclusions can be drawn on the amount of a corresponding active protein from the amount of an mRNA. There is therefore a considerable need for complementary strategies directed at proteins which, together with molecular biological investigations, can overcome the problem of the comprehensive functional analysis of genes and gene products.

2. From Protein Analysis to Proteome Analysis

Up to 1950 it was still unclear whether a certain protein possessed a defined and unequivocal covalent structure, or consisted of a very heterogeneous mixture of amino acid polymer chains. The work of Pehr Edman and Frederic Sanger on the sequence analysis of proteins and peptides demonstrated that a particular protein has a clear and unified structure.[2] The following three decades were characterized by efforts to assign certain biological functions to individual proteins. A protein was almost always isolated and purified on the basis of its biochemical activity before its covalent structure (amino acid sequence and modifications) could be elucidated in detail. Once this information was known, possible interaction partners were sought and analyzed in detail, which in turn served as the starting point in the search for further interacting molecules. The disadvantage of this reasonable, function-based, and very successfully strategy was mainly the relatively large material consumption and the tedious isolation, during which consideration always had to be paid to retention of biological activity.

The methods of protein development were under constant development, and their sensitivity increased considerably. An important milestone in this methodical advance was the development of two-dimensional (2D) gel electrophoresis in 1975.[3] Even then the great potential of this high-resolution separation method—which could separate a large number of individual proteins from even such complex mixtures as tissues, cells, or body fluids—was recognized. Protein patterns from normal and pathological states were compared, but the differences observed had solely diagnostic value since the

Friedrich Lottspeich, born in 1947, studied chemistry at the Universität Wien. In 1978 he completed his PhD with Pehr Edman at the Max-Planck-Institut für Biochemie in Martinsried, Germany, on the primary structure of fibrinogen. He worked at this same institute until 1984, when he was appointed as the leader of the independent research group "Mikrosequenzierung" at the genetics center of the Universität München. After habilitations at the Universität München and the Universität Innsbruck he returned in 1990 to the Max-Planck-Institut für Biochemie in Martinsried as leader of protein analysis. He is the author of over 550 original publications and the editor of several scientific books. His research interests center on methodical and practical approaches for protein structure elucidation and proteome analysis.

proteins identified could not be further characterized; at that time the analytical methods were either lacking or of inadequate sensitivity. It was only the appearance of special sample preparation techniques[4] and the further development of protein sequencing[5] that from the middle of the 1980s permitted analysis of proteins that had been separated by 2D gel electrophoresis. On the other hand, as a result of the spectacular success and the high development potential of molecular biology, protein chemistry had slipped from the focus of scientific interest at the beginning of the 1980s. Molecular biological investigations for solving biological questions were preferred and which were also successful, often in an astonishingly short time. New gene technology methods were being constantly developed and improved, and their application achieved an all-time high point in the sequencing projects of complete genomes.[1]

An awareness of the complexity of natural regulation networks was obtained, and the handling of a large number of complex samples in microbiology gave rise to considerable pressures for the development of automated and more rapid analysis techniques. These high-throughput methods produced a large amount of data which in turn stimulated the development of data banks and the associated software tools, and hence the birth of bioinformatics.

Towards the end of the 1980s, increasingly loud voices proclaimed that microbiological methods alone could not unravel the multiplicity and complexity of biological processes. It became clear that although the mRNA pattern can in many cases give information on switched-on or switched-off genes and gene families, the amount of mRNA allowed no conclusions to be drawn on the amount of the corresponding active protein. Incalculable processes (mRNA degradation, translation control, protein degradation, and posttranslational modifications) destroy the strict correlation between RNA and protein amount, at the mRNA level as well as at the protein level.[6] The view that proteins must also be included for a more complete picture of biological events continued to assert itself since proteins are the active players in the cell. It was precisely at this point in time that two new mass spectrometric techniques, electrospray mass spectrometry[7] (ESI-MS) and matrix assisted laser desorption/ionization mass spectrometry[8] (MALDI-MS) were first used successfully in protein analysis. With the improvement in these MS techniques and the enormously fast growing information on sequence data from the genome projects, the basis of a completely new, sensitive, and comparatively rapid protein analysis was formed. This analysis make it possible to process a large number of proteins. The old idea of differential analysis of complex protein mixtures could now be started afresh, and it now finds use in proteome analysis.

The word proteome was first used by Marc Williams in 1994 at the 2D gel electrophoresis meeting in Sienna as "the protein equivalent of a genome", and it was rapidly accepted by reason of its "convenience" and analogy to the expression "genome". This term was subsequently more closely defined as it became clear that the mere detection and compilation of all possible protein and protein variants which could be produced from a genome had very little informative value. Even the mere positioning of each possible protein within a given separation space—that is, pure mapping—will for many reasons bring only very little practical use despite enormous expenditure:

- The amount and the posttranslational modifications of a protein are decisively important for its biological activity and action.
- No biological state exists in which all possible proteins of an organism are expressed.
- Even with very good separation systems, several proteins are often located in one position. Therefore the identity of the protein of interest must be repeatedly checked analytically in the relevant experiment.
- Technical problems of proteome analysis and small differences in sample preparation and analytical protocols obstruct the comparability of "proteome maps" from different laboratories.

For these reasons the following views on proteome experiments are increasingly gaining ground: A proteome represents the protein pattern of an organism, a cell, an organelle, or even a body fluid determined quantitatively at a certain moment and under precisely defined limiting conditions. Furthermore, the proteome reflects a current metabolic status of the corresponding cell (or organism) which is determined by manifold interactions of the different molecules that are currently still difficult to analyze and by innumerable environmental parameters. Unlike the genome, the proteome is thus a highly dynamic system which is characteristically altered by changes in "environmental conditions" (for example, the change in culture conditions for a production strain or the addition of a drug to a cell culture).

Proteome analysis are only meaningful in combination with a subtractive procedure (Figure 2) in which two or more well-defined states can be compared. Changes in individual proteins for the protein patterns of these different states (e.g. a cell with and without a drug, or by comparison of cells from normal and pathological states) are observed and quantitatively evaluated. Therefore the proteome can be regarded as a unique, highly sensitive monitor for complex

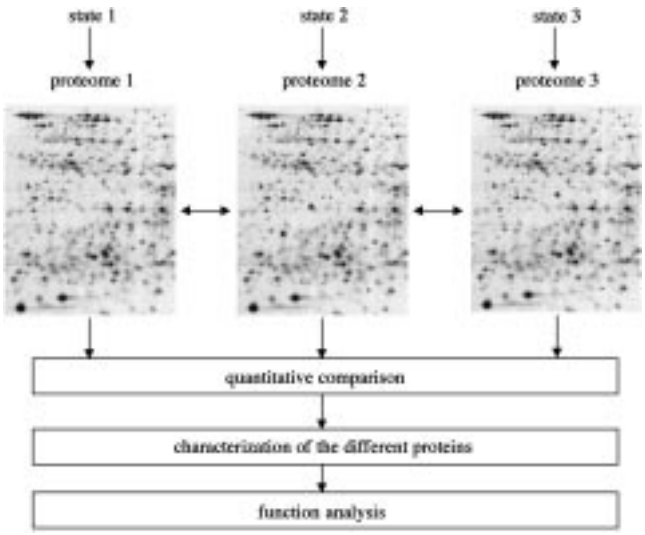

Figure 2. Schematic representation of the subtractive procedure. Differences that are characteristic of the individual starting states are recognized by the comparison of two protein patterns.

metabolic and regulatory relationships of an organism. It should be possible to draw conclusions about the participating functional networks (metabolic pathways, regulatory cascades) from the observed changes which usually involve many proteins. A functional statement can be reached by means of a correlation with the phenotype. The situation can be likened to a movie which, although it consists of a sequence of individual static pictures, obtains a significant information content from the dynamics of the changes between the individual pictures. The stills of a movie are, however, still able to give an impression of the events of the movie, and this all the better the more carefully and the more characteristically the scenes are selected for the stills. Similarly, a meaningful statement can also be achieved in proteome analysis through an investigation of the changes between different (per se static) snap shots of quantitative protein composition. However, we are only at the threshold of an understanding of the "pictures" of the proteome movie. We must learn to correlate the observed differences in protein patterns with biological effects. The better the choice of individual proteome states, the clearer the interpretation of sequence and type of changes.

Proteome analysis can thus give an new quality of answers to biological questions that currently cannot be obtained by any other technique. Only the interaction of proteome analysis with the molecular biological gene technology methods of genome analysis will give a better insight into the complex functional regulation and metabolic networks of nature.

3. Methods of Proteome Analysis

In contrast to classical protein analysis, proteome analysis has the distinct advantage that a biologically relevant statement can be obtained independent of the biological activity of the individual protein. As a result, no consideration needs to be given to protein activity during separation, which allows the use of denaturing separation conditions and hence significantly faster and more efficient analyses. However, the systematic and comprehensive analysis is much more difficult than the analysis of genes since, unlike with RNA or DNA, no amplification possibilities are available for proteins (a limited sensitivity arises from this), and the physical-chemical and biochemical properties of proteins are extremely varied. Therefore, new methods must be developed to be able to handle efficiently protein mixtures that can contain more than 10000 different protein species. Methods of proteome analysis are being developed worldwide, particularly in university research laboratories. However, because of limited resources, only single aspects of proteome analysis are being investigated in individual laboratories. This has given rise to the current situation that, although in principle all methods for the analysis of proteomes exist, the expertise for the totality of the necessary techniques (e.g. sample preparation, 2D gel electrophoresis, enzymatic and chemical cleavage, sequence analysis, mass spectrometry, bioinformatics) within a group is available in only very few places.

Proteome analysis can be divided into individual steps:

- formulation of the question and determination of the starting conditions for the proteome analysis (or for the individual states of a subtractive procedure),
- sample preparation,
- separation of the proteins,
- quantification of the proteins,
- data bank analysis of the protein pattern with bioinformatic methods,
- protein chemical characterization of the altered proteins and analysis of posttranslational modifications.

3.1 Formulation of the Question and Determination of the Starting Conditions for the Proteome Analysis

The methodology and thus the effort and costs of a proteome analysis are significantly influenced by the formulation of the question itself: A proteome analysis in the "puristic" sense should be able to provide statements on functional networks and/or involvement of certain proteins in individual reaction and control mechanisms by means of the differences between two or more snap shots of the protein patterns. The procedural possibilities are very numerous so that only a few representative and random examples will be described:

- Which proteins are involved in reaction networks or biological mechanisms? This information can serve as a basis for the selection of target proteins or marker proteins.[6b, 9, 10]
- How do chemical compounds and environmental conditions affect protein expression, and how do they alter differentiation, proliferation, and metabolism through their pharmacological and toxicological actions? What differentiates here, for example, between drugs with and without side effects? Which proteins or which protein constellations are characteristic of side effects (see also Figure 6)?[6b, 10]
- What are the molecular foundations for efficient production strains in microbiology? Can the expression of proteins be influenced on a rational basis such that higher production yields or product purities can be achieved?[10b]

In general, the difficulty in drawing functional conclusions from changes increases greatly with the number of changed proteins. Therefore, to be able to make meaningful statements, the differences between the individual states of a proteome analysis should not be too widely chosen.

A thorough consideration of the effort needed should be established before a proteome analysis is undertaken. For technical reasons—both during separation as well as during detection—a complete quantitative observation of all expressed proteins is (still?) not possible. Hydrophobic proteins and proteins that are very large, very small, very acidic, or very basic give rise to serious separation problems which prevent the preparation of a complete proteome even under otherwise optimal conditions. It is currently assumed that significantly more than 50% of expressed proteins can be quantitatively detected and analyzed in organisms with a small genome, for example yeast. Since a few milligrams of

protein material can be loaded onto a 2D gel,[11] proteins which occur in copy numbers of more than 100000 per cell can easily be directly recognized, quantified, and also characterized by protein chemistry. However, most proteins are far less well expressed and must be enriched in additional steps (see Section 3.2), which means considerably more effort. In setting the objectives of the experiment consideration must therefore be given to how far an (approximately) complete proteome analysis is at all necessary or meaningful (or even feasible), or whether the investigations can be limited to a certain group of proteins. This can be achieved by biological fractionation (e.g. of certain organelles), but it can also include a physical fractionation (e.g. precipitation). In all cases it is important that the starting material for the actual proteome analysis can be prepared reproducibly.

An important area of application for proteome analysis is the recognition of proteins which are correlated (e.g. diagnostic and therapeutic markers) with a certain state (e.g. a disease). For this purpose the protein patterns of, for example, healthy and pathological cells, tissues, or body fluids are compared. Even if here the individual states of the proteome analysis are biologically widely separated, and there are no expectations that complex functional relationships can be elucidated directly from the changes in protein patterns, the aim to recognize important diagnostic and therapeutic proteins is often achieved. These generally serve as a starting point for continuing analyses. The following aspects must be taken into consideration, however, in the evaluation of results:

- There are almost always problems in the sampling of ex vivo material which can rarely be carried out with complete reproducibility. Consequently, under certain circumstances several cell types with different proliferation and differentiation states are present.
- The "normal values" for individual proteins can differ considerably in different test subjects and must be statistically validated.
- The protein pattern will also differ significantly, even under identical limiting conditions, owing to the relatively frequent occurrence of "polymorphism". It is estimated that the genomes of two people differ in about one million base pairs; many of these differences are manifested at the protein level as amino acid exchanges. The resulting isofunctional, but physically slightly different proteins are found at different positions during proteome analysis and thus alter the protein pattern.

All these difficulties with sampling and the sample material can be ameliorated by a large statistical basis. However, the material for this is sometimes not available.

A further group of experiments is also frequently ascribed to proteome analysis since in terms of methodology it uses the same tools. This is the investigation of protein mixtures which are present, for example, in the search for interaction partners of a protein by immunoprecipitation or in the analysis of the composition of protein complexes. Here the objective is to characterize all proteins that bind to a given protein or are the components of a protein complex. Even if here too all high-throughput methods of proteome analysis are used, the significant difference—and the enormous simplification—to a real proteome analysis is that on one hand the limiting conditions for the analysis are far less complex and on the other the accurate, quantitative determination of the amount of protein present plays only a subordinate role (the absence of an absolute quantity determination prevents a simple statement on the stoichiometric composition of the complexes).

3.2. Sample Preparation

Sample preparation is the first important step of every proteome analysis. If a good description of individual proteome states is to be guaranteed, all possible experimental parameters must be registered and held constant since protein expression of a cell is highly sensitive to changes in external parameters. To be able to keep these parameters under the best possible control, cultivated cells must normally be used for proteome analysis in order to limit biological variability and to have the necessary amount of protein available. All, including unintended, changes during sampling will result in quantitative and qualitative variations in the protein pattern, which again must be compensated for by multiple analysis and complex statistical safeguards.

To be able to assess quantitative relationships correctly, care must be taken that the proteins remain completely intact during every type of sample preparation, that is, that they are neither modified nor degraded. Proteases are frequently released during cell disintegration, which can degrade proteins very rapidly and thus artificially increase the heterogeneity of the protein mixture and at the same time alter the natural quantitative relationships of the proteins to each other. This means that sample preparation must be specifically optimized for each starting material.[12] It is possible that protease inhibitors and/or high concentrations of chaotropic agents for denaturing proteolytic enzymes must be added or a low pH must be used during disintegration. In every case workup must be rapid and at low temperatures. Relatively simple samples, for example body fluids or cells without extremely stable cell membranes, are best dissolved directly in the application buffer for 2D gel electrophoresis. A specific sample preparation protocol must be individually worked out for very demanding samples. The main aim must be to make sample preparation as simple and complete as possible, but above all it must be reproducible; for the latter standardized and essentially automated procedures should be worked out and applied.

A good sample preparation can also include fractionation of the sample, which leads to a reduction in sample complexity. This is particularly advantageous for the subsequent separation of the protein mixture and quantification of the separated proteins. Biological preseparations—such as isolation of certain organelle fractions (e.g. mitochondria, membranes, cell nuclei) or physical-chemical methods such as precipitation, preparative electrophoretic procedures (e.g. free flow electrophoresis), chromatographic prepurification steps—can be used.

Sample preparation for proteome analysis is extremely difficult, mainly because of the broad spectrum of physical-

chemical properties of the individual proteins. Therefore the proteins in such a complex protein mixture, as represented by the proteome, will inevitably differ quite dramatically in their solubility properties. Consequently they behave quite differently and should ideally also be handled differently:

- Proteins that are readily soluble in water or dilute buffer cause few problems. These are mostly cytosolic proteins or proteins from body fluids.
- Proteins that have very stable secondary and tertiary structures and are poorly soluble in water can be first solubilized by the addition of chaotropic substances such as urea or guanidine hydrochloride. The danger of undesired partial and uncontrollable modifications of individual proteins can arise (e.g. by carbamylation), which causes an originally uniform protein to occur in many forms and the sample mixture to become even more heterogeneous. This can cause considerable problems during separation and quantification.
- Proteins which, when brought out of their natural environment, form large insoluble complexes are first solublized by chemical reactions, for example reduction of disulfide bridges.
- Membrane proteins, whose natural environment is lipid membranes and which aggregate very readily during isolation from the membrane and consequently become insoluble, are particularly difficult to handle. These hydrophobic proteins can only be held in solution by the action of detergents, which, however, frequently disrupt the subsequent stages of efficient protein separation, or even make them impossible.
- The high protein concentrations that are necessary for most separation and analytical procedures often involve the risk of aggregation and thus the precipitation of certain proteins. In contrast, low protein concentrations, which are generally advantageous for protein solubility, require additional steps prior to the separation and analytical procedures; this again involves the risk of protein loss.

3.3. Protein Separation

The best possible high-resolution separating techniques that are able to separate all proteins with the most diverse properties at the same time must be used for the quantitative analysis of a protein. The only high-resolution methods available in the molar mass region of proteins are electrophoretic and chromatographic techniques. Yet in a given analysis, neither of these techniques can separate much more than about 100 components. However, since a simple cell probably contains at least 10000 protein species, and moreover the amounts of proteins present in a sample can differ a factor of 10^6 or more, an sufficiently large separation space must be available. This can only be achieved by coupled, multidimensional separation procedures.

3.3.1. Electrophoretic Techniques

Proteins have a zwitterionic character and, depending upon the buffer conditions, can be positively or negatively charged. Electrophoretic procedures separate compounds according to their mobility in an electric field; the electrophoretic mobility of each protein is a characteristic value. Two high-resolution electrophoretic techniques are used in proteome analysis, isoelectric focusing (IEF) and sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS-PAGE).[3]

In IEF, the individual proteins of a mixture move to their isoelectric point in a pH gradient where they lose their net charge and thus their electrophoretic mobility. If because of thermal diffusion a protein moves from the pH region of its isoelectric charge, it gains a charge and in this moment returns electrophoretically to its "correct" position, which corresponds to the isoelectric point. Thus isoelectric focusing is a concentration end point method which forms sharp, highly concentrated protein bands. It is also compatible with (non-ionic) surfactants, which is an important prerequisite for its use in proteome analysis.

In SDS-PAGE, all protein molecules are loaded with SDS and move as negative SDS-protein complexes in the direction of the anode in an electric field and are separated in the polyacrylamide matrix according to size.

In the combination of IEF and SDS-PAGE, the 2D gel electrophoresis developed by Klose and O'Farrell in 1975,[3] the IEF gel with its highly concentrated protein bands is placed on an SDS–polyacrylamide gel and the samples are further separated according to molecular size in a second separation step. Highly resolved, two-dimensional protein patterns are formed which can be visualized by coloration and evaluated quantitatively (Figure 3). The 2D electrophoresis is

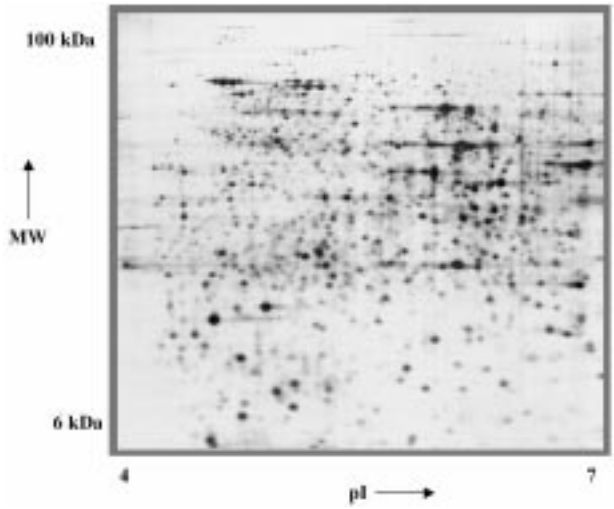

Figure 3. View of a typical 2D gel from *Saccharomyces cerevisiae* (immobiline technique, pH 4–7).

currently the only method for proteins which is able to make available a separating space of several thousand components and can separate complex protein mixtures within a few hours. However, only with the introduction of immobilized pH gradients (IPG)[13] for the isoelectric focusing dimension and through improved sample application techniques[11] over the last few years has 2D electrophoresis reached a level that leads to reproducible protein patterns. With the use of IPG gels, which are also obtainable commercially as ready-to-use

gels, many of the inadequacies of isoelectric focusing with ampholytes (cathode drift, unstable ampholyte mixtures of variable composition) could be overcome. A further, immense advantage of the immobiline technique is that gels with very narrow pH ranges (total pH gradient, for example, 1 pH unit for a gel width of 18 cm) can be prepared which can separate proteins that differ by less than 0.01 pH unit in their isoelectric point (Figure 4).[14] These gels also have the advantage that they can be loaded with a large amount of protein (up to ca. 15 mg) so that poorly expressed proteins can also be visualized directly.[11]

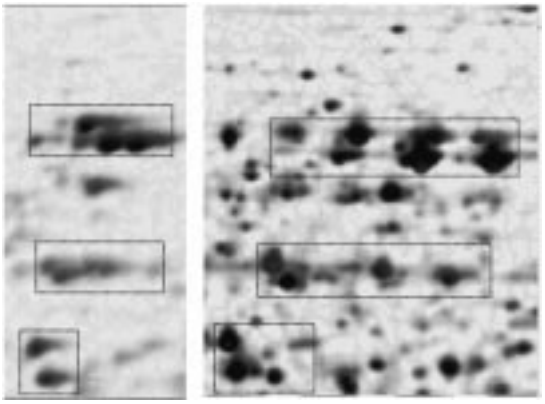

Figure 4. Split pH gradients of IPG gels improve the resolution and loadability. The proteins in the marked areas of the gel shown on the left (pH range 3–10) are much better separated on the gel shown on the right (pH range 5–7).

Several special properties make 2D gel electrophoresis such an excellent separating technique for proteome analysis:

- There are two complementary and efficient separating principles (IEF separates by charge, SDS-PAGE by molar mass).
- The technique may in principle be used for all proteins (detergent compatibility).
- The technique is parallel and therefore rapid.
- Partial regions of the proteome can be analyzed with very high resolution and high protein loading over a narrow pH gradient.

However, serious disadvantages confront the advantages of high universality and rapidity, which makes further development of this procedure imperative:

- Skilled sample preparation is necessary.
- Sample loading is quantity-limited, so that very poorly expressed proteins cannot be directly recorded.
- Protein transfer from the first to the second dimension is not quantitative and poorly reproducible.
- There is no simple quantification. Proteins must be colored with dyes which bind differently to different proteins.
- Proteins are located in a polyacrylamide gel matrix which is chemically not inert and which therefore can lead to modifications and, moreover, which prevents a direct analysis of the separated proteins.
- Automation is not available.
- The processing of the 2D gel analysis is technically demanding, and it is extremely difficult to achieve good comparability and reproducibility of the gels (at least between individual laboratories).
- Not all proteins are equally well detected. No useful information is obtained for very small ($MW < 10\,000$) and very large ($MW > 100\,000$) proteins.

In summary, an excellent resolution and an excellent compatibility with almost all protein classes conflict with an only partially adequate reproducibility and robustness. In spite of this, however, 2D gel electrophoresis is currently the only method which is used for the proteome project.

3.3.2. Chromatographic Techniques

Selectivity is achieved in high-resolution chromatographic methods (e.g. ion exchange, hydrophobic interactions, reverse-phase) by the individual interactions of protein molecules with the chromatographic surface, the stationary phase. Strongly interacting proteins can be irreversibly adsorbed onto the chromatographic surface. Unfortunately these adsorption losses are hard to predict since they depend upon the total protein concentration, the individual protein and its concentration as well as other, poorly controllable factors. The adsorbed protein changes the properties of the stationary phase so that a reproducible separation is extremely difficult.

A serious problem of all chromatographic methods is peak broadening. Every protein, even if applied to a column in very small volumes (high concentration), undergoes dilution during the chromatographic separation and elutes from the column in a concentration distribution which corresponds to a Gaussian curve. Since extremely complex protein mixtures are involved, fraction change occurs at well-defined timepoints, but is unpredictable with respect to individual proteins. Consequently, many proteins are found in several fractions, and the finally desired quantification of these proteins is made much more difficult.

A further serious disadvantage of the chromatographic techniques is that they are generally carried in series. After an initial fractionation into, for example, 100 fractions, each individual fraction must be subjected to a further separation dimension (e.g. a further chromatographic step). Even with very rapid chromatography (e.g. the total duration of an analysis 30 min including regeneration of the column), the total workup of the fractions from the first separation can require several days. Moreover, a large number of fractions are produced which have to be analyzed further. Proteins are, however, relatively labile compounds which can be rapidly modified, denatured, or degraded in solution, particularly in mixtures with other proteins. Workup periods become unacceptably long if a third chromatographic dimension must be added.

For these reasons multidimensional chromatographic separations have currently found no application in proteome analysis. However, because of the indisputable advantages of chromatographic techniques in quantification (UV detection), and automation, and the numerous possibilities for modulating separation selectivities, there is still the chance that one day multidimensional chromatography will be used as a complementary technique to electrophoresis for proteome analysis

3.4. Quantitative Recording of Separated Proteins

Since most physiological and pathophysiological processes are associated with quantitative changes to individual protein species, the central theme of proteome analysis is the most accurate determination possible of the amounts of the separated proteins. For this to be realized, the proteins must be visualized in an evaluable form.

3.4.1. Detection

The methods most frequently used are different stainings. The problem associated with all stainings is that they give a specific and unpredictable color intensity for each protein (according to its amino acid composition and the modifications that are present). The result is that no absolute statement on the amount of single proteins can be made, and protein patterns with different types of stainings cannot be compared. Furthermore, stainings are linear over a very limited range (maximum two orders of magnitude) because of the saturation effects. To cover a large dynamic region several gels with different amounts of protein must be prepared, so that an approximate calibration curve for each protein can be prepared. Only proteins which lie outside the saturation region of this calibration curve are suitable for quantification. Consequently, computer-supported image evaluation coupled with high-performance data processing is absolutely necessary. If different gels must be compared with each other and if at the same time the total protein amounts or color intensities are different, the question of standardization arises. Occasionally every protein spot is compared with the intensity of a known and constitutively expressed protein, which is itself problematic because constitutively expressed proteins can also change significantly in their amount. The better way is to relate the intensity of each protein to the total intensity of all proteins.

The most frequently used staining for proteome analysis is currently silver staining, in which the proteins are fixed in the gel with trichloroacetic acid and the gels are then placed in a silver nitrate solution. A number of silver ions are bound by the proteins and are precipitated in the form of elemental silver by reduction. The proteins darken very rapidly as a result of the high silver concentrations. The reaction is stopped by a large pH change. The silver staining can be used to detect proteins with relative sensitivity and exhibits a linear range of about $0.5-20\,ng\,mm^{-2}$. Unfortunately this staining is very difficult to reproduce since the blackening is highly dependent upon the duration of development and the temperature. A number of different protocols exist which differ from each other in duration, sensitivity, and simplicity, and which are being continually improved.[15]

Staining with the triphenylmethane dye Coomassie blue is somewhat more robust than silver staining.[16] It is, however, also significantly less sensitive (linear range ca. $50\,ng\,mm^{-2}$ to $1\,\mu g\,mm^{-2}$). Here too different staining protocols exist. The method of Neuhoff et al.,[16b] which achieves color penetration of the protein, gives the best results with respect to a quantitative evaluation.

Protein detection methods with fluorescent dyes such as SYPRO orange or SYPRO red have advantages over conventional staining with silver or Coomassie blue.[17] There are about as sensitive as silver staining but are much faster to carry out (ca. 30 min) and require no fixing of the proteins in the gel. Thus subsequent cleavage or transfer of the protein to a chemically inert membrane should be simplified. A further principle advantage of fluorescence detection is that over measurement times of different duration the fluorescence can often be detected in the presence of the fluorescence of rare proteins in the same experiment and quantified. The absolute amount of a protein cannot be derived from fluorescent staining as amino acid concentration and the type and number of translational modifications affect color intensity. Since the proteins are only visible under UV light and must be cut out, automated spot recognition and automated preparation of protein spots for the subsequent analysis is especially important in fluorescence staining.

In principle it should also be possible to allow all proteins of a proteome state to react covalently with a fluorescence reagent before their separation. In addition to an increased detection sensitivity this would bring the advantage that different states could also be treated with different fluorescent dyes. Then the proteins of the two states could be separated in the mixture under identical conditions ("multiplexing"), and the respective proteins could be assigned to the two states by detection of the different emission wave lengths. Differences would then be simpler to recognize since the technical problems of reproducible separation need not be considered.[18] Fluorescence labels with different optical but identical electrophoretic properties would need to be used. The problem with this procedure lies in the uniform covalent labeling of the proteins with the fluorescent dyes. The conversion is a chemical reaction at the functional groups of a protein. This reaction almost certainly does not go to completion in complex mixtures. Even if a reaction yield of 99 % with every single functional group were achievable, a uniform protein would provide a very heterogeneous mixture, which would also have to be separated with highly selective separation methods (primarily in isoelectric focusing). Each of the artificially generated by-products may indeed be present in a small fraction of a few thousandths, but even these minimal amounts almost always lie at the levels of naturally occurring proteins since the concentration range of different proteins in a cell spans at least six orders of magnitude.

Attempts can also be made to prevent the proteins from reacting fully, but to treat them instead with very small amounts of a fluorescence reagent. In this way only a small fraction of the protein is labeled and the main fraction of the protein remains unmodified. With a suitable choice of reagent it is possible to arrange that the separation behavior of the modified and the unmodified protein do not differ significantly.[18]

A fluorescence reaction after the first dimension of 2D electrophoresis could also be in part a way out. On the one hand, high-resolution separation is effected by isoelectric focusing, and on the other the mass heterogeneity from a few

fluorescence molecules lies below the resolution capabilities of the second dimension, SDS-PAGE.[19]

Immunological colorations with antibodies are very sensitive and can also detect very rare proteins directly. Unfortunately there are no specific antibodies for the peptide bond, so it cannot be used as a general detection tool. However, antibodies are very valuable in the detection of certain specific proteins or protein groups (e.g. tyrosine-phosphorylated proteins with a phosphotyrosine antibody). This applies similarly to lectin staining, which can be used to recognize glycosylated proteins.

Radiolabeling can be used very profitably for specific questions which allow an incorporation of radioactivity during the culturing of cells. This normally means the handling of high quantities of radioactive material, for which laboratories must be specially equipped. Thus, for example, specific proteins that are synthesized at certain time points can be detected by metabolic labeling with [^{35}S]Met or [^{35}S]Cys in pulse–chase experiments. Alternatively, a selective analysis of phosphorylated proteins (i.e., phosphoprotein partial proteome) can be achieved by labeling with [γ-^{32}P]ATP.

Even though radiolabeling is a highly sensitive detection method and quantification has been improved in its linearity by the development of photoimaging techniques, it does not reflect the absolute amount of a protein since the incorporation of the labeling is dependent upon the amino acid composition of the individual proteins.

3.4.2. Quantification

The next step after the detection of the separated proteins is their quantification. Except for amino acid analysis, no method exists to determine the absolute amount of individual proteins separated in a 2D gel. However, amino acid analysis is very sensitive towards contamination and work-intensive (even if readily automated), and it gives reliable results only with more than about 0.4 μg of protein. Thus, it is too insensitive for most proteins in a 2D gel.

Stained gels are normally measured by computer-supported laser densitometry. It has been demonstrated that only laser scanners have the required high dynamic range to provide good results. Special software packages for automatic spot recognition, quantification, and presentation of the results are used for data collection and further evaluation. The raw data and results are deposited in structured form in large data banks which can be accessed by special data bank programs for "data mining".

3.5 Data Evaluation by Bioinformatics

The large amount of data that is obtained during proteome analysis can no longer be evaluated without the massive support of specialist data bank software. Not only is rapid access to data concerning gel position, quantity, and identity of a protein important, but also the in part totally unstructured data on the origin, preparation, and work-up history of the sample needs to be readily available. Moreover, it must also be possible to include clinical data banks and literature data. Usually the desired information is deposited worldwide in different data banks which are linked by on-line connections. Part of the information is available from publicly accessible data banks by means of the World Wide Web.[1b] Other information, however, is so sensitive that it cannot be accessed publicly. The data bank software must also allow complex interrogation as, for example, which drug positively influences the expression of which proteins in which patient group, or which compounds influence the expression of certain proteins in the same way. The programs used for the quantitative evaluation of 2D electrophoresis are hopelessly overwhelmed by such questions. Special data bank programs for extensive data mining of RNA expression data and proteome data are currently a priority in the development in bioinformatics. The clear presentation of results of proteome analysis data is also receiving increased attention. These can be simple bar charts in which the protein amounts for individual proteins or protein groups from different experiments are illustrated (Figure 5). The progression of the amount of protein present at any point in time can be readily recognized, and proteins with a similar quantitative progression can be further collated by cluster and correlation analyses

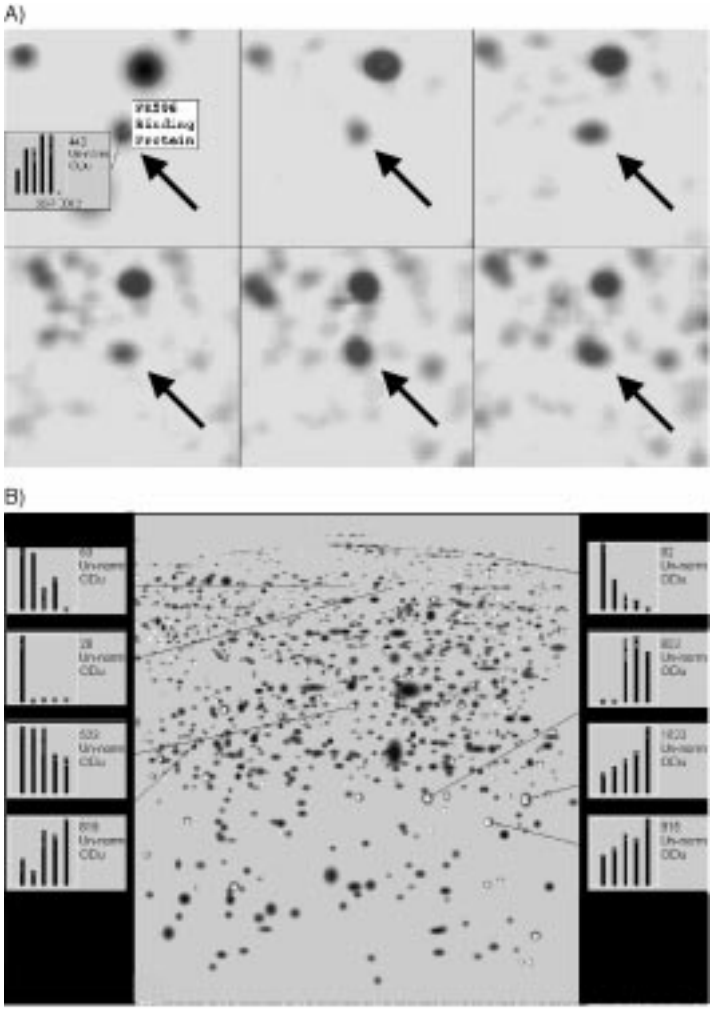

Figure 5. Illustration of proteome data. A) Change in the amount of a protein in different experiments. B) Illustration of the changes of several proteins in different experiments.

and compared.[20] Representations which depict complex relationships are also illustrative, and significance parameters can also be clearly considered (Figure 6).

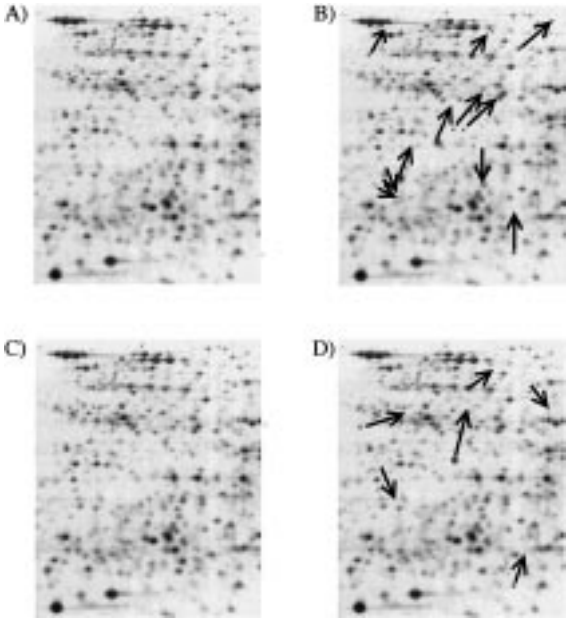

Figure 6. Illustration of a proteome analysis with consideration of statistical parameters, and schematic example for estimation of the action of a drug. A) Proteome of a normal cell. B) Proteome of a pathologically changed cell. The arrows show the changed proteins characteristic of the disease state. C) Pathologically changed cell under the influence of the ideal drug. The pathological changes are reversed. The cell appears normal. D) Pathologically modified cell under the influence of a real drug. A few pathological changes are reversed, but new changes to the proteome may also be induced, an indication of side effects. The upward or downward slope of the arrow is a measure of the increase or decrease in the respective amount of protein. The length of the arrow shows the statistical significance of these changes (unchanged proteins would all have horizontal arrows and are omitted for clarity).

3.6. Protein Chemical Analysis of Proteins Separated by Gel Electrophoresis

Methods for the analysis of proteins that have been separated by gel electrophoresis have been the subject of intensive efforts over the last 15 years. Since the proteins in a gel matrix are practically inaccessible to normal protein chemical methodology—such as sequence analysis, amino acid analysis, and mass spectrometry—the first success was to transfer the proteins from the gel onto chemically inert membranes and to immobilize them there.[4a–d] The intact, immobilized proteins can be analyzed by amino acid sequence analysis, by amino acid analysis, or more recently by mass spectrometric methods.[21, 22] However, since the analysis of intact proteins is work-intensive and rarely produces sufficient information for unambiguous protein identification, the analysis of internal fragments after enzymatic or chemical cleavage of the protein has proved to be the more efficient strategy. The procedure most generally used today within the context of proteome analysis is reproduced in Figure 7 and will be discussed in the following sections.

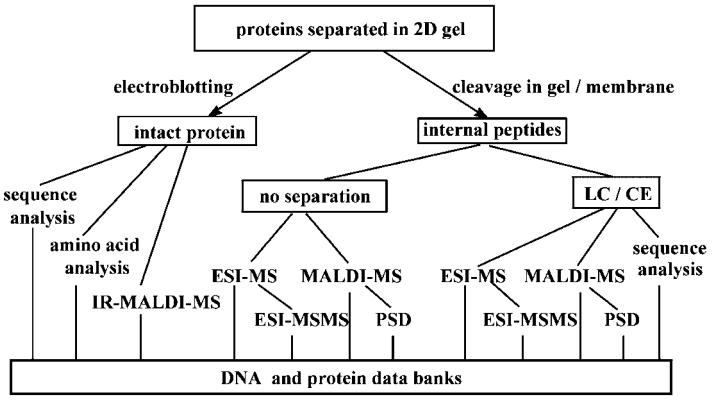

Figure 7. Schematic representation of the setup for high-throughput protein identification, and strategies for the characterization of proteins separated by gel electrophoresis.

3.6.1 Analysis of Intact Proteins

Amino Acid Sequence Analysis

The amino acid sequence analysis of blotted proteins is an automated, relatively work intensive, expensive, and slow standard procedure, and can be used without serious problems for protein amounts in the picomole range.[4a–e] The protein is frequently identified, or at least clearly characterized, by its N-terminal sequence. The achievable sequence length (maximum 40 amino acids) is sufficient for identification or homology searches in data banks, but it covers only a small part of the whole protein sequence. Therefore, closely related isoenzymes, splice variants, modifications, or point mutations outside the investigated amino terminal region of the protein cannot be recognized. The main problem is, however, that more than half of all proteins are N-terminally modified and are thus not accessible to sequence analysis.

Amino Acid Analysis

Every protein has a characteristic amino acid composition. Therefore, almost every protein can be identified with high probability in a protein data bank simply from the relationship of its amino acids to each other.[4e–h] In practice, however, the accuracy of the determination of individual amino acids is limited. Because of the drastic reaction conditions used in the hydrolysis of peptide bonds, every amino acid analysis is a compromise between total cleavage of the peptide bonds and the least possible destruction of sensitive amino acids. Therefore, only the most stable amino acids with relatively large error margins are used for a data bank search. The methods may be readily automated and is almost as sensitive as sequence analysis, but has the advantage that it can produce results even with N-terminally blocked proteins. At the same time amino acid analysis is the only method which can deliver information on the absolute amount of a protein. Unfortunately in practice these advantages are subject to a number of severe limitations which greatly restrict the general value of amino acid analysis in proteome analysis. The technique can only identify proteins whose sequence has been deposited in a data bank. With most proteins, however, the protein sequence is available only as the translation of the DNA sequence,

which often differs significantly from the sequence of the naturally occurring form of the protein. After translation, signal sequences are cleaved and proteins further processed and modified. The amino acid composition of the experimentally accessible protein thus differs from the theoretical composition deposited in the data bank. The former is normally not included in the data banks at all, or only marginally so. Because of the described technical and intrinsic problems, and because of the ubiquitous and unavoidable contamination (free amino acids from buffers, keratins, etc.), the margin of error in the determination of individual amino acids is so large that an identification of particularly poorly expressed proteins is frequently ambiguous. In practice, single amino acid exchanges or modification of an identified protein cannot be recognized.

IR-MALDI Mass Spectrometry

A significant characteristic of a protein is its molar mass. Even if the current methods are unable to determine the masses of proteins with adequate accuracy (errors in mass determination are greater than 100 ppm) to identify a protein in a data bank on the basis of its mass alone, mass information of the whole protein in association with other data is extremely valuable. For smaller proteins at least, the quality of mass determination is sufficient to recognize posttranslational modifications by a comparison of the mass calculated theoretically from the DNA sequence and the mass determined experimentally. However, the identity of a protein must be determined by means of sequence analysis, amino acid analysis, or the methods described in Section 3.6.2.

MALDI-MS is suitable for the analysis of immobilized proteins (Figure 8).[8, 21, 22] In a standard MALDI-MS preparation the proteins are embedded in a crystalline matrix of small organic molecules in which they are then vaporized and ionized by laser bombardment. The task of the matrix is to separate the individual protein molecules, absorb the laser light, and relax the energy in the solid lattice within a short period of time. Thus, an explosionlike dissipation of a small region of the solid surface and a transfer of the matrix and the protein molecules into the gas phase is achieved. With correct selection of laser energy this process is so gentle that the large, thermally labile protein molecules also remain intact. It is probable that the matrix also plays a role in the ionization of the protein molecules. Different wave lengths of laser light can be used when the interaction of the laser type (UV or IR) and the matrix (e.g. 2,5-dihydroxybenzoic acid for UV-MALDI or succinic acid for IR-MALDI) is important. The resulting protein ions are then accelerated in a time-of-flight (TOF) analyzer and the precise time of flight from ionization to detection is measured (Figure 8a).[7a] Since the time of flight at a given acceleration potential and flight path is dependent only upon the root of mass/charge (m/z), an accurate mass determination of the analyte molecule can be carried out by means of a calibration.

To analyze electrophoretically separated proteins after a transfer onto a chemically inert membrane directly with MALDI, the membrane must be incubated with the matrix immediately after electroblotting. UV-MALDI-MS usually

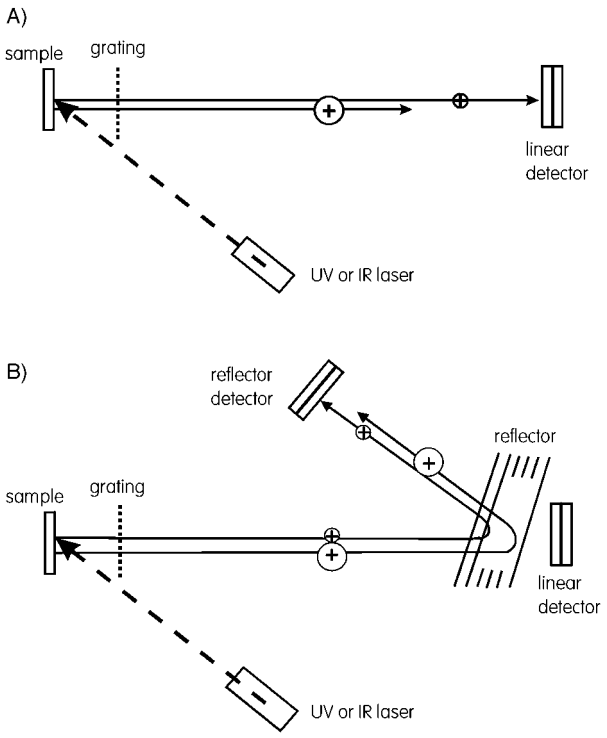

Figure 8. Schematic representation of the setup for MALDI mass spectrometry with a A) linear and B) reflector TOF detector. A) A laser pulse evaporates and ionizes the sample molecule. These are accelerated by the potential applied to the grating of the TOF analyzer. The speed of flight of the ions is proportional to $1/\sqrt{m}$ (m = mass of the ion). B) Ions can disintegrate during passage through the field-free drift path (PSD). These fragments, which carry structure information, travel with equal speed and reach the linear detector at the same time. For separation they are braked by the potential wall of the reflector, forced into reverse, and accelerated to the reflector detector. Since larger masses penetrate the reflector field more deeply, they travel along a longer path to the detector than small ions.

requires relatively hydrophobic molecules as matrix which are soluble only in organic solvents. During incubation of a membrane with such a hydrophobic matrix the protein molecules are partly dissolved and the geometric arrangement of the protein spots is lost by diffusion.[22a]

In contrast, it has been shown that during incubation of a membrane with a hydrophilic matrix the local arrangement of the blotted protein spots on the membrane and thus the resolution of the gel electrophoresis and even the intensity distribution within the protein spots is maintained.[21] Such hydrophilic matrices are used for IR-MALDI-MS, and therefore electroblotted proteins can be analyzed with high sensitivity (down to the attomole range).[21b] The limits and the intensities of the protein spots can be readily recognized, and protein mixtures within a 2D gel spot can also be recognized. Unfortunately the very different signal intensities for the individual proteins do not allow an estimation of the respective protein amounts. The development of automation and the improvements in MALDI-MS methodology with respect to ion yield and mass accuracy of larger proteins suggest that in the near future detection of electroblotted proteins will be carried out by MALDI-MS with automatic scanning of the blot membrane; a spot will be characterized by its mass as well as by its electrophoretic position.

3.6.2 Analysis of Internal Fragments

Enzymatic Cleavage and Elution of Peptides

Since the analysis of intact proteins is time-consuming and rarely delivers sufficient information, the analysis of internal protein fragments has proved to be the most efficient strategy for the characterization and identification of electrophoretically separated proteins. After enzymatic cleavage of the separated proteins, which can be carried out directly in the polyacrylamide gel matrix,[4i–o] the peptides obtained are eluted and analyzed by mass spectrometric and/or other protein chemical methods. A plethora of protocols exist, each of which show slight changes but principally fall back on one piece of work.[4i] The proteins separated by gel electrophoresis are punched out as tightly as possible, washed, and either dried or dehydrated by the addition of acetonitrile. A small volume of buffered enzyme solution is pipetted onto the shrunken gel fragments. Mainly trypsin, endoproteinase Lys-C, endoproteinase Glu-C, or endoproteinase Asp-N are used, which cut all proteins very specifically and completely. Trypsin is used very frequently, particularly if mass spectrometric analyses are to follow since it is itself relatively rapidly cleaved, and the autoproteolysis fragments serve as internal reference peptides for mass calibration. After incubation for a few hours at elevated temperature the reaction mixture, which contains the cleaved peptides partly in the supernatant and partly in the gel, is worked up differently for the subsequent analysis methods. The supernatant of the cleavage solution can be analyzed directly; the yields of larger or more hydrophobic peptides are often poor. Usually these peptides are eluted from the gel pieces with volatile acids containing organic solvents, for example trifluoracetic acid/acetonitrile (0.1/0.99 – 1/99). The further workup of the eluted peptides is guided by whether the proteome under investigation originates from an organism whose genome has already been fully elucidated. The large number of samples from proteome analysis has already led to automation of cleavage and subsequent elution of the proteins separated gel electrophoretically.[23] Commercial apparatus is available which can process automatically 50 samples per day (with sample amounts of less than 1 pmol).

Internal Sequences of Proteins from an Organism with Completely Sequenced Genome

A very simple and rapid identification of the proteins of an organism with known genome can be achieved with solely mass spectrometric methods. The eluted peptides are analyzed by MALDI-MS or by nano-electrospray MS without further separation.[7d, 24]

In general, the peptide mixture from cleavage on the gel is subjected directly to MALDI-MS, where robotic sampling systems may also be used. MALDI-MS is now so far developed that the samples can be analyzed automatically with very high mass accuracy (error less than 20 ppm). The data can be evaluated on-line so that the protein can be identified with very high probability in sequence data banks from the peptide mass pattern obtained. By this method more than 90% of the analyzed proteins can be identified directly with certainty and with high sequence coverage (Figure 9). Identification is not unambiguous with some proteins so that further analyses must be carried out. This can be achieved with MALDI-MS with post source decay (PSD) spectra.[25] The spontaneous disintegration of proteins and peptides during the flight in the field-free drift path of the TOF analyzer (metastable decomposition) is utilized here. The fragments of an ion all continue to travel with the same speed and reach the detector of a linear TOF analyzer at the same time. If the fragments in a TOF analyzer with reflector are allowed to travel up instead of against a uniformly charged

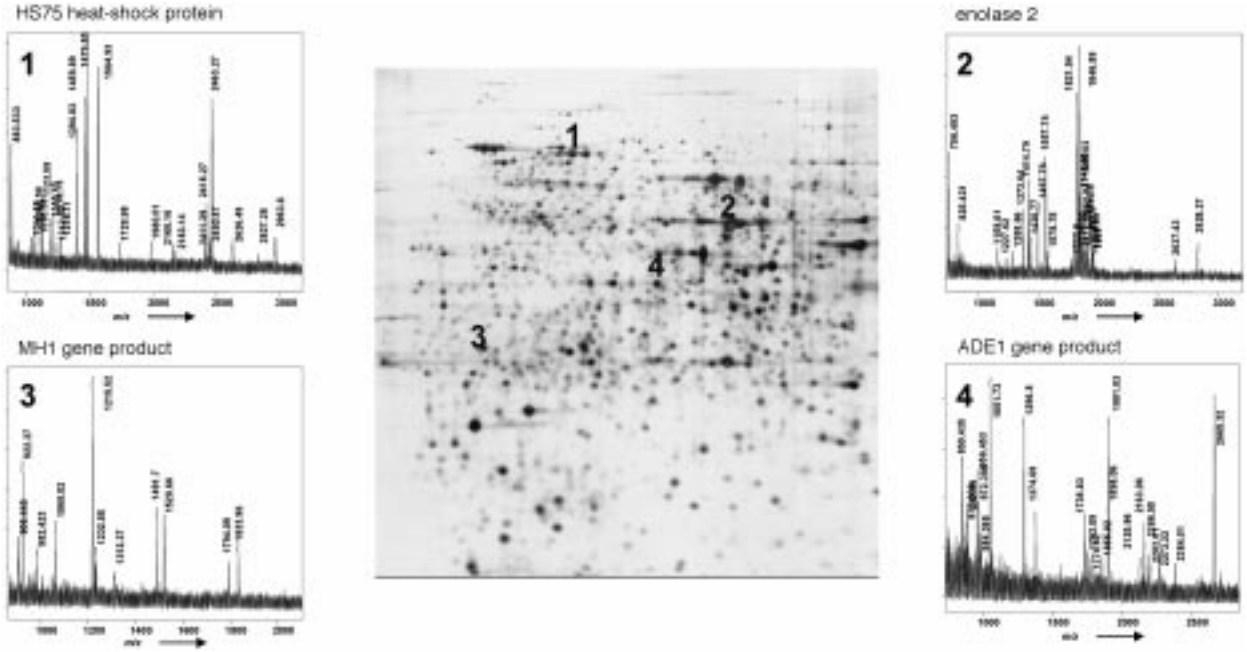

Figure 9. MALDI spectra that identify proteins separated by 2D electrophoresis. Enzymatic cleavage of a protein gives a mixture of proteins whose masses are analyzed in a data bank with computer support.

field (the reflector), the fragments come to a stop and reverse (Figure 8b). The different masses penetrate into the reflector field to different depths and so have to travel different paths, consequently reaching the detector of the reflector instrument at different times. The PSD fragments are thus separated, and their mass can be determined by calibration with reference compounds. The interpretation of the spectra is relatively complicated and tedious so that an exact evaluation of these spectra cannot be made in a proteome experiment where many samples must be analyzed. The uninterpreted spectra obtained are therefore compared automatically and on-line with a data bank of calculated spectra which were generated in silico from all theoretically possible peptides of an organism. The software then gives an identity suggestion based on similarity.[26]

The alternative to MALDI-MS analysis of the peptide mixture from cleavage in the gel is electrospray (ESI) MS (Figure 10).[7b–e] In ESI-MS the sample in liquid form is continuously sprayed in an electrostatic field to form small

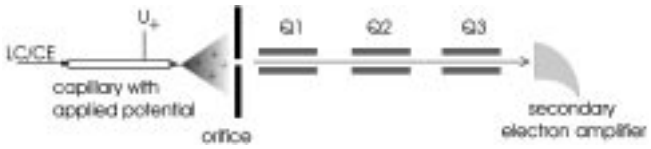

Figure 10. Schematic representation of the setup for ESI mass spectrometry. The samples in liquid form (HPLC, CE, injection) are passed continuously through the capillary, to which a high potential is applied. The resulting spray contains the (multiply) ionized molecules which are introduced into the high vacuum field of the quadrupole mass spectrometer through the orifice. For mass determination the quadrupoles Q1 and Q2 are so set that all ions can pass. The actual mass determination occurs in Q3. For structure investigations the mass filter Q1 allows only ions of a single mass to pass into Q2. Fragmentation of the sample ions occurs in Q2, which is filled with argon gas, and the masses of the resulting fragments are analyzed in Q3.

droplets, which are rapidly desolvated and the charge density on the surface of the droplets becomes even larger. After repeated, spontaneous disintegration of the droplets (Coulombic explosion), the desolvated, multiply charged molecule ions are directed to the mass spectrometer, where they are normally detected with a quadrupole mass analyzer. This type of mass analyzer is a mass filter which under preset physical conditions only allows ions with a totally defined mass/charge ratio to pass through. All other ions cannot pass the analyzer and are lost. Through continuous change of the potential at the quadrupole, ions of different masses are allowed to pass sequentially (scanning), and the intensity of the ion flow is recorded in relation to the m/z ratio. The accuracy of the mass determination allows the charge state of each mass signal to be determined from the isotope distribution, and hence multiply charged ions to be recognized and, with computerized support, the mass of singly charged ion to be calculated.

Since the observed ion flow correlates with the concentration of the sprayed sample, attempts are being made to spray highly concentrated solutions at very low flow rates of $nL\,min^{-1}$ (nanospray ESI[7d]). In addition to increased sensitivity, this has the advantage that very long measurement times are available with a sample of a few microliters.

ESI-MS also offers the opportunity to obtain at least partial structure (sequence) information from the fragmentation of individual peptides.[27] A triple quadrupole apparatus is used in which the first quadrupole is used for the selection of a peptide ion. These selected ions are directed to a second quadrupole where they collide with argon gas and are fragmented. The resulting fragments are then analyzed in a third quadruple (Figure 10). The ESI-tandem-MS fragment spectra are somewhat clearer and more convenient to interpret than MALDI-PSD spectra, but here too automated and software-controlled interpretation programs are increasingly used for proteome analysis. Labeling of the C-terminal end of the peptide with ^{18}O, by the enzymatic cleavage of the protein in the presence of $H_2{}^{18}O$, simplifies the interpretation of the fragment spectra considerably.[28] An alternative to the quadrupole apparatus are the ion traps, whereby ions are trapped in a suitable electric field and can be kept on stable paths.[29] The individual ions can be catapulted from the trap by changes in the electric conditions and then recorded at a detector. The scan speeds are up to ten times faster than with a quadrupole detector, and the ion trap is particularly suitable as a rapid detector for coupling with HPLC, where a limited amount of time is available for each substance peak. Like the triple quadrupole apparatus, the ion trap is also suitable for structural determination of peptides since individual ions can be selected in the trap and fragmented by collision with inert gas atoms. From the results of the mass spectrometric sequence analysis—which usually does not give complete peptide sequences, but only "sequence tags"—and in combination with peptide mass finger print, each protein can usually be unambiguously identified. Recent developments also allow the coupling of electron spray ion sources with TOF analyzers (orthogonal TOF apparatus, Q-TOF, Figure 11). The whole ion flow is passed through quadrupoles and hexapoles to an orthogonal acceleration system (pusher), which directs the ions to a reflector TOF analyzer. This apparatus also permits structure identification when the in-line quadrupole is used to select the ion, and the selected ion is fragmented in a subsequent collision cell and then directed into the TOF by the pusher and analyzed.

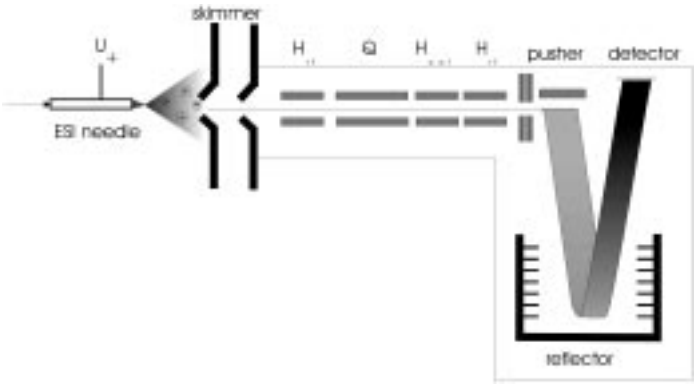

Figure 11. Schematic representation of the setup for ESI-Q-TOF mass spectrometry. The ESI ion flow is directed to the pusher through the quadrupole (Q) and the hexapole (H_{rf}) then deflected to a reflector TOF analyzer. Structure information is obtained when Q is used for the selection of one ion which is fragmented in the collision cell H_{col}. The fragments are analyzed in the TOF unit.

Internal Sequences of Proteins from an Organism with Unsequenced Genome

The strategy of identifying proteins by mass spectrometry solely by means of protein mass patterns currently has two main limitations:

- The genome sequence of the organism must be essentially known. This still has a practical significance today if one reflects, for example, that the human genome project is still not concluded, although it is anticipated that the DNA sequence of the most important organisms will be known in the near future.
- For a number of reasons, posttranslational modifications are only recognized to a limited extent.

The methods of rapid protein identification by means of peptide mass pattern fail if the sequence of the protein under investigation is not present in a data bank, if the protein is extensively modified, or if several proteins are present in one protein spot. Even if large parts of a genome are accessible only in partial sequences, EST data banks (EST = expressed sequence tags), analysis by peptide mass pattern is unsuccessful. Furthermore, difficulties in the cleavage and elution of larger peptides from the gel lead to a poor sequence coverage, and in mass spectrometry itself there are also inherent difficulties. Thus, suppression effects prevent quantification of individual peptides,[30] signals from ubiquitously present contamination (e.g. keratines) complicate the spectra, and artificial modifications during sample workup (e.g. oxidation or modifications of cystein residues) or during the measurement itself (e.g. oxidation, fragmentation) frustrate a simple assignment of the signals detected.

Additional time-consuming and slow protein-chemical microtechniques must be used for detailed analyses, such as the determination of posttranslational modifications of a protein or the characterization of proteins from organism with unsequenced genome (Figure 7). Capillary HPLC or capillary electrophoresis with coupled on-line mass spectrometry procedures are used for the separation of peptides after enzymatic cleavage in polyacrylamide gels. Such analyses, with which the protein is not only identified but also investigated, should be carried out starting from at least two different enzymatic cleavages in order to cover the total protein as far as possible with the sequence data. The end effect is that methods that produce de novo sequence information—that is, either mass spectrometric sequencing, which is cumbersome and is often associated with high uncertainty, or the classical Edman degradation, which gives unequivocal results but has a sensitivity limit of about 1 pmol of peptide starting material—must almost always be used. Both sequencing techniques are slow and currently cannot handle the large number of samples from a proteome analysis.

3.6.3. Analysis of Posttranslational Modifications

An important area of proteome analysis is the analysis of posttranslational modifications, which have a considerable effect on the functions and properties of a protein. Since this detailed protein analysis is cumbersome and tedious, only proteins for which there is an indication of posttranslational modifications should be investigated in detail. This information can be obtained from the mass spectrometric analysis of the whole protein with IR-MALDI (see Section 3.6.1). A deviation of the observed isoelectric point of a protein from that calculated from the DNA sequence is also a good indication of a modification.[31] Special mass spectrometric techniques such as precursor scan or neutral loss scan and the mass spectrometric sequencing methods MALDI-PSD and nanospray ESI MSMS are in particular used for the exact determination of the type and the position of the posttranslational modification.[32] They are supplemented by the classical structure determination procedures, mainly by the Edman sequence analysis. Partial modifications at several sites of a protein, which often occurs during phosphorylations and glycosylations, are especially difficult to analyze. In these cases separation of the peptide mixture by nano-HPLC with on-line mass spectrometric analysis must almost always be carried out. In summary, the characterization of posttranslational modifications is still demanding for protein chemists and, in spite of the enormous advances in recent years, still cannot by solved with high-throughput methods.[33]

4. Summary and Outlook

Unlike the static genome, a proteome—the quantitative protein pattern of an organism, a cell, or a body fluid under quite precisely defined limiting conditions—is highly dynamic. Several proteomes, each of which reflects a current development and metabolic state at a certain timepoint, exist for a single genome. Through appropriate selection of different states, functional conclusions can be made from the different protein patterns of the corresponding proteome with the help of bioinformatics.[13] Proteome analysis can only be used to its full potential in association with other areas of bioscience such as molecular biology, genetics, immunology, and medicine (Figure 12). At the present time proteome analysis is in statu nascendi; the methods are still immature, but are being developed at astonishing speed at all levels. Attempts are being made to improve or replace the only currently successful method, 2D gel electrophoresis, and

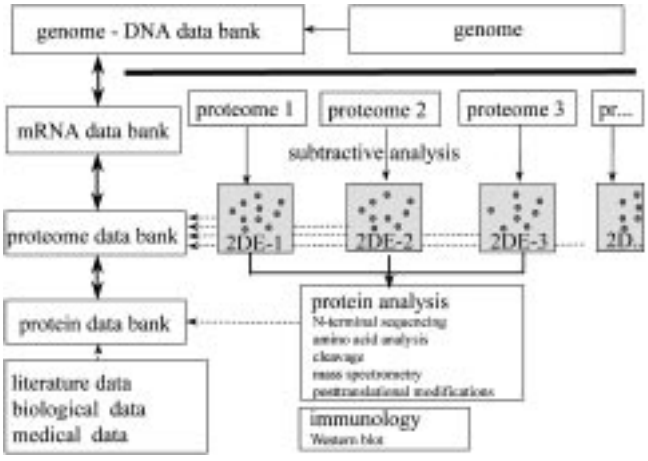

Figure 12. Proteome analysis in the context of different areas of bioscience. 2DE = two-dimensional electrophoresis.

quantification is gaining increasing significance with automation, new detection procedures, and the entry of bioinformatics. The identification and characterization of the separated proteins is most advanced. It is almost routine for proteins from organisms with totally sequenced genomes, although there is still room for improvement in terms of throughput and sensitivity. The recognition and localization of posttranslational modifications in high-throughput procedures will be the next stage in the development. Proteome analysis of organisms with incompletely established genomes is currently so difficult that in the immediate future it will remain restricted to special problems.

Since proteome analysis can be applied to a large number of complex questions, it is surely only a matter of time and financial investment before the first large practice-oriented proteome projects are started. The first realization of a comprehensive proteome project will certainly only be possible by a combination of the different and only partly available techniques with simultaneous massive automation and further development at all levels of proteome analysis; the development of new technologies is also foreseen. The large expenditure for instrumentation and the expertise required for a proteome analysis makes it likely that large proteome projects will only be carried out in specialized centers, the first of which are currently being developed.

The correlation of the proteome with mRNA expression (transcriptome) will be especially interesting since the two processes are regulated at quite different levels. The flow and changes in metabolic products and small molecules (the metabolome/fluxosome/physiome—the name has not yet been finally assigned), which have diverse feedback mechanisms to the transcriptome and the proteome, will also be more closely investigated in the future. Finally, insights into the interrelationship of genome, transcriptome, proteome, and metabolome will surely bring new knowledge of biologically relevant and very complex interconnected mechanisms of living organisms.

Received: February 16, 1999 [A 326 IE]
German edition: *Angew. Chem.* **1999**, *111*, 2630 – 2647
Translated by Dr. David Le Count, Congleton, Cheshire (UK)

[1] a) http://pedant.mips.biochem.mpg.de/; b) http://www.expasy.ch/.
[2] a) P. Edman, *Acta Chem. Scand.* **1950**, *4*, 283 – 290; b) F. Sanger, H. Tuppy, *Biochem. J.* **1951**, *49*, 481 – 490.
[3] a) P. H. O'Farrel, *J. Biol. Chem.* **1975**, *250*, 4007 – 4021; b) J. Klose, *Humangenetik* **1975**, *26*, 231 – 243.
[4] a) J. Vandeckerkhove, G. Bauw, M. Puype, J. Van Damme, M. Van Montagu, *Eur. J. Biochem.* **1985**, *152*, 9 – 19; b) R. Aebersold, D. B. Teplow, L. E. Hood, S. B. Kent, *J. Biol. Chem.* **1986**, *261*, 4229 – 4239; c) P. Madsudaira, *J. Biol. Chem.* **1987**, *262*, 10035 – 10038; d) C. Eckerskorn, W. Mewes, H. W. Goretzki, F. Lottspeich, *Eur. J. Biochem.* **1988**, *176*, 509 – 519; e) C. Eckerskorn, P. Jungblut, W. Mewes, J. Klose, F. Lottspeich, *Electrophoresis* **1988**, *9*, 830 – 838; f) M. Ploug, A. L. Jensen, V. Berkholt, *Anal. Biochem.* **1989**, *181*, 33 – 39; g) G. I. Tous, J. L. Fausnaught, O. Akinyosoye, H. Lachland, P. Winter-Cash, F. J. Vitoria, S. Stein, *Anal. Biochem.* **1989**, *179*, 50 – 55; h) S. Nakagawa, T. Fukuda, *Anal. Biochem.* **1989**, *181*, 75 – 78; i) C. Eckerskorn, F. Lottspeich, *Chromatographia* **1989**, *28*, 92 – 94; j) R. Aebersold, J. Leavitt, L. E. Hood, S. H. Kent, *Proc. Natl. Acad. Sci. USA* **1987**, *84*, 6970 – 6974; k) G. Bauw, M. Van Den Bulcke, J. Van Damme, M. Puype, M. Van Montagu, J. Vandeckerkhove, *Electrophoresis* **1990**, *11*, 528 – 536; l) M. J. Walsh, J. McDougall, B. Wittmann-Liebold, *Biochemistry* **1988**, *27*, 6867 – 6876; m) P. Tempst, A. J. Link, L. R. Riviere, M. Fleming, C. Elicone, *Electrophoresis* **1990**, *11*, 537 – 543; n) S. D. Patterson, D. Hess, T. Youngwirth, R. Aebersold, *Anal. Biochem.* **1992**, *202*, 193 – 203; o) J. Fernandez, M. DeMott, D. Atherton, S. M. Mische, *Anal. Biochem.* **1992**, *201*, 255 – 264; p) F. Lottspeich, C. Eckerskorn, R. Grimm in *Cell Biology: A Laboratory Handbook*, *Vol. 3* (Ed.: J. E. Celis), Academic Press, Orlando, **1994**, pp. 417 – 421.
[5] R. M. Hewick, M. W. Hunkapiller, L. E. Hood, J. Dreyer, *J. Biol. Chem.* **1981**, *256*, 7990 – 7997.
[6] a) L. Anderson, J. Seilhamer, *Electrophoresis* **1997**, *18*, 533 – 537; b) N. L. Anderson, N. G. Anderson, *Electrophoresis* **1998**, *19*, 1853 – 1861.
[7] a) K. Biemann, S. A. Martin, *Mass Spectrom. Rev.* **1987**, *6*, 1 – 76; b) J. B. Fenn, M. Mann, C. K. Meng, S. F. Wong, C. M. Whitehouse, *Science* **1989**, *246*, 64 – 67; c) K. Biemann, *Annu. Rev. Biochem.* **1992**, *61*, 977 – 1010; d) J. B. Fenn, M. Mann, C. K. Meng, S. F. Wong, C. M. Whitehouse, *Mass Spectrom. Rev.* **1990**, *9*, 37 – 70; e) M. Wilm, A. Shevchenko, T. Houthaeve, S. Breit, L. Schweigerer, T. Fotsis, M. Mann, *Nature* **1996**, *379*, 466 – 469.
[8] M. Karas, F. Hillenkamp, *Anal. Chem.* **1988**, *60*, 2299 – 2301.
[9] S. Müllner, T. Neumann, F. Lottspeich, *Arzneim. Forsch./Drug Res.* **1998**, *48*, 93 – 95.
[10] a) D. F. Hochstrasser in *Proteome Research* (Eds.: M. R. Wilkins, K. L. Williams, R. D. Appel, D. F. Hochstrasser), Springer, Berlin, **1997**, pp. 187 – 219; b) K. L. Wilkins, V. Pallini in *Proteome Research* (Eds.: M. R. Wilkins, K. L. Williams, R. D. Appel, D. F. Hochstrasser), Springer, Berlin, **1997**, pp. 221 – 232.
[11] a) B. Bjellqvist, J. C. Sanches, C. Pasquali, F. Ravier, N. Paquet, S. Frutiger, G. J. Hughes, D. Hochstrasser, *Electrophoresis* **1993**, *14*, 1375 – 1378; b) T. Rabilloud, C. Valette, J. J. Lawrence, *Electrophoresis* **1994**, *15*, 1552 – 1558; c) J. C. Sanchez, V. Rouge, M. Pisteur, F. Ravier, L. Tonella, M. Moosmayer, M. R. Wilkins, D. F. Hochstrasser, *Electrophoresis* **1997**, *18*, 324 – 327.
[12] T. Rabilloud, *Electrophoresis* **1996**, *17*, 813 – 829.
[13] a) B. Bjellqvist, P.-G. Rhighetti, E. Gianazza, A. Görg, R. Westermeier, W. Postel, *J. Biochem. Biophys. Methods* **1982**, *6*, 317 – 139; b) A. Görg, W. Postel, S. Gunther, *Electrophoresis* **1988**, *9*, 351 – 346.
[14] A. Görg, W. Postel, J. Weser, W. Patutschnig, H. Cleve, *Am. J. Hum. Genet.* **1985**, *37*, 922 – 930.
[15] a) C. R. Merril, J. E. Joy, G. J. Creed in *Cell Biology: A Laboratory Handbook*, *Vol. 3* (Ed.: J. E. Celis), Academic Press, Orlando, **1994**, pp. 281 – 287; b) A. Wallace, H. P. Saluz in *Cell Biology: A Laboratory Handbook*, *Vol. 3* (Ed.: J. E. Celis), Academic Press, Orlando, **1994**, pp. 289 – 298; c) H. Blum, H. Beier, H. J. Gross, *Electrophoresis* **1987**, *8*, 93 – 99; d) H. M. Poehling, V. Neuhoff, *Electrophoresis* **1981**, *2*, 141 – 147.
[16] a) W. Dietzel, G. Kopperschläger, E. Hofmann, *Anal. Biochem.* **1973**, *48*, 617 – 620; b) V. Neuhoff, R. Stamm, H. Eibl, *Electrophoresis* **1985**, *6*, 427 – 448.
[17] T. H. Steinberg, R. P. Haugland, V. L. Singer, *Anal. Biochem.* **1996**, *239*, 238 – 245.
[18] M. Ünlü, M. Morgan, J. S. Minden, *Electrophoresis* **1997**, *18*, 2071 – 1077.
[19] P. Jackson, V. E. Urwin, C. D. Mackay, *Electrophoresis* **1988**, *9*, 330 – 339.
[20] J. N. Weinstein, T. G. Myers, P. M. O'Connor, S. H. Friend, A. J. Fornace, Jr., K. W. Kohn, T. Fojo, S. E. Bates, L. V. Rubinstein, N. L. Anderson, J. K. Buolamwini, W. W. van Osdol, A. P. Monks, D. A. Scudiero, V. N. Viswanadhan, G. S. Johnson, R. E. Wittes, K. D. Paull, *Science* **1997**, *275*, 343 – 349.
[21] a) C. Eckerskorn, K. Strupat, F. Hillenkamp, F. Lottspeich, *Electrophoresis* **1992**, *13*, 664 – 665; b) C. Eckerskorn, K. Strupat, D. Schleuder, D. F. Hochstrasser, J. C. Sanchez, F. Lottspeich, F. Hillenkamp, *Anal. Chem.* **1997**, *69*, 2888 – 2892; c) K. Strupat, M. Karas, F. Hillenkamp, C. Eckerskorn, F. Lottspeich, *Anal. Chem.* **1994**, *66*, 464 – 470; C. W. Sutton, C. H. Wheeler, J. M. Corbett, J. S. Cottrell, M. J. Dunn, *Electrophoresis* **1997**, *18*, 424 – 431.
[22] a) M. Schreiner, K. Strupat, F. Lottspeich, C. Eckerskorn, *Electrophoresis* **1996**, *17*, 954 – 961; b) M. M. Vestling, C. Fenselau, *Anal. Chem.* **1994**, *66*, 47 – 477; c) J. C. Blais, P. Nagnan-Le-Meillour, G.

Bolbach, J. C. Tablet, *Rapid Commun. Mass Spectrom.* **1994**, *5*, 230 – 237; d) S. D. Patterson, *Electrophoresis* **1993**, *16*, 1104 – 1114.
[23] a) F. Hsieh, H. Wang, C. Elicone, J. Mark, S. Martin, F. Regnier, *Anal. Chem.* **1996**, *68*, 455 – 462; b) T. Houthaeve, H. Gausepohl, M. Mann, K. Ashman, *FEBS Lett.* **1995**, *376*, 91 – 94.
[24] S. D. Patterson, R. H. Aebersold, *Electrophoresis* **1995**, *16*, 1791 – 1814.
[25] B. Spengler, D. Kirsch, R. Kaufmann, E. Jaeger, *Rapid Commun. Mass Spectrom.* **1992**, *6*, 105 – 108.
[26] a) K. J. Eng, A. L. McCormac, J. R. Yates III, *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 976 – 989; b) J. R. Yates III, *Electrophoresis* **1998**, *19*, 893 – 900.
[27] M. Mann, M. Wilm, *Anal. Chem.* **1994**, *66*, 4390 – 4399.
[28] M. Schnölzer, P. Jedrzejewski, W. D. Lehmann, *Electrophoresis* **1996**, *17*, 945 – 953.
[29] a) K. R. Jonscher, J. R. Yates, *Anal. Biochem.* **1997**, *244*, 1 – 15; b) R. E. March, *J. Mass Spectrom.* **1997**, *32*, 351 – 369.
[30] R. Kratzer, C. Eckerskorn, M. Karas, F. Lottspeich, *Electrophoresis* **1998**, *19*, 1910 – 1919.
[31] a) B. Bjellqvist, G. Hughes, C. Pasquali, N. Paquet, F. Ravier, J. C. Sanchez, S. Frutiger, D. Hochstrasser, *Electrophoresis* **1993**, *14*, 1023 – 1031; b) B. Bjellqvist, B. Basse, E. Olsen, J. E. Celis, *Electrophoresis* **1994**, *14*, 1023 – 1031.
[32] a) C. Eckerskorn in *Bioanalytik* (Eds.: F. Lottspeich, H. Zorbas), Spektrum, Heidelberg, **1998**, pp. 323 – 368; b) J. W. Metzger, C. Eckerskorn, C. Kempter, B. Behnke in *Microcharacterization of Proteins* (Eds.: R. Kellner, F. Lottspeich, H. E. Meyer), 2nd ed., WILEY-VCH, Weinheim, **1999**, pp. 213 – 234.
[33] a) H. E. Meyer in *Microcharacterization of Proteins* (Eds.: R. Kellner, F. Lottspeich, H. E. Meyer) , 2nd ed., WILEY-VCH, Weinheim, **1999**, pp. 159 – 175; b) A. A. Gooley, N. H. Packer, in *Proteome Research* (Eds.: M. R. Wilkins, K. L. Williams, R. D. Appel, D. F. Hochstrasser), Springer, Berlin, **1997**, pp. 65 – 91.