Proteomics - how proteins communicate in a cell

Michael O. Hottiger

Institute of Veterinary Biochemistry, University of Zürich, Winterthurerstrasse 190, 8057 Zürich, Switzerland.

Introduction

Proteomics is the term used for the analysis of all proteins present in a cell or tissue. Its main purpose is to identify proteins and establish their function. To this end, proteomics also involves the analysis and definition of the many protein-protein interactions responsible for biological processes that are regulated by multiple proteins. A better molecular understanding of the physiological role of these processes will enhance our knowledge of the pathological changes cells and tissues are susceptible to. Ultimately, this knowledge will help in the development of new drugs against human and animal diseases.

Before the analysis of proteins can begin they must first be synthesized – a process encapsulated by the central dogma of molecular biology which claims there is a distinct flow of genetic information from genes to proteins. Deoxyribonucleic acid (DNA) contains determined nucleotide sequences, the genes, which encode the information necessary for protein synthesis. The first step in materializing this information is the process of transcription, which produces a copy of the genes in the form of ribonucleic acid (RNA). These molecules translocate from the nucleus to the cytoplasm, where translation of the 'messenger' RNA (mRNA) finally gives rise to proteins.

Although people are becoming increasingly familiar with the terms used in the study of proteins, it may still be useful to clarify a few of the most common ones. A cell's genome refers to all the genes present in the cell, and analysis of the genome is known as genomics. Similarly for RNA, the RNA species generated in a cell are known collectively as the transcriptome, and its analysis as transcriptomics. Finally, all the proteins in a cell are termed the proteome, and analysis of the proteome is proteomics. Proteomics makes no distinction between proteins located in the cytoplasm, mitochondria, nucleus, membranes, or any organelles of a cell.

Different areas of study are recognised within the field of proteomics, of which four will be discussed in detail. The first, classical proteomics, involves the identification of all proteins in a cell or tissue; the second, differential display proteomics, is the analysis of differences between two defined proteomes; the third, the analysis of the proteome's enzymatic activities, is known as chemical proteomics; the fourth is interaction proteomics, which aims to identify all protein-protein interactions in a proteome. The latter is extremely important because the identification and determination of a protein's function are together insufficient for a comprehensive understanding of biological processes at a molecular level. This fact may be explained more clearly by the following analogy.

A collection of the pieces belonging to a jigsaw puzzle reveals they are different to one another in colour and shape; each piece represents an individual protein. Since proteomics is the analysis of all proteins, each piece of a given puzzle is defined according to its colour and shape. Thus, the pieces may be sorted according to their colour, which may represent a function, or functional process, that the proteins participate in. For instance, red pieces could be nuclear proteins, green pieces could represent proteins from a membrane, and brown ones could be proteins that play a role in the cell's metabolic processes. This does not, however, provide informational content about the puzzle. Before defining the interaction between two puzzle pieces (or two proteins), the interaction partners must be defined. When 25% of the puzzle is complete, i.e. 25% of all protein interactions are solved, the information revealed in the puzzle (in the form of a picture) remains unclear. Even after 50% of the puzzle is complete, i.e. 50% of all the possible interactions

between two proteins in a cell, it is still hard to define the content. And so the analogy continues, but, unlike a jigsaw, which when 90% complete can reveal its pictorial content, the protein interactions that define the cause of a disease may be hidden in the remaining 10% of undefined interactions. Clearly then, all protein-protein interactions must be defined before there is any chance of understanding the molecular mechanisms responsible for pathological changes in a cell, and, more importantly, of eventually curing the associated disease. Multi-enzyme or multi-protein complexes regulate all biological processes, and thus it is essential to define these interactions to understand their molecular basis.

Classical proteomics – how does it work?

The methodology used for classical proteomic analysis is one- and two-dimensional polyacrylamide gel electrophoresis (PAGE) – techniques that have been around for as many as 30 years. The first step of the analysis is the extraction of proteins from either one type of cell or from a whole organ. This process is important if all the proteins of a given cell are to be analysed. Two-dimensional polyacrylamide gel electrophoresis (2-D PAGE) is the only method currently available that can simultaneously separate thousands of proteins, and it has made the technique central to proteomics technology. The first of the method's two dimensions is isoelectric focusing (IEF), a process that separates proteins along a pH gradient on which the molecules become stationary when their net charge is 0 (Figure 1). This is a protein's isoelectric point (pI). The charge of a protein is determined by its constituent amino acids, which may be positive, negative or neutral, and the sum of the amino acids' charge gives a protein its overall charge. After isoelectric focusing, the second dimension of the separation process is completed orthogonally using electrophoresis in the presence of sodium dodecyl sulphate (SDS). Surfactant SDS binds to proteins and overrides their intrinsic charge, giving all treated proteins an identical charge density and free solution electrophoretic mobility. The SDS-coated proteins migrate through a polyacrylamide gel and are separated on the basis of their molecular mass (Figure 1). It is the separation of proteins by two independent parameters (charge and mass) that makes 2-D PAGE such a high-resolution technique. Proteome studies became feasible only when mass spectrometric analysis of a single 'spot' in a polyacrylamide gel (representing a single protein species) became a reality, and the technique has been responsible for the advancement of this field in recent years.

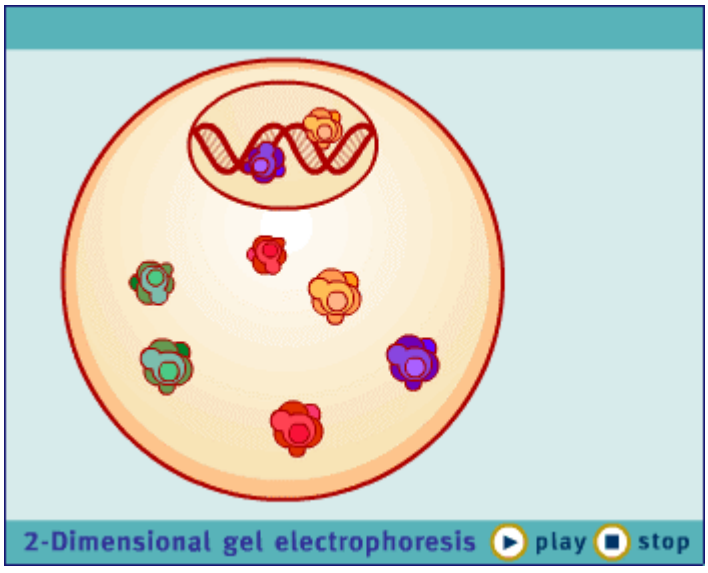

Figure 1. Two-dimensional polyacrylamide gel electrophoresis (2-D PAGE) is a highly sensitive technique used for the separation of proteins extracted from a cell or tissue. The original extraction mixture includes proteins present from the cytoplasm, nucleus and cellular membranes. The first dimension of the separation procedure is isoelectric focusing, which separates the proteins on the basis of their charge. The second dimension uses electrophoresis, in the presence of sodium dodecyl sulphate (SDS), to separate the proteins on the basis of their mass. Proteins of a single species are located in the same position on the polyacrylamide gel and may be excised from the gel for subsequent analysis by mass spectrometry. 2-D PAGE is central to proteomics research.

Identification of a protein involves the definition of one or more unique characteristics, or attributes, which are matched against protein databases. Protein characteristics are referred to either as primary or secondary – a primary characteristic is a property of the intact protein (or derived from it), a secondary property represents, or arises from, fragments of the whole molecule. Most attributes relate directly or indirectly to a protein sequence, but vary in the way in which they are generated and in the protein properties they represent. They also vary enormously in how useful they are as unique identifiers. Their most important feature is the speed with which they enable identification of a protein, because of its effect on throughput in analytical procedures. For detailed technical notes on protein identification methods the reader is referred to a book on the topic [5].

Two-dimensional PAGE analysis is currently the most powerful technique for the simultaneous study of protein expression and post-translational modification. It is not possible using the current methodology, however, to detect on a single 2-D gel all the proteins expressed by a genome. It remains to be elucidated how many copies of a protein are needed before it can be detected on a gel. Protein spots (each one of which represents a different protein species) are visualised by applying one of several different staining methods, the sensitivity of which are determined by the number of proteins that can be detected in any one spot. The highly sensitive silver stain still requires the presence of one thousand protein copies per cell, approximately 0.17 pmol protein; subsequent analysis of the protein, however, is difficult. Immunoblotting, which uses a combination of high affinity monoclonal antibodies and enhanced chemiluminescence (ECL), will allow detection of proteins with as few as 10 copies per cell, corresponding to as little as 1.7 fmol protein. Thus, the proportion of the full complement a protein extracted from a single cell type or tissue that can be detected by 2-D PAGE depends on the copy number, the quantity of the protein loaded onto the gel, and the method of detection. Specific sub-cellular fractions may be enriched prior to their loading on a gel to increase the number of detectable low abundance proteins. For relatively small genomes, such as those derived from baker's yeast (*Saccharomyces cerevisiae)* which contains 6,034 genes, and *E. coli* that has 4,285 genes, up to 40-50% of their proteome can be displayed on a polyacrylamide gel that has the dimensions 160mm x 180mm x 1.5mm. Using the same type and size of gel, no more than 20% of expressed genes from mammalian cells are detectable, representing 50,000 genes and corresponding to about 10,000 -15,000 polypeptides or isoforms. Many proteins with low copy numbers have important regulatory functions in cells; their separation in quantities sufficient for subsequent analysis is important in proteome studies, and represents a challenge to the two-dimensional separation techniques.

Proteins in a proteome can undergo different co- and post-translational modifications, all of which influence a protein's charge, hydrophobicity, conformation and stability. The one-gene-one polypeptide paradigm has consequently become outdated. In both eukaryotes and prokaryotes the polypeptide translation of many genes is modified to create multiple gene products from a single gene sequence. Numerous different types of modification have already been identified, but most have not been localised on a sufficiently large number of proteins to allow the construction of extensive modification-specific databases that can then be used to predict all modifications from the analysis of gene sequences. Once again, 2-D PAGE demonstrates its usefulness in its ability to separate many of the protein isoforms generated by co- and post-translational modifications. This is an important feature for proteome projects and provides scientists with a powerful tool to investigate how co- and post-translational modifications influence protein structure and function, and to determine whether the expression of particular isoforms is under developmental, or disease, control. Protein isoforms often produce a trail of spots in 2-D PAGE because of differences in their pI and/or apparent mass, which is an indication of glycosylation or phosphorylation.

Differential display proteomics

Differential display proteomics is the study of differences between two proteomes. Most physiological and pathological processes are associated with quantitative differences in gene products or proteins. However, investigations at the DNA or RNA level do not provide a complete picture of these differences, due to the absence of a direct correlation between the abundance of mRNA and the abundance of protein in a cell. Since the development of 2-D PAGE many studies on protein quantification have been undertaken. Polyacrylamide gels are treated with a silver stain and the protein spots are localised by scanning the gel. Photographs of different gels are compared to identify and analyse the differences. For example, a normal, healthy tissue produces its own unique proteome, called Proteome 1. The same tissue may become infected by a bacterium that alters the proteome, giving rise to Proteome 2, but when a virus infects the tissue, different set of genes are activated and a different protein pattern is produced, this is Proteome 3. Quantitative pattern analysis reveals differences between Proteomes 1 and 2, and between Proteomes 1 and 3. Analysis of the spots to identify the quantitative differences in proteomes provides information about the function of the proteins during bacterial and viral infections. Some proteins may be highly expressed during an infection while the synthesis of others may be reduced or inhibited. The reproducibility of 2-D PAGE reference maps, using two-dimensional gels, has prompted several groups to make their maps available on the World Wide Web. Examples may be found on the Expert Protein Analysis System (ExPASy) proteomics server of the Swiss Institute of Bioinformatics (SIB). Not all the differences between infected and uninfected cells, or tissues, can be analysed, however.

A great deal of research has been carried out on the analysis of tumours in an attempt to find markers specific to the different tumour types. Markers are proteins detected in higher concentrations in transformed tissue than in normal, healthy tissue. Scientists hope to find some markers that will confirm the presence of a tumour, and others that will provide prognostic possibilities. However, difficulties remain in the staining of proteins to determine their concentration, and until these are resolved the 2-D PAGE databases cannot be extended to consider protein abundance as a component of pathological responses.

Chemical proteomics

In this proteomics approach, all cDNA's from a given tissue are fused to a tag, such as glutathione-S-transferase (GST) or histidine, and are expressed and subsequently attached to a matrix. This matrix can be a so-called chip. Once a chip is established with thousands of proteins expressed in a tissue, all the fusion proteins are analysed for their ability to bind chemical substances and catalyse chemical reactions. The aim of this approach is to assign progressively a specific function to each protein. The disadvantage of this analysis is that the necessary experiments are conducted *in vitro*, not in the protein's natural environment, which may influence the measured enzymatic activity of a protein.

Interaction proteomics

Interaction proteomics is a very important area of proteomics. Several methods have been established to identify protein-protein interactions within a proteome, one of which is the yeast two-hybrid system (Figure 2). The first step in this system is to introduce into a modified yeast cell (reasons for the modification are given later) the gene encoding a specific protein of interest. The yeast cell is able to activate this gene and generate the mRNA as well as the protein encoded by the gene. The modified yeast cell is propagated to produce a population of identical cells, into which the genes of all the other proteins of this cell are introduced. For technical reasons, each yeast cell will usually take up only one gene. Due to the presence of this second gene the yeast cell now produces a second protein. If the two introduced proteins interact with one another the

earlier modification of the yeast cell causes it to turn blue; if, on the other hand, they do not interact the cell remains unchanged . The blue yeast cell is isolated and propagated so that the second of the introduced genes can be isolated (Figure 2). The technique has enabled the identification of a new interacting partner for the protein of interest. Using the yeast two-hybrid system, millions of different protein-protein interactions can be tested in one experiment. Its limitation, however, is that only two protein combinations can be analysed at once, thus if a complex contains four, five, or ten different proteins the analysis can only be completed by testing one protein after the other.

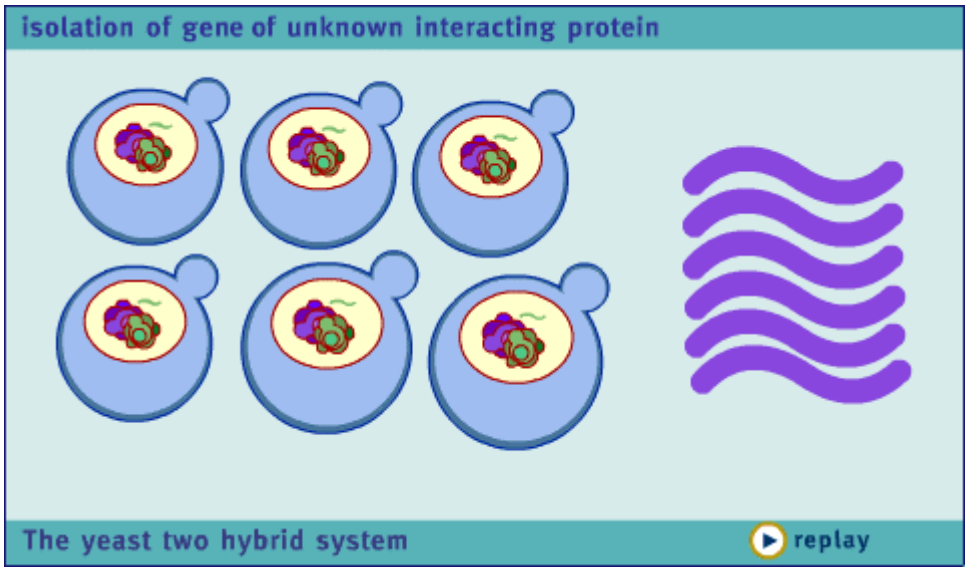

Figure 2. The yeast two-hybrid system is used in interaction proteomics to identify protein-protein interactions in the proteome. The gene of a protein of interest is incorporated into the genome of a modified yeast cell, which subsequently produces the associated protein. The modified yeast cell, containing the gene of interest, is multiplied to produce a population of identical cells. The genes of all the other proteins are introduced. Due to technical reasons each yeast cell takes up only one of these genes. The second gene is incorporated into the cells' genome and a second protein is produced. The second protein either does (blue yeast cell) or does not (neutral-coloured yeast cell) interact with the protein of interest. After multiplication of the blue yeast cell, the gene of the new interacting protein is isolated.

In summary, the different proteomic approaches use different methods, all of which have their technical limitations. It should become a goal for scientists to develop new techniques, and to simplify the existing ones so that they become standard laboratory methods.

Making use of proteomics

For a gene to translate its information into a protein it must first be activated, either conservatively or upon stimulation. The proteins that regulate gene expression are called transcription factors. To illustrate the usefulness of proteomics, the following example is taken from an investigation of the transcription factor nuclear factor kappa B (NF-κB). This transcription factor, which consists of two subunits with molecular weights of 65 kD and 50 kD, respectively, binds to specific sequences in front of the genes [1], but is unable to bind to DNA in the non-stimulated cell because it is sequestered in the cytoplasm in a tight comlex together with an inhibitor of NF-κB, called I-κB [1]. Treatment of cells with extracellular stimuli, including cytokines, such as IL-1 or TNF-α, bacterial lypopolysaccharides (LPS) and phorbol esters (PMA), and the potent oxidants ultraviolet (UV) and gamma irradiation, lead to the rapid phosphorylation of I-κB by a high molecular weight I-κB-kinase complex [3]. Phosphorylation leads to the ubiquination of I-κB and to its subsequent dissociation from NF-κB, which allows NF-κB to translocate to the

nucleus and bind to specific binding sites on the DNA. Once bound to the DNA, NF-κB activates the genes that encode a number of different proteins known to be important to the immune system – they are involved in inflammatory processes, cell-cell interactions, stress responses, and are regulators of apoptosis (programmed cell death) [4]. Thus, the transcription factor plays several crucial roles in the regulation of genes involved in immune and inflammatory responses in mammals. The NF-κB response occurs in virtually all cell types in combination with a variety of co-activators. Because NF-κB alone is not capable of activating its genes when bound to DNA, it is no surprise that the exact genes activated will also vary depending on the cellular context. These co-activators are believed to link enhancer-bound transcription factor-like NF-κB to components of the basal transcription machinery, which then transcribe the gene to generate the mRNA copy. The aim of the investigation was to identify these different components by using proteomic approaches.

Before analysing the components of the NF-κB complex, they had to be isolated. Only after induction of the transcription factor, so that it moved into the nucleus, was a nuclear extract prepared. The extract contained a mixture of all the proteins present in the nucleus, including NF-κB in its complex. The complex was isolated from the other proteins by using oligonucleotides (short DNA sequences that contain the binding sites for NF-κB), which were biotinylated so that they could be attached to avidin-coated beads. The nuclear extract preparation was incubated with the beads, which lead to the trapping and sequestration of the complex through the attachment of NF-κB to its binding site on the oligonucleotide. Several washing steps gave rise to a clean preparation of the complex.

The isolated complex was analysed using 2-D PAGE. For the first of the two dimensions the proteins were denatured and separated according to their charge. Subsequent treatment with SDS gave all proteins a negative charge, so that in the second dimension they were separated according to their mass. Finally, the constituent proteins of the complex were isolated and identified by mass spectrometry. The biological significance and physiological function of each component in the complex will be verified by further analysis.

What is the importance of this study? The results have increased our knowledge of gene regulation and the way in which genes are activated. Critical, and previously unknown, components of a gene regulatory process have been identified, and can be analysed further. It is known that NF-κB can be induced by a number of different stimuli so it should be possible to establish exactly which components are induced by these stimuli. It is not yet clear which sets of proteins are present in the complex after induction of NF-κB. The results have provided a better understanding of the processes regulated by NF-κB, through its activation of different genes and their proteins, and are known to include changes to cell physiology, such as induced cell death (apoptosis), immunological reactions, such as asthma or arthritis, and the induction of cancers.

Why do we need proteomics?

Today we are living in a so-called post-genomic era. We have sequenced the genome of dozens of bacteria and virus species, yeast (*Saccaromyces cerevisiae*), and recently also of man. Without genomics these achievements would not have been possible. DNA contains four different nucleosides – adenosine (A), guanosine (G), cytidine (C) and thymidine (T) – and the evolution of genomics has enabled us to determine the sequence of these nucleotides in a given genome. The question is, now that we have all this information why do we need proteomics?

A simple example will help explain its importance. A sequence of letters, which in this example are random letters of the alphabet, represents a random sequence of the different nucleotides:

thevetsaysthedogowneristhebestoneonthisplanet

Applying genomics the sequence of these letters can be defined, but the real challenge is in determining the information that they encode. Soon computers are likely to analyse the genes in a sequence, by looking for repetitive sequences or by comparing genomes of different species. In the present example, analysis of the sequence of letters reveals the following content:

the vet says the dog owner is the best one on this planet

Analysis of this sequence would identify the first word 'the' as a repetitive one, because it occurs a further two times in the given sequence. 'The' appears to have a regulatory role, while 'dog' and 'planet' appear to be genes. However, at first glance it is not easy to define the information in the sequence. Its interpretation is dependent on the accent given to the words, or genes, when reading the series. There are two ways in which the sequence can be read, both of which have a different meaning. With the help of punctuation marks, the two meanings are indicated below:

the vet says, that the dog owner is the best one on this planet

the vet, says the dog owner, is the best one on this planet

Continuing the analogy, let us imagine that the first of these two interpretations is the way in which the genes are pronounced and accented in a muscle cell, while the second example is found in a liver cell. Although both contain exactly the same sequence of letters, packaged into the same genes, the muscle cell expresses different genes to those in the liver cell. Differences in expression give rise to cell specificity. Although the cells contain exactly the same information at the DNA level, they differ at the protein level and therefore have different proteomes.

The analysis of all genes – genomics – is context independent because fluctuations in DNA are extremely small and do not significantly affect the genes. The estimated number of genes for higher eucaryotes is 30,000, based on the human genome. The analysis of RNA – transcriptomics – is context dependent because genes can be spliced differently to give different species of mRNA. This increases the estimation for the number of genes to 90,000. Proteomic analysis is more context dependent due to the co- and post-translational modifications that can occur. The proteome, unlike the genome, is not a fixed feature in an organism – it changes with the state of development, the tissue and even with the environment in which an organism is living. Hence, there are many more proteins in a proteome than genes in a genome, especially in eucaryotes. This fact makes one of the famous dogmas in biology, the one-gene-one-enzyme hypothesis of Beadle and Tattham, no longer tenable [2]. The estimated number of proteins generated from 30,000 genes is 600,000, and from studies on the human genome it is believed that approximately 25% of all proteins and their functions are known. Functions cannot be assigned to the remaining 75% of proteins and their respective isoforms. Clearly, very little is known at a molecular level about the processes taking place in a cell.

In an attempt to remedy this situation, various efforts have been undertaken by different organisations. The Faculty of Veterinary Medicine at the University of Zurich, for example, has established a new Proteomics Center, which provides groups within the faculty access to proteomic technology in their respective research projects. A second initiative at the University of Zurich is the recently founded company DUALSYSTEMS BIOTECH, a spin-off from the Veterinary Biochemical Institute. The aim of the company is to provide a genetic platform for protein-protein interactions, and it offers numerous services, such as the yeast 2-hybrid system for the identification of protein-protein interactions. The company's research and development section is currently working on the development of a new screening system for membrane proteins.

Analysis of the human genome, and those of some animal species, reveals there are approximately 30,000–40,000 protein-coding genes (See Nature's Genome Gateway). Presently only 500 proteins are being used as therapeutic targets, yet an estimated 8,000 proteins are believed to have therapeutic potential. The difference in these figures represents a huge commercial opportunity for the pharmaceutical industry, which has an interest in the rapid

identification and analysis of the entire proteome of a given cell to identify other therapeutic targets and to develop new drugs. The instrument to develop these drugs is proteomics.

Proteomics will determine our future – how?

A molecular understanding of all the information present in any given genome requires an interdisciplinary approach. Research in many different fields, such as structural biology, cell and molecular biology, biochemistry, genetics and informatics, will all need to contribute, as will the veterinary practitioner. Only if attention is focused on determining the functions and interactions of proteins will their biological significance be revealed.

A combination of the different approaches in proteomics will provide a better understanding of physiological processes and how they regulate one another. In turn, this information will lead to a better understanding of pathological changes related to disease. The ultimate goal in medicine is to prevent or cure disease before permanent damage has occurred, or before side effects become evident. Treatment of diseases, if not curative, should provide palliative care and relief from the symptoms. The choice between a preventive course of action or a treatment relies on a clear recognition of the exact disease status of the patient. A diagnostic procedure helps to establish a patient's condition and to classify patients into categories. To date, a patient's diagnosis has always been carried out by physicians, who take the patient's history, perform a physical examination and, if necessary, they will analyse samples, for instance blood. Sometimes more sophisticated procedures are carried out. In the future, however, proteomics will help define and improve the process of disease diagnosis. For example, markers could be used to define certain tumours so that a diagnosis can be established, a prognosis defined, and a treatment decision made. Nowadays, the prognostic knowledge, aided by diagnostic categories, is inadequate and tailor-made prognostic evaluation is increasingly required for individual patients. For example, the metastatic potential of a tumour is most likely to be unique to a given patient. Tumours usually evolve over a period of time and can be quite different from one patient to the next, even when they have identical clinical and pathological findings. The complexity of interactions between the environment, genes and their products is tremendous. An approach complementary to genomics is required in clinical situations to better understand epigenetic regulation and to provide a more holistic approach in medicine. The application of proteomics for the analysis, for example, of body fluids and tissue biopsies would be particularly useful in disease diagnosis, and could offer valuable prognostic possibilities. New proteomic approaches will also enable the identification of new therapeutic targets, which could be used to identify small chemical compounds in high-throughput screening procedures. They can also be further developed for new drugs, which will allow the generation of new therapeutic concepts for the benefit of man and animals. Finally, new molecular understanding of how biological processes are regulated in cells will allow the development of diagnostic tools and drugs for diseases, which are as yet unknown.

Acknowledgements

Professor M.O. Hottiger would like to thank the following members of the Institute of Veterinary Biochemistry, University of Zürich, Switzerland: I. Stagljar and S. Hasan for their critical reading of the manuscript, and R. Imhof and P. Hassa for their technical assistance, and other members of the Institute of Veterinary Biochemistry who provided helpful advice.
Professor M.O. Hottiger is grateful for the support of the Kanton of Zürich.

References

1. Baldwin, A.S. Jr. (1996) The NF-kappaB and I kappa B proteins: new discoveries and insights. Annu. Rev. Immunol. 14, 649-83.

2. Beadle, G.W. and Tatum, E.L. (1941) Genetic control of Biochemical Reactions in Neurospora. Proc. Natl Acad. Sc. 27, 499-506.

3. Karin, M. and Ben-Neriah, Y. (2000) Phosphorylation meets ubiquitination: the control of NF-kappaB activity. Annu. Rev. Immunol. 18, 621-63.

4. Li, N. and Karin, M. (2000) Signaling pathways leading to nuclear factor-kappa B activation. Methods Enzymol. 319, 273-9.

5. Link, A.J. (1998) 2-D gel proteome analysis protocols. Ed. A.J Link. The Humana Press Inc., Totowa, New Jersey, USA.