Protein expression analysis: From 'tip of the iceberg' to a global method

Peter James
Wallenberg Laboratory II, Lund University, P.O. Box 7031, SE-220 07 Lund, Sweden
Tel.: +41 765 585802;
E-mail: peter.james@elmat.lth.se

In this review I will describe the advances that have recently been made in 'traditional' two-dimensional gel based protein expression analysis. A major jump has been made toward the automation of gel image analysis and comparison, one of the major bottlenecks in the analysis chain as well as the automation of spot excision and preparation for mass spectrometric analysis. Currently the gel-based 'proteome mapping' approach is highly effective and 300 gels and over 10,000 spots a week can be analysed. Very recently, viable alternatives to the use of two-dimensional gel electrophoresis have emerged and these approaches are discussed here. In combination with the recently developed stable isotopic tagging methods for peptide quantitation and new mass spectrometers, this emerging technology will be a rapid and highly effective alternative to gel-based methods with few of the latter's shortcomings.

1. Introduction

1.1. The starting impulse, whole genome availability

Recently biological vernacular has been expanded with a series of 'omes': The genome; the DNA sequence of an organism, the transcriptome; the mRNA being expressed at a given time in a cell and the proteome; the protein equivalent. The latest in the family has been dubbed the metabolome and is a catchall term for all small molecules that are a product of enzymatic and chemical activity within the cell. In contrast to the genome, which is fairly inert, the latter three molecular groups are highly dynamic and vary greatly according to the endo- and exogenous conditions and throughout the life cycle of an organism. The human genome for example consists of the forty-six chromosomes, which encode somewhere between 30 and 100,000 genes. These can either directly transcribed 1:1 or can be recombined into various different combinations by gene rearrangements as with T-cell receptors and immunoglobulins. The nucleotide sequence of the human genome (at the first draft) is now available in databases, yet only a small fraction of the genes found have a known role. The mouse genome is roughly the same size as that of the human, about 3.1 billion base pairs and as of March 2001, the coverage of the public area mouse sequencing effort was around 95% with x3 coverage. This will greatly facilitate the interpretation of the human sequence since only about 5% of the human genome contain genes and the gene sequences in mouse and human that encode the same proteins show a high degree (85%) of sequence identity. The DNA sequences in the vast regions between genes are much less similar (50% sequence identity or less). The availability of complete genome sequences and extensive Expressed Sequence Tag (EST [1]) libraries potentially allows the entire potential protein complement of organisms to be defined. Interest is now focussed on trying to interpret the massive influx of new sequence data and to understand how the vast array of chemical species in the cell interact with one another to create the molecular machinery of the cell. The focus of biological problem solving must now move from a reductionist to a global approach and methodologies must be developed to allow genome wide monitoring of gene expression at the mRNA, protein and metabolite (protein activity) levels.

1.2. The development of genome-wide expression studies

The dynamic expression of genes as mRNA, the transcriptome can be followed both in a quantitative and qualitative manner. This has been made possible by the development of a variety of mRNA expression analysis methods that allow genome wide studies [2,3] to be carried out. A prerequisite to these large scale mRNA expression studies and to the sequencing of genomic DNA is a high degree of automation. The key to the development of these large scale mRNA expression stud-

Disease Markers 17 (2001) 235–246
ISSN 0278-0240 / $8.00

ies were technological advances in analytical biochemistry such as the development of polymerase chain reaction amplification (PCR), shotgun sequencing, and fluorescent-tagged DNA capillary sequencing [1,4,5]. In order to obtain the reproducibility and data accuracy at the high-throughput levels necessary for the assembly of these complex data sets, the process had to be automated by robotics and new algorithms for data assembly and sequence evaluation had to be developed. The combination has made the genome sequencing not only possible, but also almost routine. A 128 capillary array DNA sequencer can produce 128,000 bases per run, giving an output of 3 million bases, -equivalent to a bacterial genome per day. Reproducibility is essential for a statistical analysis of these complex data sets and automation and high throughput data accumulation are essential for such large undertakings. DNA and mRNA are physico-chemically very homogenous and 'easy' to handle, can be amplified by polymerase chain reaction methods and are hence amenable to automation. mRNA analysis methods such as DNA hybridisation arrays and Serial Amplification of Gene Expression have made the quantitative and qualitative analysis of mRNA into an extremely high-throughput technique.

1.3. The old 'newcomer' proteomics

The same cannot be said for methods for global analysis of the protein complement of the cell, which is in its infancy. Since proteins are vastly more physico-chemically diverse than nucleic acids, a universal separation method is unlikely to be found and this is further compounded by the lack of an amplifying method analogous to PCR. The only partially satisfactory methods for analysing the state of expression of the majority of proteins in a cell are those based on two-dimensional polyacrylamide gel electrophoresis (2D-PAGE [6]). Proteins are separated in the first dimension according to their isoelectric point, i.e. by migrating to a point in the gel where the pH causes the net charge on the protein to become neutral. In the second dimension they are separated according to their mobility in a porous gel, which is proportional to the amount of detergent, Sodium Dodecyl Sulphate bound which is approximately mass dependent. However 2D-gel technology suffers many drawbacks. The separation on a single gel can show up to 10,000 species, however many of these are due to post-translational modifications and the number of gene products being visualised is probably only of the order of 1,000. The increase in reproducibility that has been brought about by the introduction of commercial immobilised pH gradient first dimension gels (IPG [7]) allows very accurate and quantitative comparative 2D gel mapping. Detailed 'proteome maps' can be created with advanced computer imaging programs and then analysed by subtractive or cluster methods to find relationships between the protein spots. The weakness of 2D-PAGE lies in it's inability to deal with certain classes of proteins, mostly highly hydrophobic ones (membrane and cytoskeletal especially) and those with isoelectric points at either extreme of the pH scale (such as acidic hyperphosphorylated and alkaline DNA binding proteins). There are also problems with quantitation due to the low dynamic range of stains. These and other problems must be solved before proteomics can truly become a global approach. I will address the advances being made in the 2D field and then compare these with the new non-gel based methods under development.

1.4. What is proteomics?

Before going in to detail as to how protein expression analysis can be automated, one should define the term Proteomics. Originally it was defined as the protein complement of the genome, however since the whole genome is never expressed in a cell, a more restricted definition must be used. The proteome is the set of gene products and their covalent modifications that occur within a given type of cell at a specific stage and time in its development. Proteome analysis can be subdivided into expression proteomics which analyses protein expression and modification and cell-map proteomics which attempts to define all protein-protein interactions occurring in a cell under given conditions [8]. Expression proteomics relies heavily on quantitative 2D-PAGE to map protein expression in defined cells and is used to follow how protein expression changes in response to perturbation, be it genetic modification or environmental. Cell-map proteomics can be carried out either by high-throughput genetic screening using two-hybrid systems [9] or by isolation and characterization of protein complexes.

These new methods for proteome and gene expression analysis are quantitative and will allow new systematic approaches to investigation the function and regulation of unknown genes. N.L. Anderson has defined three major areas for the analysis of gene function and regulation: molecular anatomy (protein composition of cells and tissues); molecular pathology (analysis of disease in terms of changes in protein expression and modification); and molecular pharmacol-

ogy/toxicology (the effects of drugs and xenobiotics on protein expression and modification). A fourth area, molecular physiology, can be added, the change in protein expression in response to changes in the cells micro- or macro-environment.

2. Why bother with proteomics?

In order to fully understand the workings of such a complex system as a cell, global analyses (both spatial and temporal) of transcription, translation, post-translational modifications and metabolites must be carried out. There is an obvious need to complement the well-established genome-wide mRNA expression methods with global analyses of protein expression and post-translational modification [10,11]. There have been very few comprehensive analyses of the correlation between mRNA profiles and protein expression in any biological system [12–14]. The initial evaluations seem to indicate that there is only a significant correlation between mRNA and protein levels for half of the genes being expressed. The reason(s) for this discrepancy is entirely unknown at the moment. There are several key objections to the reduction of biological studies to following changes in mRNA: (i) the level of mRNA does not allow one to predict the level of protein expression, (ii) protein function is controlled by many post-translational modifications, and (iii) protein maturation and degradation are very dynamic processes which dramatically alter the final amount of active protein independent of mRNA level. A large-scale protein expression study would be an invaluable aid to understanding this phenomenon as well as for identifying markers missed by mRNA studies. These studies can also indicate defects in cell signalling mechanisms by showing the changes occurring, for example, in phosphorylation patterns. This would be an important tool in understanding the mechanism underlying the development and progression of a disease. Finally, in a similar vein to analysing to the relationship of mRNA to protein, the level of metabolites in a tissue is only partially related to the protein expression profile [15]. The analysis of metabolite profiles may provide a very useful tool for diagnostics and prognostics [16].

The success of the genomes projects when measured by the sheer amount of sequence data that has been generated is immense. However the number of genes for which a function can be assigned is rather meagre and hence the discipline functional genomics was created to describe the analysis of gene expression and function. The genome of the yeast *Saccharomyces cerevisiae* contains at least 6,200 genes. Despite intensive genetic work over the past years, 60% of yeast genes have no assigned function and half of those encode putative proteins without any homology with known proteins [17]. In order to describe the functions of the yeast genes, a systematic large-scale approach is being taken using a combination of mutant generation and analysis by transcriptomics, proteomics and metabolomics [18].

3. Advances in the automation of 'traditional' two-dimensional gel electrophoresis based analysis

3.1. Prefractionation and gel-running

The extremely high degree of complexity of eukaryotic tissues often requires that a pre-fractionation step be carried out in order to reduce the complexity and allow the resolution and analysis of minor components. When dealing with a mixed cell population such as a tissue, pre-fractionation of cells using a fluorescence activated cell sorter [19] can allow small sub-populations to be specifically isolated, greatly increasing the sensitivity of the analysis. Similarly, preconcentration of the proteins to be analysed can be carried out using methods orthogonal to 2D gel separation such as native PAGE [20] or by an affinity preenrichment such as heparin chromatography for DNA binding proteins [21], immobilised metal ion affinity chromatography for phosphoproteins [22] or by antibody precipitation to select for a specific protein complex. Alternatively a series of increasingly powerful solubilising buffers may be used to obtain a series of protein fractions [23] or the various cell compartments and/or organelles may be isolated [24]. All of the above methods can be automated and before any large-scale study is started, a systematic study of sample preparation reproducibility should be carried out.

An alternative to prefractionation, if sufficient material is available, is the use of a series of overlapping narrow pH range first dimension strips. This allows greatly increased loading amounts and greater separation efficiency. These are called zoom gels and can be used to determine the pI of a protein to within 0.001 pH units on a narrow pH range gel covering say 0.5 pH units over a 20 cm separation range. A non-orthogonal approach which can be combined with zoom gels is the prefractionation of large amounts of cell extract into defined pH ranges using isoelectric membranes mounted

to form a series of pH defined chambers [25]. This is commercially available under the name IsoPrime and was intended for the purification of individual proteins but has proven a useful first step for zoom gels which can save large amounts of material which would otherwise be lost off the ends of the gels as well as improving the loading and separation of the zoom gels.

Up-to-now, no fully automated commercial method is available for running multiple 2D gels. A detailed study using a prototype semi-automated instrument clearly showed the advantages of mechanical reproducibility [26]. Until recently the only developments in large scale automated production of 2D gels has been carried out in commercial enterprises such as Large Scale Biology, Proteome Systems and Oxford Glycosciences and given the developments in non-gel based systems, such systems are unlikely to be developed.

3.2. Automated gel matching and analysis

Recently a commercial version of the fluorescence-based Differential Gel Electrophoresis (DIGE) technique described by Unlu et al. [27] has become available. The major advance in the technology offered by DIGE is that the entire gel-image analysis procedure can be fully automated. There is no longer any need for time-consuming manual gel matching or editing. The basis of the technique is the covalent fluorescence labelling of the samples with cyanine dyes prior to electrophoresis and the generation of a universal master gel (see Fig. 1). For example, say an experiment involves the analysis of 10 normal breast tissue and 40 tumour samples. A master pool sample is created by mixing half of the protein extracts from all of the samples together and labelling the mixture with dye 1 (red). Sample 1 is labelled with dye 2 (blue) and sample 2 with dye 3 (green) and after labelling both are mixed with the master sample and loaded on the gel. By using a fluorescence detector, a gel image can be obtained from each sample according to the marker dye and thus intra-gel matching of the three samples is trivial since identical spots occur always at the same place since the dyes have almost identical masses and the same charge. Since each gel contains a master image which contains all possible protein spots found in all samples, inter-gel matching becomes trivial and can carried out automatically without user intervention. This technique coupled with the use of zoom gels will greatly extend and speed up the traditional approach to gel-based proteomics. A single person can easily run and analyse up to 250 samples a week using a four-dye system.

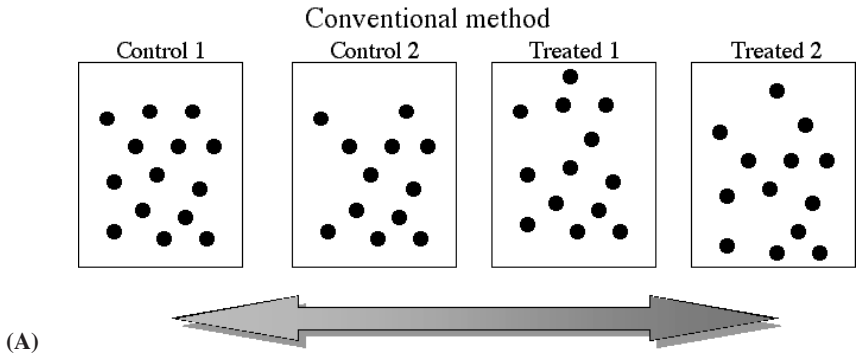
3.3. Spot-cutting and preparation for analysis

Once the data system has matched, quantitated and analysed a gel series, a spot cut-list can be generated based on the criteria fed into the statistical analysis program by the user. Currently there are three commercially available spot-cutting systems available (Amersham-Pharmacia Biotech, BioRad, and Genome Solutions). I will describe the Amersham-Pharmacia approach since it is currently the most fully automated though the others differ only in the degree of automation, and not the approach. The gels are run on a plastic backing to allow ease of handling by robots. After imaging, the stained gels are placed in a 'gel hotel', a temporary storage area which prevents them drying out and cracking. Each gel is picked up in turn by a robot arm and then identified by a bar-coding system and the appropriate cut-list is downloaded from the gel analysis system. The plastic backing has two landmark spots that are at either end of the gel. The spots serve to allow an automatic alignment of the gel once it is placed on the X-Y cutting board with the image of the gel used for the analysis to generate the spot-cutting file. The spots are then cut out with a cutter head and placed in bar-coded 96 well plates. The robot arm transfers the plate to a liquid handling station where the spots are washed (destained if non-fluorescent) and the plate is placed in a drying station. The plate is then returned to the liquid handling station for the addition of enzyme. After digestion the spots are extracted and an aliquot is spotted onto a mass spectrometer target plate for subsequent protein fingerprinting. The rest of the extract is kept cooled in the 96 well plate ready for transfer to an autosampler for HPLC-MS/MS analysis should the protein fingerprinting not deliver a high confidence result. All of this runs in a fully automated manner in a closed environment, allowing around 1,500 spots to be prepared for MS analysis per day.

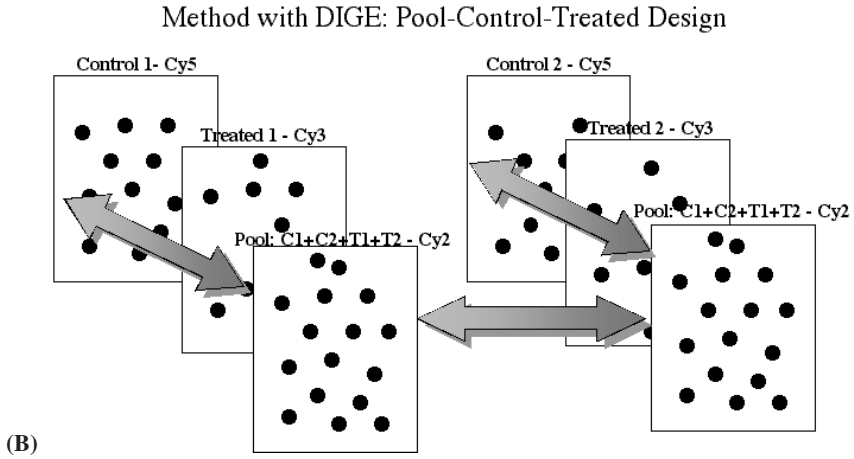
The group of Hochstrasser [28] has described an alternative approach to protein identification on 2D gels. After image analysis, the entire 2D gel is rehydrated with trypsin and allowed to digest before being electroblotted through an immobilised trypsin membrane onto a PVDF membrane. The membrane can then be soaked with matrix and analyses directly by scanning in a MALDI mass spectrometer.

3.4. Hierarchical mass spectrometric analysis

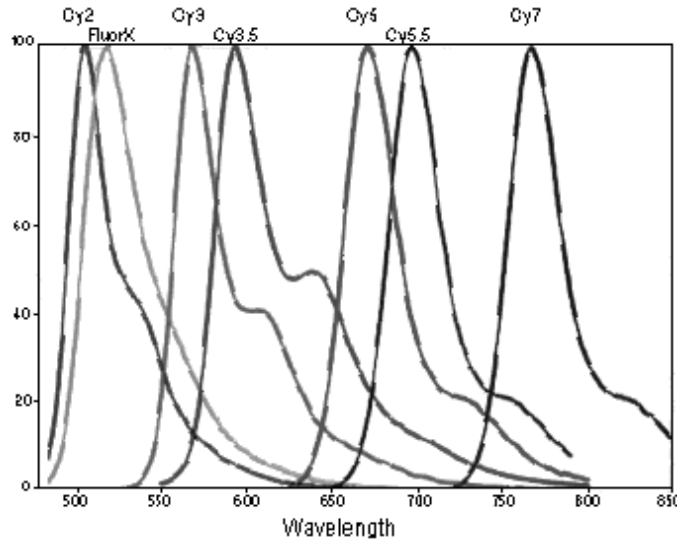
Mass spectrometer manufacturers have also been focussing on increasing throughput. For gel-based pro-



(A)



(B)



(C)

Fig. 1. (A) Using conventional technology, matching all four gels is very time consuming (ca. 4 hours) and difficult as there is a lot of gel to gel variation. Each gel contains a different amount of proteins in different locations and one cannot tell if there are real differences or if it is due to gel to gel variation. (B) Since the samples are running on the same gel, intragel matching is 100% efficient and automatic. Since there is a pool image in every gel, only these have to be matched to linked the samples and these are identical so matching is very easy and can be automated (ca. 1 min.). (C) The excitation and emission spectra of the various cyanine dyes are shown.

teome analysis, a two-tiered approach is most commonly used. An aliquot of the digestion extract (between one third and one tenth of the total) is used for protein fingerprinting by Matrix-assisted laser desorption and ionisation mass spectrometry (MALDI MS). The MALDI target plates produced by the gel cutting system are stacked in a storage array and a robot arm automatically loaded into the mass spectrometer. Spectra are automatically obtained for each position and the proteolytic masses are stripped out and used to search the database for proteins that generate a similar theoretical mass profile. This occurs in real-time and one spot can be analysed/identified per minute, matching the output of the cutter system. Those spots not identified by protein fingerprinting at a high enough confidence level can be automatically scheduled for MS/MS analysis. The plate containing the remaining digest is loaded into an autosampler for injection onto a LC-MS/MS system. Here the throughput is slower and only 96 samples can be analysed/identified a day. However two different non-commercially available systems have been described by the group of Barry Karger [29,30] which allow rapid electrospray sampling in the order of tens of seconds per sample. The rate-limiting factor here becomes the accumulation of enough MS/MS spectra and the database search. Realistically, when applied to protein digests, a throughput of one sample per minute is obtainable.

3.5. Data analysis

Possibly the most critical aspect of the automation procedure is the development of a robust Laboratory Information Management System (LIMS) to deal with the logistics of handling large numbers of samples and collecting and collating the results from the mass spectrometry and gel analyses with the sample types. The LIMS system should also facilitate the scheduling and running of all the samples so that user intervention and error introduction is kept to an absolute minimum. Above this, another layer of software must be available which allows one to interrogate the data from a higher level, cross-matching sets of experiments e.g. match results of 100 sample analysis of liver cirrhosis with a set of results from the result of liver damage by a prescribed medicine. Several firms are developing high-level analysis software based on Oracle data structures. This will allow one to analyse different data types, for example to correlate the findings from a proteomic, transcriptomic and metabolomic analysis of a disease progression and to match these to genotypic and demographic data.

4. The development of non-gel based proteomics approaches

Currently there is no universal method allowing the separation and visualisation of all the proteins and their modifications in a cell [31]. Attempts have been made to carry out multi-dimensional chromatographic separation of proteins, however these methods suffer the same drawbacks as the 2D electrophoresis in not being able to handle low abundance and poorly soluble proteins. However one area in which this approach is useful, is for the analysis of the low molecular weight components of human fluids such as urine, blood, cerebrospinal and synovial fluids. Vast numbers of bioactive peptides have been identified and the development of libraries of peptide profiles may become a very useful diagnostic tool [32,33]. However by digesting the proteins into smaller fragments, they are much easier to handle since they are much more homogenous in their physico-chemical properties [10,13]. This removes the problem of very large or small proteins or membrane proteins since once reduced in size to peptides, a good deal, if not all of the peptides can be separated by standard chromatographic means. Thus sample preparation and handling can be simplified and do not have to be optimised for each cell or tissue type as is the case for 2D PAGE. Also, since HPLC can be directly coupled to mass spectrometers and hence to peptide identification and quantification, the entire process can potentially be completely automated with no user intervention necessary.

4.1. Basic requirements

What is the total amount of protein needed to observe all peptides in a cell and what degree of separation is needed? If one assumes the maximum sensitivity level for peptide detection and MS/MS is 1 fmol. There are thus 6×10^{-23} moles of a single copy protein per cell; hence 1.6×10^{7} cells (0.25 mg protein extract) are needed, assuming no losses, to obtain 1 fmol signal. Thus the first dimension separation will have to be carried out on a 500 μm ID column and the second dimension can then be done with a 150 μm column. Given the human genome is assumed to have 30,000 genes, of which 10% are expressed in any one cell line at a given time and assuming there are on average 20 variants of each protein due to alternative splicing, post-translational modification etc. there will be approximately 200,000 tryptic peptides per cell given an average protein molecular weight of 50 kDa. In order

to avoid too much signal suppression, one should aim to have a separation method that produces individual spectra containing 10 peptides or less. Given 10 fractions from the first dimension and a second dimension flow rate of 200 nl/min, the peak width will be about 5 sec. Thus a single gradient will have to be around 2.7 hours if a maximum of 10 peptides are to be observed per scan on average, giving a total analysis time of 27 hours. If MS/MS analysis is required for all peaks, then the time increases to 100 hours.

4.2. Initial reports on 2D peptide chromatography based proteomics

The basic requirements of such a system are an effective two-dimensional chromatographic separation of peptides with a rapid HPLC-MS/MS analysis and subsequent peptide identification using the fragmentation spectra. The group of John Yates using the yeast ribosome as a model system [34] has made an initial proof of principle that this approach is viable. The first dimension is a strong cation exchange material run in tandem with a second dimension reversed-phase C18 material. The unique feature of this construction is that the separation phases can be packed into the same column, with the first 5 cm being SCX and the last 10 cm up to the nanospray tip being C18 material. The digest material is loaded onto the column and the peptides that do not bind to the SCX column are caught on the C18 column are eluted with a reverse-phase gradient of 0–80% acetonitrile. Then the peptides are eluted from the SCX onto the RP column by successive cycles of step elution with increasing amounts of ammonium acetate following by a reversed phase gradient. The eluent is directly electrosprayed into an ion-trap mass spectrometer programmed to carry out as many MS/MS analyses as possible. A dynamic exclusion rule is built into the analysis that excludes a peptide mass from being analysed by MS/MS more that once in a defined time window to prevent highly abundant peptides swamping the analysis. The method allowed the identification of all of the predicted ribosomal and ribosomal associated proteins (> 100).

Recently, the group has extended this method they term MudPIT (MultiDimensional Protein Identification Technology) to the analysis of the yeast proteome [10]. In a twenty-seven hour chromatography run, 5,540 peptides could be identified by their MS/MS spectra, corresponding to 1,484 proteins, representing a very significant part of the yeast proteome in logarithmically growing cells. The most complete annotated yeast 2D analysis covers only 410 spots corresponding to 282 gene products. The distribution of identified proteins ranged from very low copy number proteins such as transcription factors, through a significantly high number of membrane proteins (131) as well as the extremes of mass (from < 10 to > 550 kDa) and pI (from < 3.9 to > 12.5).

4.3. Expression quantitation by stable isotope labelling

A commonly used technique in protein analysis is the use of radiolabelled amino acids (such as ^{35}S methionine or less commonly cysteine) to increase the detection sensitivity of 2D gel electrophoresis and to allow pulse-chase experiments to determine rates of protein synthesis and degradation. Recently the use of stable isotopes for MS analysis of proteins has been an area of intense interest. One approach is to use isotopically depleted media (i.e. media low in heavy isotopes such as ^{15}N, $^{2 \text{ and } 3}$H and ^{13}C). This allows the isotopic cluster of an ion to be collapsed into a single peak, thereby increasing sensitivity and accuracy [35]. Whole-cell isotopic labelling has been used as a method to quantitate relative changes in protein expression and modification [36]. Cells are grown in either a 'light' medium that consists of compounds with a normal isotope distribution or in 'heavy' medium that is highly enriched in ^{15}N (95%). The protein extracts from the cells grown under different conditions are pooled and partially separated, usually by 2-PAGE. The spot of interest is excised, digested and the extracted peptides analysed by mass spectroscopy. This allows one to differentiate between the peptides originating from the two cell pools since the ^{15}N incorporation shifts one of the pools upwards by one mass unit. The peptides thus appear as doublets and can be quantitated by the relative height of the peaks. This ratio was found to be linear over an abundance ratio of two orders of magnitude. This procedure also lends itself well to defining changes in the level of post-translational modifications. This method is essentially limited by material costs since it is impractical to label entire animals and is often not applicable to eukaryotic cell cultures due to the need for serum derived factors usually obtained from foetal calf serum.

An alternative approach is to use post-separation isotopic labelling for relative quantification. A very promising alternative to 2D gel analysis as a comprehensive method for comparative proteomics was recently describing by the group of Ruedi Aebersold and

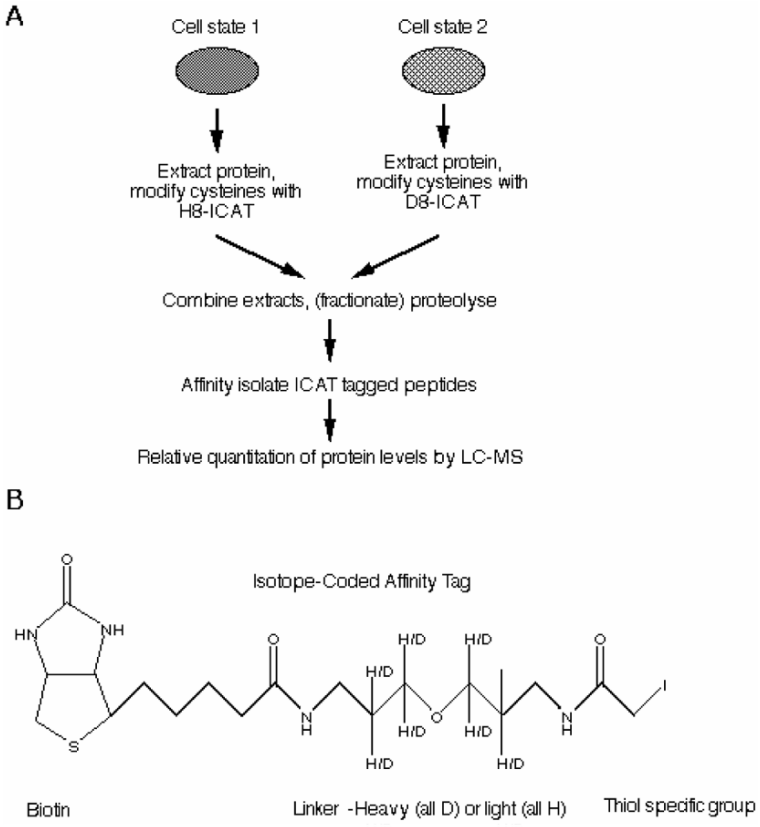


Fig. 2. (A) This shows schematically how the ICAT (isotope coded affinity tagging) technology works. (B) This shows the chemical structure of the isotopic reagents.

is called Isotope-coded affinity tagging (ICAT [37]). Essentially whole cell protein extracts are digested and labelled with either a 'light' or 'heavy' deuterated biotin label which has a thiol specific reactive group (Fig. 2). The mixture is digested and the biotinylated peptides are recovered using an avidin affinity column. This much simpler peptide mixture is then analysed by LC-MS using an RP column. The isotopically labelled pairs of peptides elute almost simultaneously and the isotope peak ratios can give the relative amount of each. It is then necessary to perform MS/MS analysis of the peptides to determine their sequence and thus identify the parent protein. An analogous method for isotopic labelling for quantitation has been described by ourselve for proteins isolated by 2D SDS-PAGE but is can also be used directly for peptide mixtures [38]. The proteins are denatured and then succinylated prior to digestion (Fig. 3). The distinct cell digests are either labelled with D4 or H4 nicotinic acid prior to HPLC-MS analysis. Those peptides showing a change in expression or modification level are then chosen for MS/MS analysis, either in a second HPLC run or dynamically in the first. The combination of any of these isotopic labelling approaches for protein expression quantitation with the MuDPIT technology described above should provide the basis for a broadly applicable non-gel proteomics method and represents the most viable alternative to quantitative 2D-PAGE available today.

4.4. Protein chips

The concept of the proteome, if restricted to the set of proteins being expressed in a cell at any given time, yields a fairly static picture. In reality, proteins can only exert their functions in a cell as a result of highly dynamic interactions with other proteins. The cell can be regarded as a series of interacting molecular machines that are formed from large protein complexes [39]. The spatial and temporal modulation of these interactions is the key to defining cell functions in molecular terms. Mass spectrometry is now being explored as a tool to explore the dynamics of protein interactions. Following changes in the phosphorylation state of proteins,

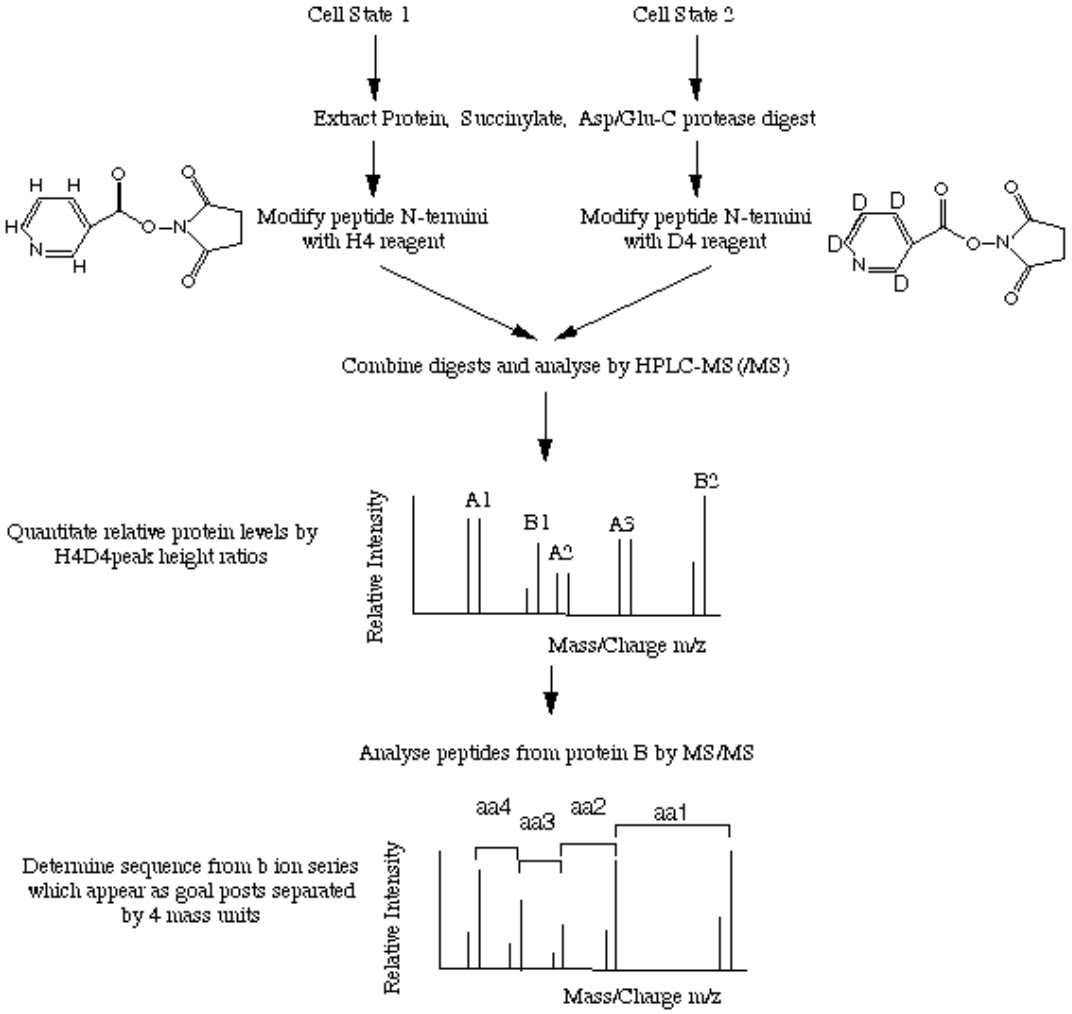


Fig. 3. This show schematically the N-terminal isotopic labelling methodology.

especially those involved in signalling, can help define the set of interactions occurring. A more direct method of defining protein:protein contacts is by direct observation of the complexes and determination of the binding affinities. One method that has recently been developed is surface-enhanced laser desorption/ionisation (SELDI) affinity mass spectrometry [40]. The MALDI target is chemically modified to allow attachment of a 'bait' molecule (analogous to the bait protein used in the yeast two-hybrid system) which is then used to fish for prey proteins, i.e. proteins that bind to the immobilised molecule. The surface can then be washed and the target placed in the mass spectrometer to analyse what has bound to the immobilised molecule. A similar approach has been put forward to map the epitopes of antibodies by immobilising the antibodies on a target and presenting a digest of the target protein. The non-binding peptides are washed away leaving the peptides that form the epitope. The completion of the human genome will allow the development of antibodies against all of the theoretical open reading frames and these should be commercially available in the not-to-distant future. There are many groups working on the development of chip-based systems like those used for mRNA analysis. It should be possible to interface these chips directly with a mass spectrometer and determine which proteins have bound to the bait, beit an antibody, other protein or ligand. In order to identify the bound proteins, either the spot has to be digested which may be problematic if very small amounts of material are bound or the protein can be fragmented directly in the mass spectrometer (see later section on FT-ICR mass spectrometers).

4.5. Mixed approaches

2D-PAGE separates proteins according to their isoelectric points in the first dimension and by mass in the second. Since stable isotopic labelling methods have been introduced for quantitation by mass spectrometry, it is no longer necessary to run a second dimension SDS-PAGE gel for quantitation by staining and scanning. It would seem logical therefore that one should replace the very low mass accuracy and resolution second dimension gel with a high resolution mass spectrometer. Initial steps in this direction have been carried out using direct scanning of first-dimension IPG strips with an Infra Red Laser [41]. If one wishes to identify the proteins, one can then digest the proteins in situ and repeat the MALDI analysis to obtain protein fingerprints [41].

One can even replace the first dimension gel by capillary isoelectric focussing which can be directly connected to a mass spectrometer with an electrospray ionisation interface [42]. Fourier-Transform Ion Cyclotron Resonance mass spectrometry (FT-ICR MS) allows not only high accuracy mass measurement of the eluting proteins as well as quantitation by isotope distribution but the proteins can be rapidly identified directly using MSn techniques. It has already been demonstrated that intact proteins can be identified from their fragmentation in an FT-ICR by using a combination of the exact intact mass with a series of sequence tags extrapolated from the MSn experiments [43]. Recently Li and Marshall have demonstrated on-line identification of proteins by LC/ESI FT-ICR MS. A normal scan is first used to extract the exact mass of the intact protein and on alternate scans infra-red multi-phonon dissociation (IRMPD) is used to fragment one selected m/z ion from the protein. The intact mass is used with a wide mass window to select a subset of the database entries. The list of mostly b- and y-fragment ions, as well as any small sequence tags obtained from IRMPD, is then matched against this set to identify the protein [44]. Jensen et al. [42] described the analysis of *E. coli* cell extracts by capillary IEF-FT-ICR using total injections of only 300 ng of protein, which is equivalent to 3 million bacteria (or 3,000 human cells). 400–1,000 putative proteins were found with a mass range between 2 and 100 kDa. The sensitivity is now coming into the range where it will be possible to analyse individual cells. Why look for the needle in the whole haystack if you know which bale it is in? For example, instead of analysing a whole tissue biopsy, individual cell types can now be isolated by laser capture micro-dissection to select only those cells showing a morphology typical for cancer [45].

5. Summary

The main drawback of the non-gel techniques as described is the lack of quantitation of protein expression levels. However MuDPIT is fully compatible with the isotopic labelling techniques described below and should form the basis for a comprehensive proteome analysis tool. The length of time required for the separation is also somewhat limiting if it is to be extended to human proteome analysis. There are however new mass spectrometers under development that may solve this problem by allowing extremely rapid scanning rates (1000s of spectra per second) which are compatible with high speed chromatographic methods. The new mass spectrometers should also show an increase in the dynamic range of detection (and absolute sensitivity) from the current 3–4 orders to seven orders of magnitude, in line with the range of protein expression found in cells. The combination of fast scanning and chromatography could reduce the time from 27 hours to less than 2 hours within a few years. An alternative to on-line analysis of peptides after multi-dimensional separations has been described by the group of Barry Karger [46] using a vacuum deposition interface for coupling capillary electrophoresis with MALDI-TOF MS. The eluent together with matrix is deposited on a moving tape in the evacuated source chamber of a TOF mass spectrometer. The advantage of the method is that the interesting peptides (determined by isotopic ratios) can be analysed by MS/MS after post-run analysis, greatly reducing the number of MS/MS spectra to be accumulated.

Maybe now we are verging on the edge of being able to harness the flood of information coming from the genome projects, to put it in order using proteome and microarray/SAGE projects, in a way that we may finally see how all the fine threads are pulled together to make the biochemical web which defines life. As an amateur detective once succinctly put it [47]:

"My dear fellow", said Sherlock Holmes, "life is infinitely stranger than anything which the mind of man could invent. If we could fly out of that window and hover over this great city, gently remove the roofs, and peep in at the queer things which are going on, the strange coincidences, the plannings, the cross-purposes, the wonderful chain of events, working through generations, and leading to the most outré results, it would make all fiction with its conventionalities and foreseen conclusions most stale and unprofitable."

References

[1] M.D. Adams, J.M. Kelley and J.D. Gocayne et al., Complementary DNA sequencing: expressed sequence tags and human genome project, *Science* **252** (1991), 1651–1656.

[2] V.E. Velculescu, L. Zhang, W. Zhou, J. Vogelstein, M.A. Basrai, D.E. Jr. Bassett, P. Hieter, B. Vogelstein and K.W. Kinzler, Characterization of the yeast transcriptome, *Cell* **88** (1997), 243–251.

[3] D.A. Lashkari, J.L. DeRisi, J.H. McCusker, A.F. Namath, C. Gentile, S.Y. Hwang, P.O. Brown and R.W. Davis, Yeast microarrays for genome wide parallel genetic and gene expression analysis, *Proc Natl Acad Sci USA* **94** (1997), 13057–13062.

[4] R.K. Saiki, S. Scharf, F. Faloona, K.B. Mullis, G.T. Horn, H.A. Erlich and N. Arnheim, Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia, *Science* **230** (1985), 1350–1354.

[5] J.R. Scherer, I. Kheterpal, A. Radhakrishnan, W.W. Ja and R.A. Mathies, Ultra-high throughput rotary capillary array electrophoresis scanner for fluorescent DNA sequencing and analysis, *Electrophoresis* **20** (1999), 1508–1517.

[6] P.H. O'Farrell, High Resolution Two-Dimensional Electroporesis of Proteins, *J. Biol. Chem.* **250** (1975), 4007–4021.

[7] B. Bjellqvist, K. Ek, P.G. Righetti, E. Gianazza, A. Görg, R. Westermeier and W. Postel, Isoelectric focusing in immobilized pH gradients: principle, methodology and some applications, *J. Biochem. Biophys. Meth.* **6** (1982), 317–339.

[8] W.P. Blackstock and M.P. Weir, Proteomics: quantitative and physical mapping of cellular proteins, *Trends Biotechnol.* **17** (1999), 121–127.

[9] M. Fromont-Racine, J.C. Rain and P. Legrain, Toward a functional analysis of the yeast genome through exhaustive two-hybrid screens, *Nat Genet.* **16** (1997), 277–282.

[10] M.P. Washburn, D. Wolters and J.R. Yates 3rd., Large-scale analysis of the yeast proteome by multidimensional protein identification technology, *Nat Biotechnol.* **19** (2001), 242–247.

[11] H. Zhou, J.D. Watts and R. Aebersold, A systematic approach to the analysis of protein phosphorylation, *Nat Biotechnol.* **19** (2001), 375–378.

[12] L. Anderson and J. Seilhamer, A comparison of selected mRNA and protein abundances in human liver, *Electrophoresis* **18** (1997), 533–537.

[13] S.P. Gygi, Y. Rochon, B.R. Franza and R. Aebersold, Correlation between protein and mRNA abundance in yeast, *Mol Cell Biol.* **19** (1999), 1720–1730.

[14] P.S. Nelson, D. Han, Y. Rochon, G.L. Corthals, B. Lin, A. Monson, V. Nguyen, B.R. Franza, S.R. Plymate, R. Aebersold and L. Hood, Comprehensive analyses of prostate gene expression: convergence of expressed sequence tag databases, transcript profiling and proteomics, *Electrophoresis* **21** (2000), 1823–1831.

[15] O. Fiehn, J. Kopka, P. Dormann, T. Altmann, R.N. Trethewey and L. Willmitzer, Metabolite profiling for plant functional genomics, *Nat Biotechnol.* **18** (2000), 1157–1161.

[16] P. Duez, A. Kumps and Y. Mardens, GC-MS profiling of urinary organic acids evaluated as a quantitative method, *Clin Chem.* **42** (1996), 1609–1615.

[17] A. Goffeau, B.G. Barrell and H. Bussey et al., Life with 6000 genes, *Science* **274**(546) (1996), 563–567.

[18] S.G. Oliver, M.K. Winson, D.B. Kell and F. Baganz, Systematic functional analysis of the yeast genome, *Trends Biotechnol.* **16** (1998), 373–378.

[19] P.S. Madsen, M. Hokland, J. Ellegaard, P. Hokland, G.P. Ratz, A. Celis and J.E. Celis, Major proteins in normal human lymphocyte subpopulations separated by fluorescence-activated cell sorting and analyzed by two-dimensional gel electrophoresis, *Leukemia.* **2** (1988), 602–615.

[20] G.L. Corthals, M.P. Molloy, B.R. Herbert, K.L. Williams and A.A. Gooley, Prefractionation of protein samples prior to two-dimensional electrophoresis, *Electrophoresis* **18** (1997), 317–323.

[21] M. Fountoulakis, H. Langen, S. Evers, C. Gray and B. Takacs, Two-dimensional map of Haemophilus influenzae following protein enrichment by heparin chromatography, *Electrophoresis* **18** (1997), 1193–1202.

[22] J. Porath, J. Carlsson, I. Olsson and G. Belfrage, Metal chelate affinity chromatography, a new approach to protein fractionation, *Nature* **258** (1975), 598–599.

[23] M.P. Molloy, B.R. Herbert, B.J. Walsh, M.I. Tyler, M. Traini, J.C. Sanchez, D.F. Hochstrasser, K.L. Williams and A.A. Gooley, Extraction of membrane proteins by differential solubilization for separation using two-dimensional gel electrophoresis, *Electrophoresis* **19** (1998), 837–844.

[24] N.G. Anderson, Preparative zonal centrifugation, *Methods Biochem Anal.* **15** (1967), 271–310.

[25] B. Herbert and P.G. Righetti, A turning point in proteome analysis: sample prefractionation via multicompartment electrolyzers with isoelectric membranes, *Electrophoresis* **21** (2000), 3639–3648.

[26] M.G. Harrington, K.H. Lee, M. Yun, T. Zewert, J.E. Bailey and L. Hood, Mechanical precision in two-dimensional electrophoresis can improve protein spot positional reproducibility, *Appl Theor Electrophor.* **3** (1993), 347–353.

[27] M. Unlu, M.E. Morgan and J.S. Minden, Difference gel electrophoresis: a single gel method for detecting changes in protein extracts, *Electrophoresis* **18** (1997), 2071–2077.

[28] W.V. Bienvenut, J.C. Sanchez, A. Karmime, V. Rouge, K. Rose, P.A. Binz and D.F. Hochstrasser, Toward a clinical molecular scanner for proteome research: parallel protein chemical processing before and during western blot, *Anal Chem.* **71** (1999), 4800–4807.

[29] H. Liu, C. Felten, Q. Xue, B. Zhang, P. Jedrzejewski, B.L. Karger and F. Foret, Development of multichannel devices with an array of electrospray tips for high-throughput mass spectrometry, *Anal Chem.* **72** (2000), 3303–3310.

[30] C. Felten, F. Foret, M. Minarik, W. Goetzinger and B.L. Karger, Automated high-throughput infusion ESI-MS with direct coupling to a microtiter plate, *Anal Chem.* **73** (2001), 1449–1454.

[31] S.P. Gygi, G.L. Corthals, Y. Zhang, Y. Rochon and R. Aebersold, Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology, *Proc Natl Acad Sci USA* **97** (2000), 9390–9395.

[32] K. Wagner, K. Racaityte, K.K. Unger, T. Miliotis, L.E. Edholm, R. Bischoff and G. Marko-Varga, Protein mapping by two-dimensional high performance liquid chromatography, *J Chromatogr A.* **893** (2000), 293–305.

[33] P. Schulz-Knappe, H.D. Zucht, G. Heine, M. Jurgens, R. Hess and M. Schrader, Peptidomics: the comprehensive analysis of peptides in complex biological mixtures, *Comb Chem High Throughput Screen* **4** (2001), 207–217.

[34] A.J. Link, J. Eng, D.M. Schieltz, E. Carmack, G.J. Mize, D.R. Morris, B.M. Garvik and J.R. Yates 3rd., Direct analysis of protein complexes using mass spectrometry, *Nat Biotechnol.* **17** (1999), 676–682.

[35] T. Solouki, M.R. Emmett, S. Guan and A.G. Marshall, Detection, number, and sequence location of sulfur-containing amino acids and disulfide bridges in peptides by ultrahigh-resolution MALDI FTICR mass spectrometry, *Anal Chem.* **69** (1997), 1163–1168.

[36] Y. Oda, K. Huang, E.R. Cross, D. Cowburn and B.T. Chait, Accurate quantitation of protein expression and site-specific phosphorylation, *Proc Natl Acad Sci USA* **96** (1999), 6591–6596.

[37] S.P. Gygi, B. Rist, S.A. Gerber, F. Turecek, M.H. Gelb and R. Aebersold, Quantitative analysis of complex protein mixtures using isotope-coded affinity tags, *Nat Biotechnol.* **17** (1999), 994–999.

[38] M. Munchbach, M. Quadroni, G. Miotto and P. James, Quantitation and facilitated de novo sequencing of proteins by isotopic N-terminal labeling of peptides with a fragmentation-directing moiety, *Anal Chem.* **72** (2000), 4047–4057.

[39] *Cell* **92** (1998), whole issue.

[40] H. Kuwata, T.T. Yip, C.L. Yip, M. Tomita and T.W. Hutchens, Bactericidal domain of lactoferrin: detection, quantitation, and characterization of lactoferricin in serum by SELDI affinity mass spectrometry, *Biochem Biophys Res Commun* **245** (1997), 764–773.

[41] R.R. Ogorzalek Loo, J.A. Loo and P.C. Andrews, Obtaining molecular weights of proteins and their cleavage products by directly combining gel electrophoresis with mass spectrometry, *Methods Mol Biol.* **112** (1999), 473–485.

[42] P.K. Jensen, L. Pasa-Tolic, G.A. Anderson, J.A. Horner, M.S. Lipton, J.E. Bruce and R.D. Smith, Probing proteomes using capillary isoelectric focusing-electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry, *Anal Chem.* **71** (1999), 2076–2084.

[43] E. Mortz, P.B. O'Connor, P. Roepstorff, N.L. Kelleher, T.D. Wood, F.W. McLafferty and M. Mann, Sequence tag identification of intact proteins by matching tanden mass spectral data against sequence data bases, *Proc Natl Acad Sci USA* **93** (1996), 8264–8267.

[44] W. Li, C.L. Hendrickson, M.R. Emmett and A.G. Marshall, Identification of intact proteins in mixtures by alternated capillary liquid chromatography electrospray ionization and LC ESI infrared multiphoton dissociation Fourier transform ion cyclotron resonance mass spectrometry, *Anal Chem.* **71** (1999), 4397–4402.

[45] M.R. Emmert-Buck, R.F. Bonner, P.D. Smith, R.F. Chuaqui, Z. Zhuang, S.R. Goldstein, R.A. Weiss and L.A. Liotta, Laser capture microdissection, *Science* **274** (1996), 998–1001.

[46] J. Preisler, P. Hu, T. Rejtar and B.L. Karger, Capillary electrophoresis – matrix-assisted laser desorption/ionization time-of-flight mass spectrometry using a vacuum deposition interface, *Anal Chem.* **72** (2000), 4785–4795.

[47] A.C. Doyle, *The complete Sherlock Holmes*, Penguin Books, London, 1930, pp. 190.