Chimia 55 (2001) 354–358
© Schweizerische Chemische Gesellschaft
ISSN 0009–4293

Proteomics: From Protein Identification to Biological function

Jan van Oostrum*, Sjouke Hoving, Dieter Müller, Patrick Schindler, Michel Steinmetz, Harry Towbin, Hans Voshol, and Urs Wirth.

Abstract: The term 'proteome' describes the expressed protein complement of a genome. The largely invariant genome of an individual or organism determines its potential for gene- and protein expression but does not specify which proteins are expressed in the various types of cells in an organism or individual or their level or extent of post-translational modification. Furthermore, the proteomes of cells are directly affected by environmental factors, such as stress or drug treatment, or by aging and disease. This review briefly describes a selection of the main technologies applied by proteome sciences and their applications to study complex biological problems.

Keywords: Mass spectrometry · Protein interactions · Proteomics · Two-dimensional gel electrophoresis

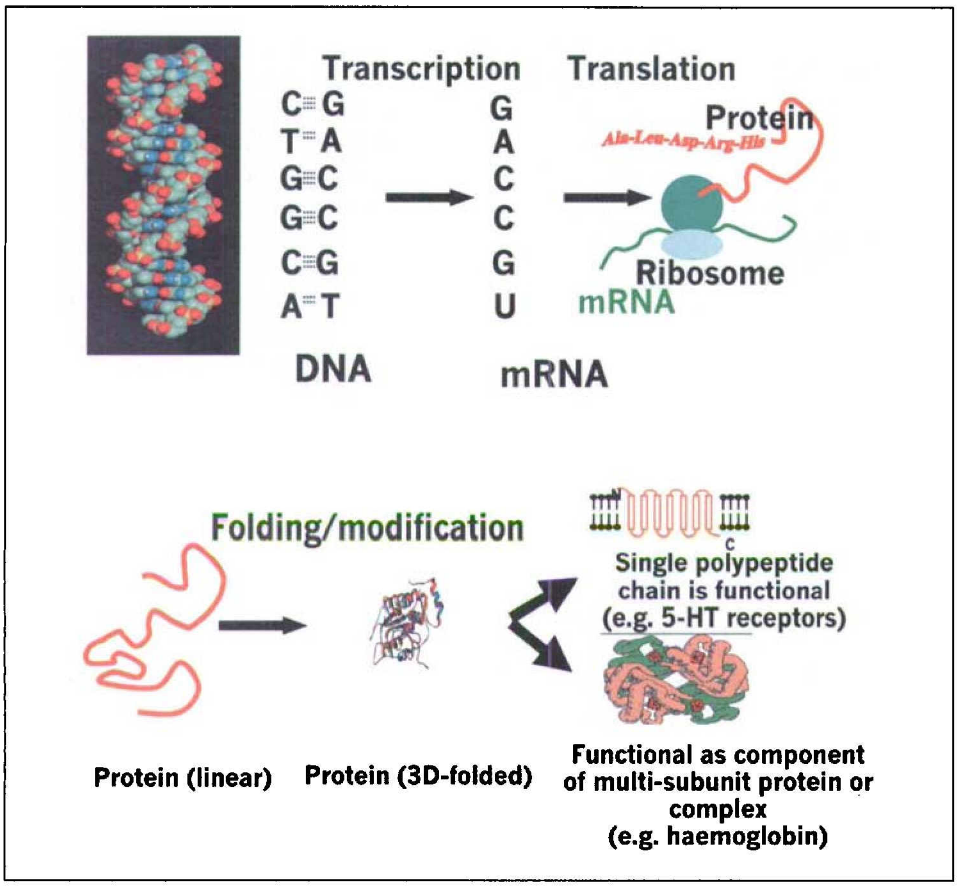
1. Introduction

The sequencing of the genomes of several organisms is proceeding at a rapid pace. Recently the sequencing of the human genome was largely completed, allowing an estimation of the number of encoded genes. The human body is currently thought to contain ~30 000–50 000 genes potentially encoding more than 100 000 different proteins. Temporal and spatial specificity is achieved by regulation of protein expression at several points during DNA transcription, processing to mRNA and subsequent translation into polypeptide chains. And only after folding and post-translational modifications, such as phosphorylation and glycosylation, are functional proteins formed (Scheme). Based on available evidence, the average number of proteins derived from a gene increases with the complexity of the organism, ranging from prokaryotes with one to two proteins/gene to three protein copies/gene for lower eukaryotes like yeast. For higher eukaryotes such as the fruit fly Drosophila and humans at least ten protein copies/gene are expected on average (Table 1). The resulting degree of complexity necessitates a range of highly efficient, sensitive and rapid analytical techniques to provide qualitative and quantitative information. These techniques are the basis of proteomic sciences, which involve the identification, and characterisation of proteins as a complement to genomics. It is realistic to expect that proteome analysis will ultimately lead to the identification of new therapeutic targets and surrogate markers and provide completely new insights in the initiation and progression of disease. There are several experimental phases during differential proteome analysis. First, the proteins expressed by a particular organism, tissue or cell under 'normal' conditions need to be separated and compared to protein expression in the same organism, tissue or cell under different conditions, *e.g.* after drug treatment or in a disease state. The separation and visualisation of these complex protein mixtures is commonly performed using two-dimensional polyacrylamide gel electrophoresis (2D-PAGE) [1], but recent advances in micro-HPLC and capillary electrophoretic methods may develop into alternatives to 2D-PAGE in the future [2]. Image analyses of the 2D-PAGE protein patterns of two or more proteomes then allow the analysis of quantitative changes in protein expression. Mass spectrometry is the method of choice for identifying first the nature of the differentially displayed protein spots as well as for additional characterisation of the identified proteins, for example with respect to differences in posttranslational modification.

2. 2D-PAGE

Two-dimensional gel electrophoresis is at the core of proteome technologies, as it is currently the only method capable of simultaneously separating complex mixtures of thousands of proteins found in biological samples. The first dimension of 2D-PAGE is isoelectric focusing, during which proteins are separated in an immobilised pH gradient (IPG) until they reach the pH of the stationary phase where their net charge is zero, also referred to as the isoelectric point (pI) of the protein. In the second dimension, the proteins are further separated orthogonally by electrophoresis in the presence of sodium dodecyl sulphate (SDS-PAGE) based on their relative molecular mass. Standard 2D gels (20 x 25 cm) covering a pH gradient from 3–10 in their first dimension allow routine separation and

*Correspondence: Dr. J. van Oostrum
Novartis Pharma AG
Functional Genomics Area
WSJ-88.10.01
CH–4002 Basel
Tel.: +41 61 324 73 29
Fax: +41 61 324 49 70
E-Mail: jan.van_oostrum@pharma.novartis.com



Scheme. The path from genomic information to functional proteins.

Table 1. Proteomes in different organisms

		Size of Genome [bp]	Number of Genes	Number of Proteins[a] [x number of genes]
Bacteria	*E. coli*	4.6 x 10^6	4.3 x 10^3	1–1.5 x
Yeast	*S. cerevisiae*	12 x 10^6	5.9 x 10^3	2–3 x
Insects	*Drosophila*	1 x 10^8	1.4 x 10^4	5–10 x
Mammals	Human	3 x 10^9	3 - 5 x 10^4	> 10 x

[a] Best guess based on current estimates.

visualisation of about 2000 proteins. It is expected that this set comprises many 'housekeeping' proteins and fewer of the perhaps more interesting regulatory molecules. One strategy to overcome this problem is to use a larger gel format in which up to 9000 proteins are reportedly detectable [3]. Another way to visualise low-abundance proteins is by using high protein loads in conjunction with multiple (partially overlapping) narrow IPG ranges for the first dimension [4]. IPG's of 1–3 pH units width, referred to as ultrazoom gels, combining high resolution with high sample loading capacity (Fig. 1). This method allows the routine separation of over 15 000 protein spots and the detection of proteins present down to a copy number of a few hundred per cell. (Table 2).

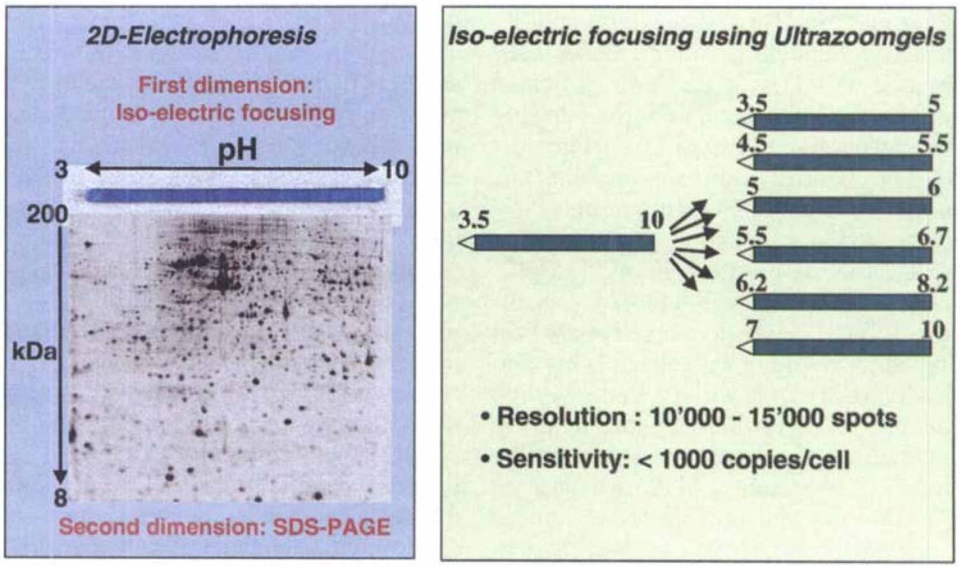

2D-Electrophoresis
First dimension: Iso-electric focusing
pH
3
10
200
kDa
8
Second dimension: SDS-PAGE
Iso-electric focusing using Ultrazoomgels
3.5
10
3.5
5
4.5
5.5
5
6
5.5
6.7
6.2
8.2
7
10
• Resolution : 10'000 - 15'000 spots
• Sensitivity: < 1000 copies/cell

Fig. 1. Principles of 2D-PAGE and IEF zoomgels.

Table 2. What can we see?

1. 3 mg protein, corresponding to 3 x 10^7 cells assuming an average MW of 60000 and a Coomassie detection limit of 15 ng
15 ng = 0.25 pmol protein = 1.5 x 10^{11} molecules
In 3 x 10^7 cells → 5000 molecules/cell
2. Ag / fluorescence detection limit = 1 ng/protein
330 molecules/cell
3. MALDI detection limit = 2 fmol (120 pg)
40 molecules/cell

3. Protein Identification

Visualisation of protein spots on gels is easily achieved by various post-separation staining methods, such as Coomassie Blue, Silver or Sypro Ruby (fluorescence) stains, where the best methods achieve a sensitivity of about 1 ng of protein. Pre-separation radioactive metabolic labelling allows even lower detection limits, but is not practical, for example, in the study of post-mortem human tissue. Protein identification based on the position of a spot on a 2D-PAGE alone is unreliable. Protein identification nowadays relies on mass spectrometric analysis of enzymatic hydrolysates prepared by in-gel digestion of protein spots selected from 2D-PAGE (trypsin is frequently used as one of proteolytic enzymes). The resulting peptide mixture can then be analysed by HPLC/ESI-MSMS [5] or by a frequently applied, two-stage sequential approach using MALDI-MS and ESI-MSMS [6].

For mass spectrometric analyses the individual peptides have to be transferred into the gas phase, ionised, separated by their mass to charge ratio (m/z) and detected with high sensitivity. After calibration with known standards, a typical mass spectrum is a graph of ion intensity as a function of m/z. Apart from high resolution Fourier transform instruments, time-of-flight (TOF) analysers are typically used in conjunction with matrix-assisted laser desorption/ionisation (MALDI). Triple quadrupole, ion trap or quadrupole-TOF mass spectrometers are best suited for coupling with electrospray ionisation (ESI). In MALDI-MS the peptide mixture is co-crystallised with a large excess of a light absorbing matrix and by pulses of laser light (*e.g.* nitrogen laser at 337 nm) the analyte is desorbed, ionised and transferred into the gas phase. In general, protonated peptide ions are formed in high yield and sensitivities in the lower attomole range have been reported for test peptides [7] . The peptide masses are measured with high mass accuracy (typically better than 30 ppm) and this set of masses derived from a protein spot is then screened against the set of expected tryptic masses for each protein or open reading frame in comprehensive protein databases. For medium size proteins (about 40 kDa) only about eight to ten peptide masses are required for unambiguous identification. The complete MALDI approach, including sample preparation, data acquisition, data evaluation and database searching can be automated and is particularly successful for proteins from organisms with a completely sequenced genome [8]. If insufficient peptide masses are detected for unambiguous protein identification, a more specific approach based on electrospray tandem mass spectrometry has to be applied. ESI-MSMS yields actual amino acid sequence information rather than a 'mass fingerprint' of the protein. In this method a solution of the peptide mixture is sprayed from a needle kept at high voltage. The liquid disperses into small, highly charged droplets, which evaporate and explode into several generations of smaller offspring droplets to finally liberate protonated peptide ions. Using a miniaturised version of electrospray, nanoelectrospray allows long term MSMS measurements with the sensitivity required for the identification of proteins derived from 2D gels (femtomole level). From the mixture of these mostly multiply charged peptide ions, one species of interest is separated in a first mass analyser (usually quadrupole filter). In the next step, fragmentation is induced by collisions with nitrogen or argon and the resulting peptide fragments are analysed in the second mass analyser (TOF or quadrupole). In general, fragmentation occurs at several peptidic amide bonds and thus provides at least partial amino acid sequence information. By the application of a tailored database search algorithm (sequence tag approach), it is possible to identify proteins on the basis of one or two peptides only. Based on this highly selective approach almost all human proteins can be identified via their corresponding expressed sequence tag (EST) database entries [9]. The methods for interfacing these high-sensitivity 'downstream' analytical techniques to 2D-PAGE have matured and the complete approach can now be reliably applied in a largely automated fashion at the femtomole-level of gel-isolated proteins. Protein identification is only the first step in protein characterisation, and detailed analysis has to consider post-translational modifications as a major factor influencing protein structure and function.

4. Identification *versus* Characterisation

The characterisation of posttranslationally modified proteins and protein isoforms, which have been separated by 2D-PAGE, is as important as protein identification. Especially important is the detection of the phosphorylation status, which is frequently essential for the modulation of protein function. Whereas only a small number of peptides are sufficient to uniquely identify the protein(s) contained in a 2D-PAGE spot, the characterisation process ideally requires the analysis of all peptides derived from the protein. For example, in a differential proteome study of Taxol-treated *versus* untreated human 697-cells, several distinct protein spots within the pI range 5–6 and apparent molecular mass range 18–23 kDa were identified as stathmin with protein spots A-F being up-regulated (Fig. 2A). Differences in post-translational modifications could have been the reason for the different electrophoretic mobilities and phosphorylation was an obvious possible explanation, since stathmin contains four known phosphorylation sites at Ser15, -24, -37 and -62. To separate potential phosphorylation sites and to achieve optimal sequence coverage a double enzymatic in-gel digestion strategy with trypsin and parallel Glu-C treatment was employed (Fig. 2B). Detection of metastable signals for loss of phosphoric acid in MALDI-MS together with nanoelectrospray partial sequencing allowed the determination of the prevailing phosphorylation status of all seven stathmin isoforms [10]. Spot G corresponds to the unphosphorylated species, spot F mainly to Ser37 monophosphate and spot E to Ser24, -37 diphosphate. Spots D and C are isomeric triphosphates (Ser24, -37, -62 and Ser15, -24, -37) and the remaining spots A and B are phosphorylated at all four sites (Ser15, -24, -37 and -62). They are probably derived from the two previously reported isoforms of stathmin, differing by a yet unknown modification.

5. Interaction Proteomics

Protein identification and characterisation, accomplished by the methods described, sometimes provides first conclusions about biological function. However, since only a fraction of the potential gene products are known proteins, an understanding of the function of the unknown proteins encoded by the genome, or proteins whose function is elusive, remains a challenge. One route to increase understanding of the important aspects of their biological role is to identify interacting partners and thus to establish the position of the individual proteins within the network of cellular pathways. Ideally, interaction analysis combines the identification of interacting proteins as well as the characterisation of interacting sites. Both of these questions can be addressed simultaneously by the analysis of covalently cross-linked complexes by the MS technology developed for proteome analysis [11]. Fig. 3 illustrates the approach of chemical cross-linking with the identification of tubulin as a binding partner of stathmin in a cross-linked complex. A cellular extract was treated with EDC (1-ethyl-3-(3-dimethylaminopropyl) carbodiimide) and subjected to 2D-PAGE separation. Stathmin and stathmin-containing complexes were subsequently visualised after 2D-Western blotting and probing with anti-stathmin antibodies. Besides the predominantly non-phosphorylated and mono-phosphorylated isoforms of stathmin, two protein spots in close proximity at higher mass were detected and identified as complexes between stathmin and alpha-tubulin.

6. From Characterisation to Function

Since the major isoforms differ mainly in the extent of phosphorylation, most of the information obtained by 2D-PAGE is retained in a simpler separation system using native 1D electrophoresis or 1D isoelectric focusing, followed by Western blotting using anti-stathmin antibody

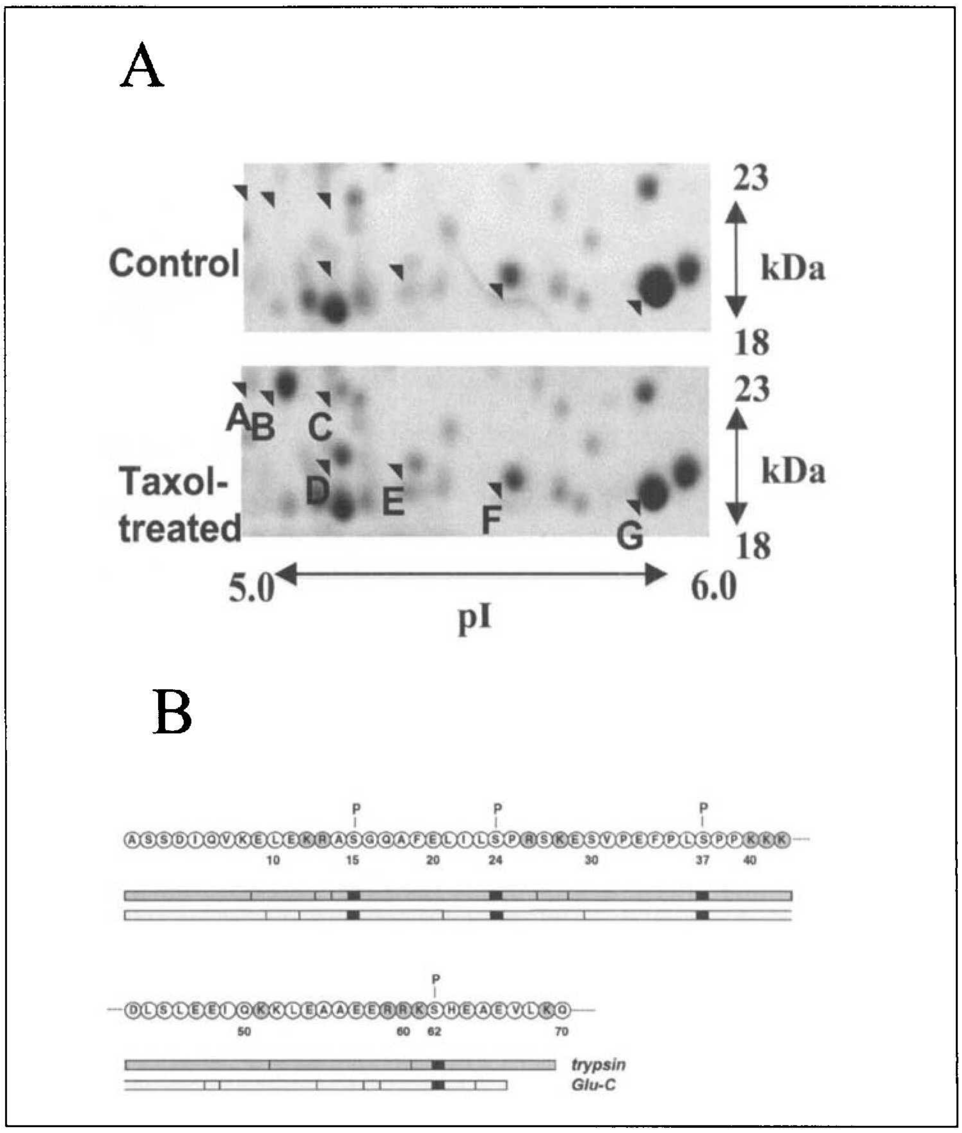


Fig. 2. A) Detection of differentially expressed stathmin isoforms by 2D-PAGE. B) Stathmin sequence with enzymatic cleavage sites.

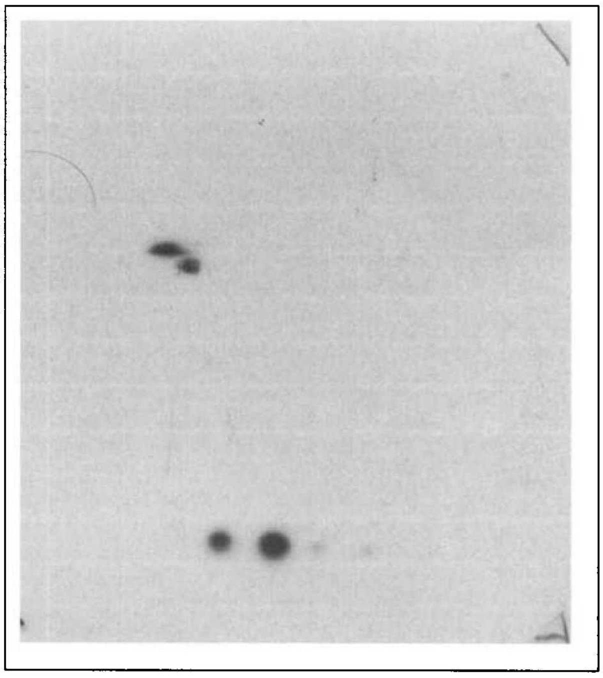


Fig. 3. 2D-PAGE of a 697 pre-B lymphomas cell extract cross-linked with EDC, after immunoblotting and probed with stathmin antibodies, reveals the expected 'free' stathmin isoforms and stathmin in high molecular weight complexes cross-linked to tubulin.

detection. These systems allow, for example, different isoforms to be correlated with the action of Taxol in time-course experiments (Fig. 4). In this way good correlation between G2/M cell cycle arrest and the appearance of tri- and tetraphosphorylated stathmin species after Taxol treatment of SW-2 cells was observed. In these species, positions 15 and/or 62 are additionally phosphorylated. Interestingly, phosphorylated isoforms are proposed to be involved in the regulation of microtubule dynamics. To detect whether Ser15 and Ser62 take part in the formation of a potential tubulin/stathmin complex, we prepared several specific stathmin constructs and tested these for tubulin sequestering activity (Fig. 5). Removal of the 39 N-terminal and the 9 C-terminal amino acids had no influence, while residues 40-139 were essential for sequestering tubulin. The individual regions 40-109 and 75-148 are practically inactive, but a mixture of both reconstituted part of the original activity. Ser62 is therefore the only phosphorylation site involved in direct binding to tubulin. For further localisation of the interacting area of stathmin and tubulin, the synthetic peptide 54-72 was added in a 100-fold excess to the complex in a competitive inhibition experiment. Stathmin activity was completely suppressed- and tubulin depolymerization was as slow as in its absence. The region around Ser62 is therefore essential for stathmin-tubulin binding and represents a potential target for drug candidates.

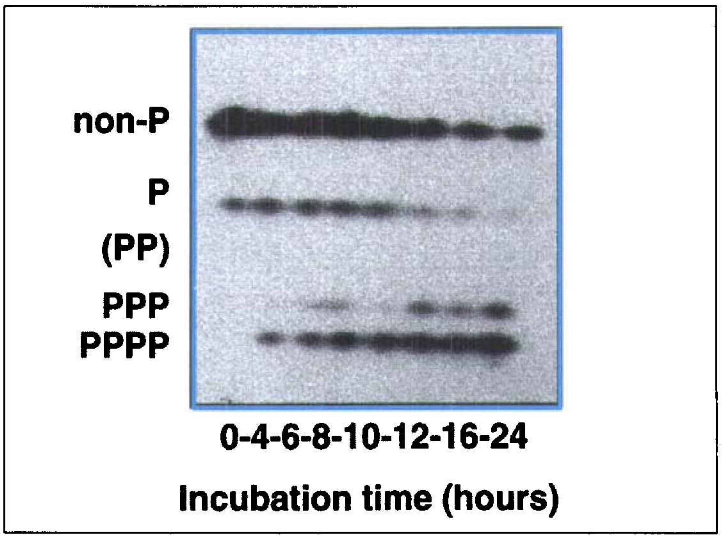


Fig. 4. Detection by IEF-Immunoblotting of phospho-isoforms of stathmin in paclitaxel-treated (100 nM) 697 pre-B lymphoma cells. Proteins were separated by IEF on IPG sheets (pH 4-7) for parallel analysis, IEF-Immuno-blotted and probed with stathmin antibodies (1:3000).

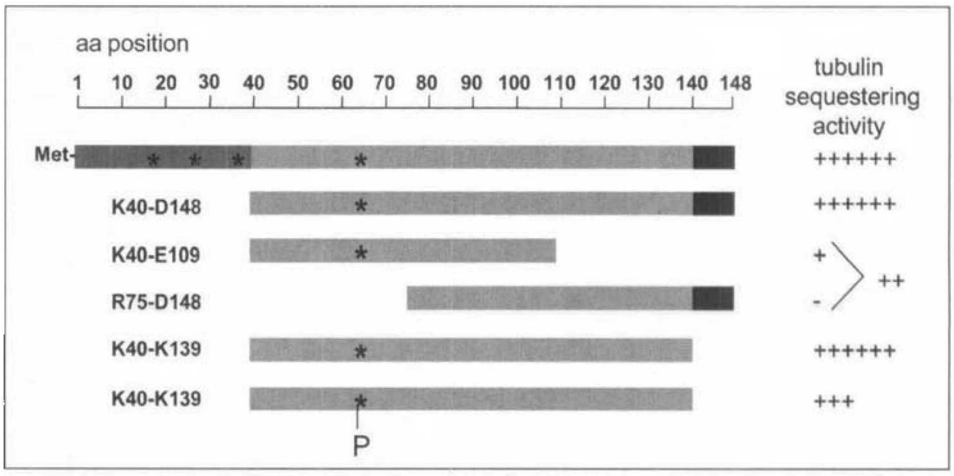


Fig. 5. Stathmin constructs and corresponding tubulin sequestering activity. Stars mark the phosphorylation sites, 15, 24, 37 and 62. First line corresponds to wild-type plus N-terminal methionine.

7. Conclusions

The analysis of posttranslationally modified peptides using MALDI-MS and nanoelectrospray MSMS is feasible even in complex mixtures. However, use of a double enzymatic treatment, such as trypsin and Glu-C in parallel, is essential to separate potential phosphorylation sites of the various isoforms of the protein. On the basis of phosphopeptide characterisation, subsequent experiments established a relationship between Taxol action and reduced stathmin-tubulin complex formation by phosphorylation of Ser62. The binding site was mapped to a region surrounding Ser62 by selected constructs and by competitive action of peptide 54-72. Interaction proteomics, *e.g.* the analysis of binding/interaction partners, may develop as a powerful follow-up technique after differential display proteomics to establish the position of the individual proteins within the network of cellular pathways and to provide a more mechanistic insight into the functionality of the protein complexes. In summary, this study demonstrates that the step from differential display proteomics to functional analysis of potential drug targets is quite feasible.

Received: February 9, 2001

[1] A. Goerg, W. Weiss, in 'Proteome Research: Two-dimensional Gel Electrophoresis and Identification Methods', Ed. T. Rabilloud, Springer Verlag Berlin, Heidelberg, New York, **2000**, p. 57.

[2] S.P. Gygi, B. Rist, S.A. Gerber, F. Turecek, M.H. Gelb, R. Aebersold, *Nature Biotech.* **1999**, *17*, 994–999.

[3] J. Klose, *Methods Mol Biol.*, **1999**, *112*, 147.

[4] S. Hoving, H. Voshol and J. van Oostrum, *Electrophoresis* **2000**, *21*, 2617.

[5] A. Ducret, I. Vanoostveen, J.K. Eng, J.R. Yates, R. Aebersold, *Protein Sci.* **1998**, *7*, 706.

[6] A. Shevchenko, O. Jensen, A. Podtelejnikov, F. Sagliocco, M. Wilm, O. Vorm, P. Mortensen, A. Shevchenko, H. Boucherie, M. Mann *Proc. Natl. Acad. Sci. USA* **1996**, *93*, 1444

[7] J. Gobom, E. Nordhoff, E. Mirgorodskaya, R. Ekman, P. Roepstorff, *J. Mass Spectrom.* **1999**, *34*, 105.

[8] H.W. Lahm, H. Langen, *Electrophoresis* **2000**, *21*, 2105.

[9] M. Mann, *TIBS* **1996**, *21*, 494.

[10] D.R. Müller, P. Schindler, M. Coulot, H. Voshol, J. van Oostrum, *J. Mass Spectrom.* **1999**, *34*, 336.

[11] D. Müller, P. Schindler, H. Towbin, U. Wirth, H. Voshol, S. Hoving, M.O. Steinmetz, *Anal. Chem* **2000**, in press.