Hogeschool van Arnhem en Nijmegen

Techniek

Applied Sciences

Bio-Informatica

Course 5b: Proteomics

Reader 539

Hoofdfase

Voltijd

HAN www.han.nl

Proteomics

Introduction

Although the number of genes can be extensive for higher eukaryotes (> 20.000 genes), the number of proteins for any higher organism is at least an order of magnitude larger than the number of genes. There are several explanations for the high amount of proteins. First, different protein products can be generated from one coding DNA sequence by alternative splicing of transcripts prior to translation. Secondly, and even more importantly, most proteins are modified after translation.
Lower eukaryotes and prokaryotes do not have such a high number of genes but after translation, like in higher eukaryotes, proteins can be modified (although not so extensively).

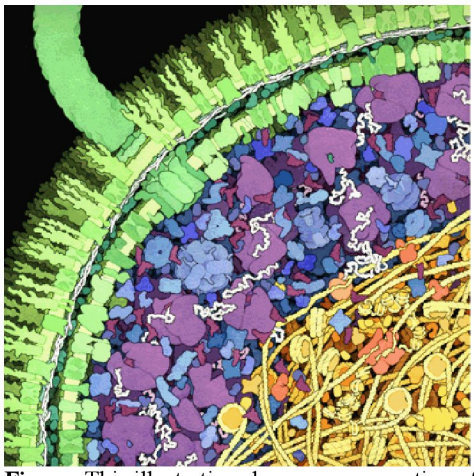


Figure: This illustration shows a cross-section of a small portion of an *Escherichia coli* cell. The cell wall, with two concentric membranes studded with transmembrane proteins, is shown in green. A large flagellar motor crosses the entire wall, turning the flagellum that extends upwards from the surface. The cytoplasmic area is coloured blue and purple. The large purple molecules are ribosomes and the small, L-shaped maroon molecules are tRNA, and the white strands are mRNA. Enzymes are shown in blue. The nucleoid region is shown in yellow and orange, with the long DNA circle shown in yellow, wrapped around HU protein (bacterial nucleosomes). In the center of the nucleoid region shown here, you might find a replication fork, with DNA polymerase (in red-orange) replicating new DNA. © *David S. Goodsell 1999*. Picture taken from: http://mgl.scripps.edu/people/goodsell/illustration/public/ecoli-icon.gif

Distinct versions of any given protein translation product can be generated by modifications of the polypeptide chain. More than 200 amino acid residue modification have been observed in proteins and most, but not all, are due to enzyme-mediated reactions. Prominent examples include glycosylation, ADP-ribosylation, phosphorylation or isoprenylation and acetylation, but also directed (usually limited) proteolysis and linkage with other proteins, such as ubiquitin. Many individual polypeptides do not acquire biological activity until they have undergone specific, and often multiple, non-covalent associations. They may require interactions with cofactors, other polypeptides, nucleic acids or membranes, or be transported to particular locations within cells or tissues.

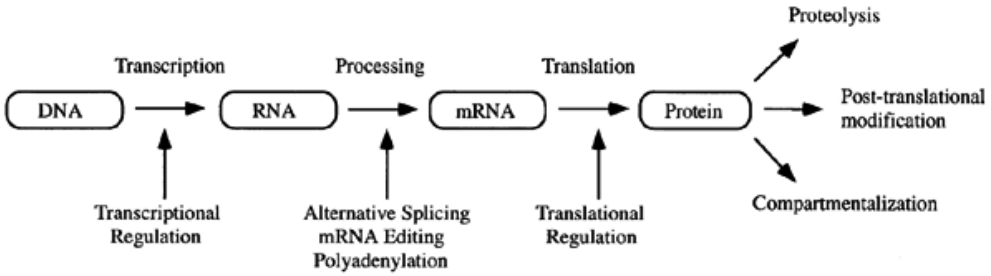


Figure: Mechanisms by which a single gene can give rise to multiple gene products. Multiple protein isoforms can be generated by RNA processing when RNA is alternatively spliced or edited to form mature mRNA. mRNA, in turn, can be regulated by stability and efficiency of translation. Proteins can be regulated by additional mechanisms, including posttranslational modification, proteolysis, or compartmentalisation. *Picture taken from: http://mmbr.asm.org/content/66/1/39/F2.small.gif*

Proteomics is the systematic study of the many and diverse properties of a set of proteins. Proteomics tries to answer analytical questions about the change in concentration and distribution of proteins in the organism, the expression profiles of proteins in different tissues and the identification and localisation of proteins. These questions are closely connected with more functional questions, which aim to elucidate interactions between different proteins, or between proteins and other molecules. Moreover, the increasing knowledge of 3-dimensional structures of proteins at an atomic level combined with the results from proteomic studies can elucidate the functional role of proteins.

Proteomics-research is complex

- Protein-concentrations vary over a large dynamic range (up to a million times). This makes it difficult to analyse only those proteins that you are interested in, as they are most often present in a low concentration. This means that sampling is very important (obtaining a subset of proteins you are interested in).
- Proteins undergo post-translational modifications.
- Protein concentration is determined by the rate of turnover, synthesis and breakdown.
- Protein activity may depend on cellular location, on co-factors or complex formation with other proteins or molecules.
- Sample material, from which to isolate proteins, is often limited and variable.
- Proteins cannot be multiplied like you can multiply DNA or RNA (using PCR).
- Protein-samples easily degrade (proteases are everywhere).
- Proteomes are not static; there is developmental and temporal specificity and influence from disease and drug perturbations.

Sub-areas of proteomics

1. Protein characterisation and genomic annotation
   - sub-cellular localisation
   - post-translational modifications
   - protein function
2. Expressional studies (protein quantification)
3. Global protein-protein interactions
4. High-throughput structural assessment

Proteomics promises to address al these questions, but it is not easy given the enormous variety of proteins present in higher organisms. However, the ability to analyse proteins on a cell-wide basis is increasing with the technological advances in mass spectrometry, multidimensional liquid chromatography, protein chips, and gel based separation methods.

Computation plays a critical role in proteomic analysis, having the responsibility of for example linking complex mass spectral data to gene and protein sequence data in a meaningful way. Though useful software tools exist, the informatics lags behind the rapidly advancing chemical and analytical methods in proteomics. Nowadays, significant advancements must occur in the handling, storage, and analysis the copious, complex data that is starting to be generated by efforts to study proteomics.

What is the "function" of a protein?

A major aspect of proteomics is the identification and characterisation of proteins. These proteins can be those that are involved in certain processes or proteins that are present in some tissue at a certain time point. One of the things researchers want to know of these proteins is their function. Each protein is a gene product that interacts with the cellular environment in some way to promote the cell's growth and function. Thus, in principle, function is defined as the role of a protein in a cell. However, protein function means different things to different people, because the concept of function can be considered from different perspectives. Function can be described in biophysical terms (e.g. kinase), biochemical terms (being a part of a certain pathway), biological terms (e.g. "development"), pathological terms (tumour suppressor), etc. Therefore, before predicting "function", it should be clarified what it is that is attempted to predict.

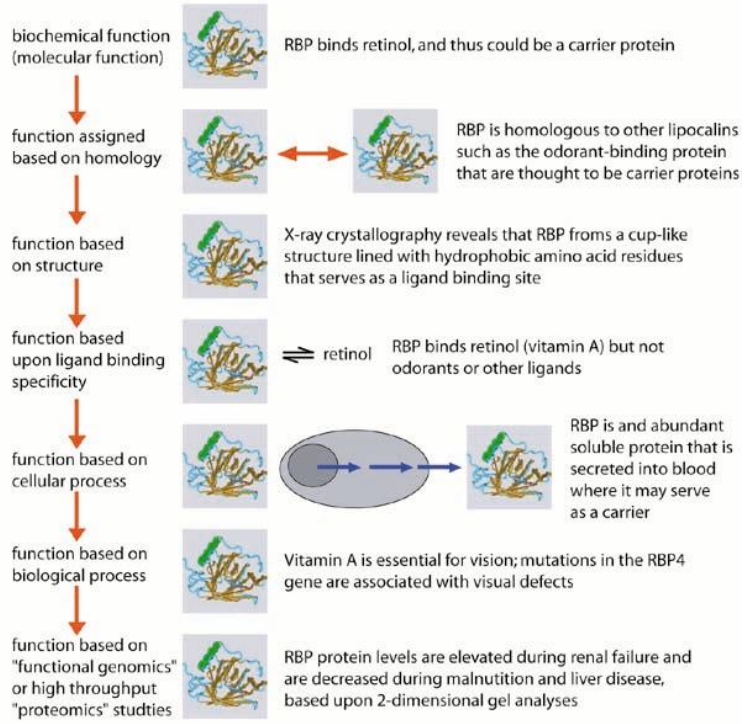


Figure: Protein function may be analysed from several perspectives. Retinol-binding protein (RBP) is used as an example. Retinol is vitamin A. When you only look at the function based on the biological process you would think that RBP is essential for vision. In addition, when you look at the function based on proteomics you would find that RBP levels are elevated in renal failure and RBP levels are decreased in liver disease. From this you can conclude that RBP is not only essential for vision, but that something is also going on in the liver. Of course RBP is important for vision, but vision is not the function of RBP. But what is the function of RBP? RBP binds retinal and therefore could be a carrier. RBP is homologous to lipocalins like the odorant binding protein (OBP). OBP binds a variety of odorants in nasal mucus, suggesting that the binding properties of the protein are central to its function. RBP only binds retinol. However, the biological function of a protein is not known from its ligand binding properties alone. RBP is secreted into the bloodstream. Therefore the actual function of RBP probably has something to do with binding vitamin A and transporting it to the eyes. *Figure adapted from Bioinformatics and Functional Genomics, Pevsner.*

-A protein has a biochemical function synonymous with its molecular function. For an enzyme, the biochemical function is to catalyse the conversion of one or more substrates to product(s). For a structural protein such as actin or tubulin, the biochemical function is to influence the shape of a cell. For a transport protein, the biochemical function is to carry a ligand from one location to another. In principle each protein has a biochemical function.
-Functional assignment is often made based upon homology. If a hypothetical protein is homologous to an enzyme, it is often provisionally assigned that enzymatic function. This type of functional assignment is best viewed as a hypothesis that must be tested experimentally.
-Function may be assigned based upon structure. If a protein has a three-dimensional fold that is related to that of a protein with a known function, this may be the basis for functional assignment. Note, however that structural similarity does not necessarily imply homology, and homology does not necessarily imply functional equivalence.
-All proteins function in the context of other proteins and molecules. Thus a definition of a protein's function may include its ligand (if the protein is a receptor) its substrate (if the protein is an enzyme), its lipid partner (if the protein interacts with membrane) or any other molecule with which it interacts.
-Many proteins function as part of a distinct biochemical pathway such as the citric acid cycle.
-Proteins function as part of some broad cell biological process. All cellular processes require proteins in order to function, and each individual protein can be defined in the context of the broad cellular function it serves.

Controlled vocabularies such as Gene Ontology (GO) can convey a rigorous framework for discussing function.

Gene ontology

Biologists require the ability to use biological information from a variety of sources, and to be able to integrate this information in order to make biologically meaningful discoveries. The rapidly increasing amounts of many types of biological data, stored in numerous and diverse biological databases using different accessions and annotations with inconsistent terminology (like different names for the same gene product), have made it difficult for the average biologist to identify and easily query biological data.
Recent years have seen a growing trend towards the adoption of ontologies for the management of biological knowledge. Ontologies represent a powerful means to analyse and integrate biological data. They provide a formal specification of how to represent the objects, concepts and other entities that are assumed to exist in some area of interest, and the relationships that hold among them. Various ontologies and controlled vocabularies have grown out of the effort to provide a shared language for communicating biological information. Biological ontologies aim to overcome the semantic heterogeneity (different words or descriptions for the same molecular function, biological process or cellular component, such as apoptosis and cell death) commonly encountered in molecular biology databases.

The Gene Ontology (GO, http://www.geneontology.org/) is by far the most widely used bio-ontology. It aims to formalise our knowledge about biological processes, molecular functions and cell components in three (mutually independent) hierarchies. The vocabularies of these hierarchies are structured in a classification that supports 'is-a' and 'part-of' relationships.

- Molecular function describes tasks performed by individual gene products. In other words, it describes biochemical activities (including specific binding to ligands or structures) performed by individual gene products, or assembled complexes of gene products, at the molecular level. It only describes what is done, it does not specify where

or when, or in what context the action takes place. Examples of broad functional terms are ‘catalytic activity’, ‘transporter activity’ or ‘binding’. Examples of narrower functional terms are ‘6-phosphofructokinase activity’, ‘Na^+-K^+ antiporter activity’ or ‘EGF receptor binding’

- Biological process refers to the broad biological goals that a gene product or an assembled complex of gene products is associated with. High-level processes such as ‘cell death’ can have subtypes, such as ‘apoptosis’, but also subprocesses, such as ‘apoptotic chromosome condensation’.
- Cellular compartment describes locations at the levels of subcellular structures and macromolecular complexes. It is arranged in such a way that specific terms are part of a large group or compartment that can be found inside a cell. Examples of cellular components include ‘nuclear inner membrane’ or ‘mitochondrion’, but also ‘ubiquitin ligase complex’.

Biological processes, molecular function and cellular component are all attributes of genes, gene products or gene-product groups. Terms from these three disciplines may be assigned independently to genes, gene products or gene-product groups as they represent independent attributes and are therefore clarifying in many situations. The relationships between a gene product (or gene-product group) to biological process, molecular function and cellular component are one-to-many. This reflects the biological reality that a particular protein may function in several processes, contain domains that carry out diverse molecular functions and participate in multiple alternative interactions with other proteins, organelles or locations in the cell. Thus any protein may participate in more than one molecular function, biological process, and/or cellular component.

Genes and gene products are assigned to GO categories through a process of annotation. The author of each GO annotation supplies an evidence code that indicates the basis for that annotation. For example IGI stands for inferred from genetic interaction (experiments in which one gene provides information about the function, process, or component of another gene). Another example is IEA, which stands for inferred from electronic annotation (annotations based on ‘hits’ in searches such as BLAST, but without confirmation by a curator) in contrast to ISS, which stands for inferred from sequence or structural similarity (BLAST results that are reviewed for accuracy by a curator)

Ontology terms do not represent an individual item, but the associated set – that is, not a particular protein kinase from yeast but all protein kinases. As well as being described by their relationships, terms in an ontology contain an unique identifier (ID), such as GO:0016301. These ontology IDs have two components: a letter code that specifies the ontology type and a number. Thus GO:0016301 represents kinase activity in the Gene Ontology. These IDs can be used in two ways: to link a biological database to ontologies and to connect different biological databases.

The ontology dataset can be provided as a flat file, XML or MySQL database file. The ontology structure can be represented as a graph. The Gene Ontology structure can be represented graphically as a directed acyclic graph (DAG). DAGs differ from hierarchies in that a ‘child’ (more specialised term) can have many ‘parents’ (less specialised terms). The parent is a more general term then the child and as the depth on the three increases both parent and child terms get more specialised. In addition, a child term may be an instance of its parent term, in which case the graph is labelled ‘is-a’, or the child term may be a component of the parent term (a ‘part-of’ relationship).

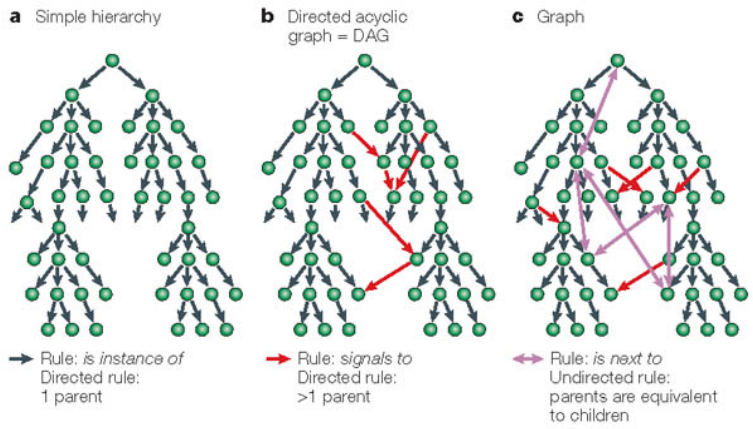


Figure: Different possibilities of graphical representations of ontologies. A) simple hierarchy: every child has only 1 parent. B) Directed acyclic graph: a child can have more parents. C) Graph: A child can have more than one parent and parents may be equivalent to children. Picture taken from: http://www.nature.com/nrg/journal/v5/n3/images/nrg1295-i1.jpg

For example, the biological process 'DNA ligation' has two parents, DNA repair and DNA recombination. This is because ligation is needed to put pieces of DNA together again after DNA repair as well as after DNA recombination. The DAG graph representation is called directed because DNA-ligation cannot be a parent of DNA-repair or DNA-recombination.

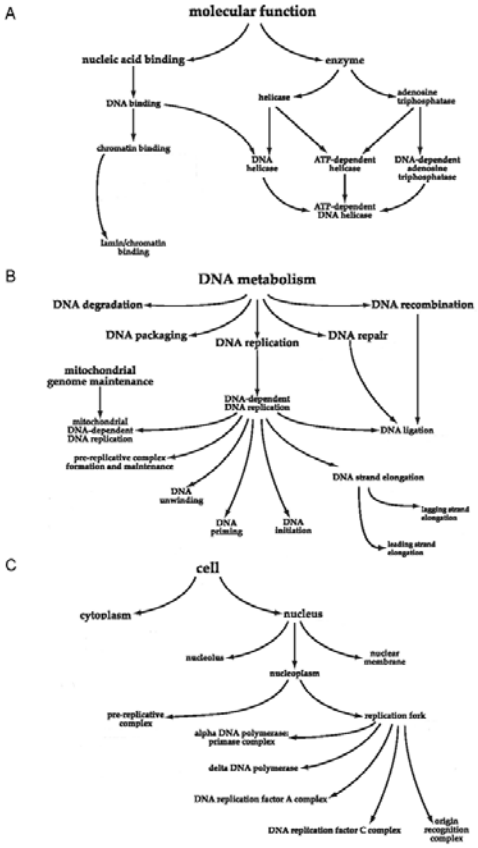


Figure: Examples of Gene Ontology represented as directed graphs. A) Molecular function. B) Biological process. C) Cellular component. Picture taken from: http://www.nature.com/ng/journal/v25/n1/abs/ng0500_25.html

The Gene Ontology is being developed by the Gene Ontology Consortium. The Consortium was begun by scientists associated with three model organism databases: the *Saccharomyces* Genome

Database (SGD), the *Drosophila* genome database (Flybase) and the Mouse Genome Informatics Databases (MGD/GXD). Subsequently databases associated with many other organisms have joined the GO Consortium. The GO database relies on these databases in which each gene or gene product is annotated with GO terms. Thus the GO represents an ongoing cooperative effort to unify the way genes and gene products are described. This means that the ontologies within the Gene Ontology are by no means complete, there are being updated and expanded all the time.

The Gene Ontology Consortium project has three major goals:

1. to develop a set of controlled, structured vocabularies – known as ontologies- to describe key domains of molecular biology, including gene product attributes and biological sequences
2. to apply GO terms in the annotation of sequences, genes or gene products in biological databases
3. to provide a centralised public resource allowing universal access to the ontologies, annotation data sets and software tools developed for use with GO data

The key to the general use of ontologies will be access to the data in biological databases that are annotated with knowledge in these ontologies. Many biological databases are now incorporating ontology IDs and using them to annotate data objects. The more that this is done, the more useful these resources will be for the community.

2D-gel electrophoresis

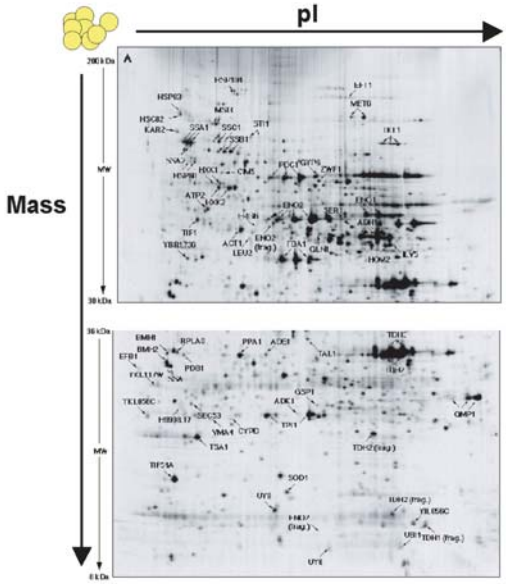


Figure: Saccharomyces cerevisiae SWISS-2D PAGE maps, high molecular weight (top) and low molecular weight (bottom). Picture taken from: http://world-2dpage.expasy.org/swiss-2dpage/images/publi/yeast-high.gif
http://world-2dpage.expasy.org/swiss-2dpage/images/publi/yeast-low.gif

The classical method of gene expression studies in terms of proteomics is two-dimensional gel electrophoresis (2DE), which is often combined with subsequent protein identification by mass spectrometry (MS). This approach allows the generation of catalogues of expressed proteins in a cell or tissue of interest. Nevertheless, the identification of proteins using this approach is limited among others by the sensitivity of the protein complement to extract preparation, running conditions and gel composition. Furthermore, the proteins separated by 2DE are obtained in denatured form and in limited amounts; their further functional characterization is not possible.

For functional characterization, for example by biochemical experiments, the expression of the protein of interest in recombinant form is usually required.

Proteins possess a charge and thus migrate when introduced into an electric field. Proteins are denatured and electrophoresed through a matrix of acrylamide that is inert (so it does not interact with the protein) and porous (so that the proteins can move through it). The velocity of a protein as it migrates though an acrylamide gel is inversely proportional to its size and thus a complex mixture of proteins can be separated in a single experiment. Proteins are almost always electrophoresed through acrylamide under denaturing conditions in the presence of the detergent sodium dodecyl sulfate (SDS), so this technique is commonly abbreviated SDS-PAGE.
By combining SDS-PAGE with an initial separation of the proteins based on their charge, the proteins in a protein mixture can be much more separated. Separation of proteins based on their charge can be done by isoelectric focussing. A gel matrix (or strip) is produced that contains ampholytes spanning a continuous range of pH values, usually between pH 3 and pH 11. Each protein is zwitterionic, and when electrophoresed it migrates to the position at which the total net charge is zero. This is the isoelectric point (abbreviated pI) at which the protein stops migrating. A complex mixture of proteins may thus be separated based upon charge.
Thus the first dimension of 2DE is based on isoelectric focusing, by which the proteins are separated according to their pI in pH gradient polyacrylamide gels (first dimension). In the second dimension proteins are separated according to their molecular weights by SDS-PAGE. Visualisation of the separated proteins is achieved by different staining techniques. Colour density and size of the detected spots enable protein quantification.
Normalisation of the multiple gels in an experiment can be done by using the spot intensity of one or several housekeeping proteins like actin or spiking controls. Of course global normalisation techniques using the total of all valid spots or the total pixel density of the gel image can also be used. T-tests can then be applied to check the validity of the differences measured between the different samples.

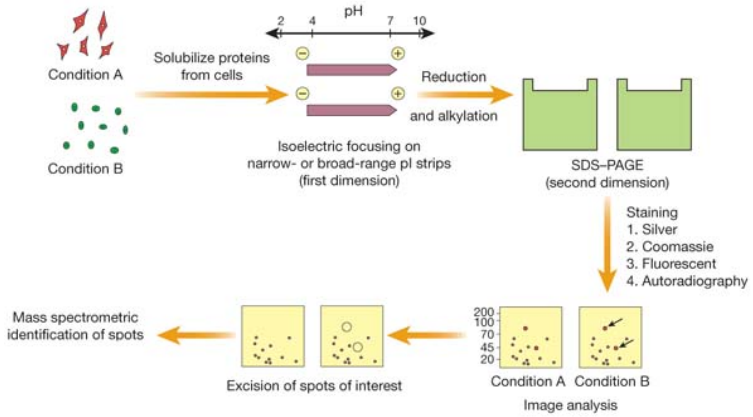


Figure: Scheme showing the 2DE approach. Cells (or tissue) derived from two different conditions, A and B, are harvested and the proteins solubilised. The crude protein mixture is then applied to a first dimension gel strip that separates the proteins based on their isoelectric points. After this step, the strip is subjected to reduction and alkylation and applied to a 'second dimension' SDS–PAGE gel where proteins are denatured and separated on the basis of size. The gels are then fixed and the proteins visualized by silver staining. Silver staining is less quantitative than Coomassie blue but more sensitive and is also compatible with mass spectrometric analysis. After staining, the resulting protein spots are recorded and quantified. Image analysis requires sophisticated software and remains one of the most labour-intensive parts of the two-dimensional gel approach. The spots of interest are then excised and subjected to mass spectrometric analysis. Picture taken from: http://cmbi.bjmu.edu.cn/cmbidata/proteome/method/2DE.gif

With 2DE thousands of protein spots can be visualised, resulting in a global view of the state of the proteome. By comparing the 2D spot patterns from different samples, changes in individual

proteins can be detected and quantified. This allows identification of protein markers that are characteristic of a specific physiological or pathological state of a cell or tissue. In addition to changes in protein levels, changes in post-translational processing, such as phosphorylation can also be detected and studied.
Identification of the large numbers of proteins separated by 2DE is most commonly achieved by automated matrix-assisted laser desorption/ionisation time-of-flight mass spectrometric (MALDI TOF-MS) peptide mapping followed by extensive database searches In cases, where more structural information is required from the separated proteins, nano-liquid chromatography (LC)-electrospray ionisation (ESI)-MS/MS is often employed.

However, the approach is far from perfect.

- It is time consuming and laborious, because each spot has to be extracted, digested and analysed individually. Automation is difficult.
- Large amount of protein needed (compared to other proteomics techniques)
- 2DE is not able to resolve and depict in a single gel all of the proteins though to be present in a (mammalian) cell. Improvements:
    - larger gels
    - use of narrower pH gradients to zoom in
    - use of higher and lower percentages SDS-PAGE gels
    - use of radio-labelling or fluorescence labelling to identify low abundance proteins.
- Dynamic range of proteins is very large (1 to 10^5 or 10^6). So highly abundant proteins will overshadow low abundant proteins. For example, 50% of the protein content of yeast is the product of 100 genes.
    - Enrichment or pre-fractionation strategies necessary to reach the less abundant proteins.
- For finding hydrophobic proteins or membrane proteins fractionation is required, as they are not easily extracted using normal protein extraction procedures
- There is a high degree of gel-to-gel variation in spot patterns, making it difficult to distinguish any true biological variation from experimental variation. Improvements:
    - Use of protein purification kits and pre-cast gels.
    - In-gel standards for normalisation / DIGE

2D difference gel electrophoresis (DIGE)

Samples are labelled prior to electrophoresis with fluorescent dyes (Cy2, Cy3 or Cy5). Samples are then mixed prior to IEF and resolved on the same 2D gel. Due to the different fluorescent colours it is possible to compare to samples on the same gel directly. This will result in a large reduction in technical variation between samples. For example protein loss during loading of the mixed sample on the pH strip will be the same for both samples within a single DIGE gel. Thus the relative amount of proteins between the two samples will be unchanged. Furthermore, the number of gels run in an experiment is reduced. Instead of an extra experimental sample, each sample in the experiment can also be compared with a reference sample containing a mixture of all samples in the experiment. This pooled standard can then be used to normalise protein abundance measurements across multiple gels in an experiment. As a consequence, each gel will contain an image with a highly similar spot pattern, simplifying and improving the confidence of inter-gel spot matching and quantification.
Normalisation is based on the assumption that the majority of all spots in the experiment have not changed from one experiment to the other. When this is true, a histogram representation of the $\log^{10}$ of all spot volume ratios will resemble a normal distribution. The individual volume ratios are then normalised by aligning the log volume ratio histogram around zero.

Mass spectrometry

Mass spectrometry is a powerful analytical technique that is used to identify unknown compounds, to quantify known compounds, and to elucidate the structure and chemical properties of molecules. Detection of compounds can be accomplished with very minute quantities (as little as 10^{-12}g, 10^{-15} moles (femtomoles) for a compound of mass 1000 Daltons). This means that compounds can be identified at very low concentrations (one part in 10^{12}) in chemically complex mixtures.

Mass spectrometers use the difference in mass-to-charge ratio (m/z) of ionised atoms or molecules to separate them from each other. Mass spectrometry is therefore useful for identification of molecules based on the chemical and structural information the MS instrument provides.

A mass spectrometer is an instrument that measures the masses of individual molecules that have been converted into ions, i.e., molecules have to be (electrically) charged. The charge can be positive (one H^+ added to the molecule) or negative (one H^+ abstracted from the molecule). Since molecules are very small, it is not convenient to measure their masses in grams. In fact, the mass of a single hydrogen atom is approximately 1.66 X 10^{-24} grams. A more convenient unit for the mass of individual molecules is the Dalton (Da for short). It is defined as follows: 1 Da=(1/12) of the mass of a single atom of the isotope of carbon-12(^{12}C). This follows the accepted convention of defining the ^{12}C isotope as having exactly 12 mass units.
A mass spectrometer does not actually measure the molecular mass directly, but rather the mass-to-charge ratio of the ions formed from the molecules. The unit used for charge is the magnitude of the charge on an electron. The charge on an ion is then denoted by the integer number z of the fundamental units of charge. The mass-to-charge ratio m/z therefore represents Daltons per fundamental unit of charge.

MS spectra:

A mass spectrum is often depicted as a simple histogram. For example, the Figure shows a mass spectrum of the simple molecule carbon dioxide, CO_2. In this example, all the ions are positively charged. (It is possible to generate and detect negative ions as well.) The ionised CO_2 molecule (or molecular ion) appears at m/z 44. The ion is singly charged thus the m/z ratio is 44/1. The ion mass of 44 Da can be calculated from C=12 and O=16. It is possible to break up or to fragment (some of) the CO_2 molecules in the ionisation procedure. This fragmentation is dependant on the amount of energy used for ionisation. When fragmentation occurs some ions appear in the spectrum at m/z values less than the m/z value that corresponds to the molecular mass of CO_2. Cleavage of a carbon-oxygen bond in the molecular ion to produce ionised carbon monoxide or ionised atomic oxygen result in the fragment ions at m/z 28 and 16; loss of two neutral oxygen atoms results in an additional fragment at m/z 12 for carbon. The molecular ion is designated as M^+ or CO_2^+ and the fragment ions are designated as CO^+, O^+ and C^+.

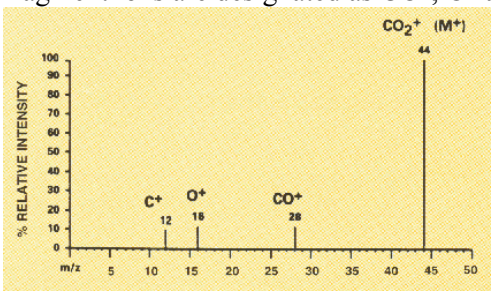


Figure: Mass spectrum of carbon dioxide, CO_2. Molecular ion is seen at m/z 44.
Picture taken from: http://www-scf.usc.edu/~lijuanmo/tutorial.files/image002.jpg

Nowadays, mass spectrometers can measure the mass of a molecule with high accuracy. It is important to have high resolution and accuracy as any biological molecule will display several MS peaks in a normal MS spectrum due to the presence of isotopes. In nature a biological molecule can contain one or more atoms with a higher isotopic mass (C^{13} or N^{15}). Therefore, the mass spectrum will show several peaks representing the different isotopic masses of the molecule. For small molecules the mono-isotopic mass, the first peak (called the C^{12}-peak) will be the highest peak. The average mass represents the centroid of the various peaks.

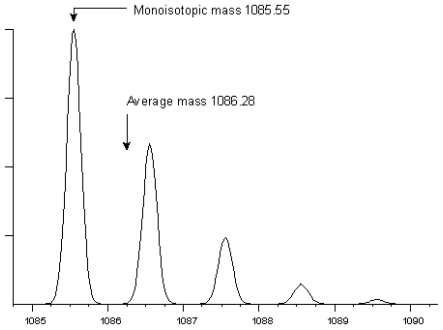


Figure: MS spectrum of a peptide. Picture taken from: http://www.matrixscience.com/images/iso_dist1.gif

For peptides and proteins, the difference between an average and a monoisotopic weight is approximately 0.06%. This is a significant difference when even the most modest instruments are capable of measuring the mass of a small peptide with an accuracy of 100 ppm (1 ppm = $1x10^{-6}$, this implies that a peptide with a mass of 1000 Da is measured with a precision of 0.1 Da). Therefore, for small peptides always the monoisotopic masses are used for identification.
The first peak will not be the largest one for proteins. Actually, for large proteins, the chance of observing the monoisotopic peak is almost zero (see the spectrum of BSA below). In order to be able to identify the first peak with a certain amount of confidence there should be sufficient mass resolution to separate the various isotopes and a high signal to noise ratio. Therefore, for larger proteins, the average mass value is used for identification. Most MS instruments have a certain cut-off point until when they report monoisotopic weights. Above that cut-off they report average mass values.

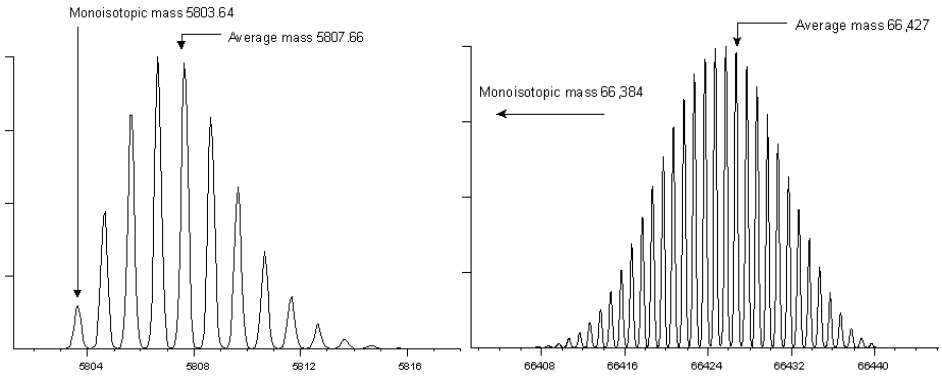


Figure: MS spectra of two proteins. Left a relatively small protein (insulin). Right a large protein (BSA). For the identification of both proteins, the average mass is used. Picture taken from: http://www.matrixscience.com/images/iso_dist2.gif http://www.matrixscience.com/images/iso_dist3.gif.

Obtaining information on molecules using Mass Spectrometry can be divided into three stages:

- Sample preparation and creation of gas-phase ions (**ionisation**)
    - For biological samples to be analysed by MS, the molecules must be charged and vaporised. This is accomplished by converting the (usually uncharged) molecules into gas phase ions. The two most common methods for ionisation are electrospray

ionisation (ESI) and matrix-assisted laser desorption/ionisation (MALDI). In both cases, peptides are converted to ions by the addition or loss of one or more protons. ESI and MALDI are "soft" ionisation methods that allow the formation of ions without significant loss of sample integrity. This is important because it enables accurate mass information to be obtained about proteins and peptides in their native states. MALDI ion sources are most commonly coupled with time-of-flight (TOF) mass analysers, whereas ESI can also be coupled to ion trap or triple-quadrupole MS spectrometers.

- Separation of the ions in space or time based on their mass-to-charge ratio
  - Once the proteins are ionised their mass and charge must be determined. The ions are first collected and then they go through the **mass analyser** where they are sorted according to their mass-to-charge (m.z) ratio. Ion-trap, quadrupole and time of flight (TOF) analysers are used most often as mass analyser.
- Measuring and analysing the quantity of ions of each mass-to-charge ratio
  - The time it takes individual groups of identical molecules to move through the mass analyser to the **detector** is measured. In the detector the ion flux is converted to a proportional electrical current. The data system records the magnitude of these electrical signals as a function of m/z and converts this information into a mass spectrum. A mass spectrum is a graph of ion intensity (of all ionisable components in a sample) as a function of mass-to-charge ratio. From the mass spectrum, the molecular weight and the structure of the molecule that was being analysed can be determined.
  - Peptides, either natural or from a protein digest, can be analysed either with MALDI-MS or ESI-MS mass spectrometry as monoisotopic masses with a mass accuracy of 100 ppm (1 ppm = $1x10^{-6}$) or less. This implies that a peptide with a mass of 1000 Da is measured with a precision of 0.1 Da. If only the peptide masses are of interest, MALDI-TOF mass spectrometry is mainly used because the spectra are easy to interpret. In the positive ionisation mode the peptide acquires one proton (H^+) and is analysed as a singly charged protonated mass. The practical amount of a peptide required for MALDI-TOF analysis is a few pmols. Electrospray mass spectrometry of peptides is a little bit more complicated because peptides acquire multiple charges depending on their size. Typically peptides are analysed as singly-, doubly- and triply charged molecules. In tandem (ESI-MS/MS) mass spectrometric analysis it is possible to obtain partial sequence data from peptides (See: MS-MS).

Mass Spectrometers

Most mass spectrometers consist of four basic elements: (i) an ionisation source, (ii) one or more mass analysers, (iii) an ion mirror, and (iv) a detector. The names of the various instruments are derived from the name of their ionisation source and the mass analyser

Ionisation sources

Matrix-Assisted Laser Desorption/Ionisation (MALDI)

When MALDI is used, the samples of interest are solidified within an acidified matrix of small energy-absorbing molecules and spotted on a metal MALDI plate (which is held under vacuum).

Figure: MALDI sample plates. Modern MALDI sample plates (size ~ 12x10 cm) can hold 400 samples or more. Picture taken from: http://www.anagnostec.eu/uploads/pics/20060228111335sapmle_plate.jpg.

A laser in a specific UV range is targeted to places on the plate where a sample is located. The matrix molecules absorb and spread the energy from the laser thereby vaporising the matrix and simultaneously ejecting the analytes into the gas phase where they acquire charge. A strong electrical field between the MALDI plate and the entrance of the MS Time of Flight (TOF) tube forces the charged analytes to enter the TOF tube with different speeds based on their mass-to-charge (*m/z*) ratios. Since a robot can perform sample application, the entire process including data collection and analysis can be automated. Another advantage of MALDI over ESI is that samples can often be used directly without any purification after in-gel digestion.

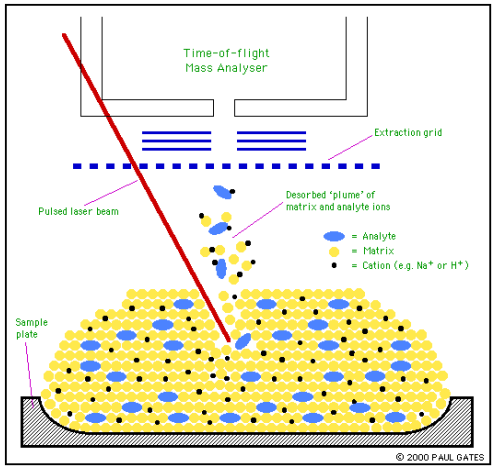


Figure: Schematic overview of MALDI – TOF mass spectrometry.
Picture taken from: http://wwwmethods.ch.cam.ac.uk/meth/ms/theory/maldi.html

Electro-Spray Ionisation (ESI)

The Electro-Spray Ionisation process is achieved by spraying a solution (such as the effluent of a HPLC column) through a high voltage (4.5 kV) charged needle at atmospheric pressure directed towards the inlet of the mass spectrometer into the source chamber, causing small droplets to form. The spray device creates a fine mist of charged droplets, which go through a process of solvent evaporation (see figure: N2 drying gas inlet) to remove the solvent. As a result only charged analytes are left in the gas phase and these enter the MS analyser.
In most cases the electric charge on the needle is positive, thus forming positive ions of the biological molecules. Actually, many mass spectrometrists in biological sciences use the phrase: "think positive". An exception is made for molecules already carrying a negative charge by itself (e.g. glycoproteins, phosphorylated proteins etc.), which can also be measured in the negative mode (a negative needle potential, analysing negative ions). Molecules that ionise easily in positive mode often are more difficult to ionise in the negative mode and vice versa. It is therefore useful to apply both ionisation methods to your sample.

ESI is particularly useful for large biological molecules that are difficult to vaporize or ionise. A slight disadvantage of the ESI source is that the signal intensity is flow dependent. In other words, the higher the eluent flow (the more solvent !), the lower the amount of ionisation and therefore

the lower the MS signal. With nanospray ionisation, the microcapillary tube has a much smaller outlet and lower flow-rates. The low flow rates reduce the amount of solvent used , thereby increasing the quality of the signal and increasing the signal to noise ratio.

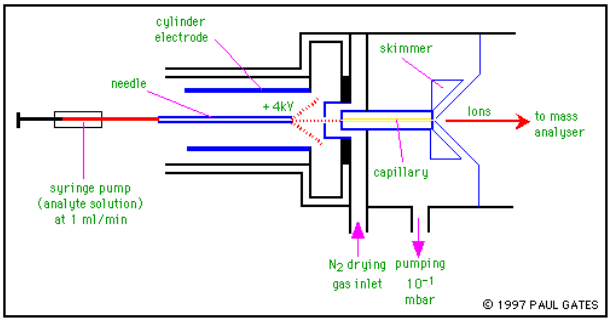


Figure: An ESI ion source. *Picture from* http://www-methods.ch.cam.ac.uk/meth/ms/theory/.

For ESI, there are several ways to deliver the sample to the mass spectrometer. The simplest method is to load individual microcapillary tubes with sample (sub-microliter range). When peptides require some form of purification after in-gel digestion, this can be accomplished directly in the microcapillary tubes ("zip-tips"). The drawback to both the purification and manual loading of microcapillary tubes is that it is tedious and slow. As an alternative, electrospray sources can easily be connected in line with liquid chromatography (LC) or high pressure LC (HPLC) systems that automatically purify and deliver the sample components to the mass spectrometer allowing high-throughput and on-line analysis of peptide or protein mixtures. It is easy because ESI is directly compatible with the solvents that are used to solubilise peptides in LC or HPLC. Typically, proteins in a complex mixture are separated by ionic (separation based on charge) or reverse phase (separation based on hydrophobicity) column chromatography and subjected to MS analysis. In this way it is possible to separate and identify proteins with identical mass and to identify proteins from complex protein mixtures, e.g. protein complexes isolated by immunoprecipitation.

Mass analysers

Ion-trap: The ion-trap mass spectrometer uses three electrodes to trap ions in a small 3D electric field. The mass analyser consists of a ring electrode separating two hemispherical endcap electrodes. Ions are trapped by applying specific AC voltages to the electrodes. A mass spectrum is obtained by changing the electrode voltages to eject the ions from low mass to high mass (in fact, m/z) from the trap.
In contrast to a quadrupole mass analyser, in which ions are discarded before the analysis begins, the main advantage of an ion trap mass analyser is the ability to allow ions to be "stored" and then selectively ejected from the ion trap, increasing sensitivity.

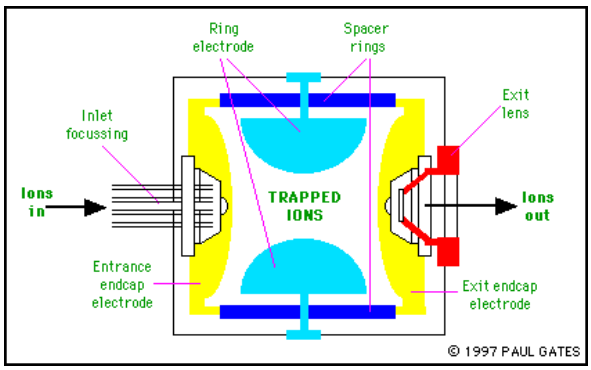


Figure: An ion trap mass analyser. *Picture from http://www-methods.ch.cam.ac.uk/meth/ms/theory/iontrap.html.*

Quadrupole: A quadrupole mass filter consists of four parallel metal rods (the quadrupole) in high vacuum arranged as in the figure. Two opposite rods have an applied potential of (U+Vcos(wt)) and the other two rods have a potential of -(U+Vcos(wt)), where U is a DC(direct current) voltage and Vcos(wt) is an AC (alternating current) voltage. The applied voltages affect the trajectory of ions travelling down the flight path centred between the four rods. For given DC and AC voltages, only ions of a certain mass-to-charge ratio pass through the quadrupole filter and all other ions are thrown out of their original path. Doing this for a series of mass-to-charge values, finally yields a complete spectrum. A complete mass spectrum can be obtained by monitoring the ions passing through the quadrupole filter as the voltages on the rods are varied. If multiple quadrupoles are combined, they can be used to obtain information about the amino acid sequence of a peptide.

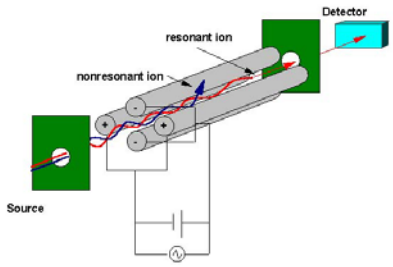


Figure: Quadrupole mass spectrometers consist of an ion source, ion optics to accelerate and focus the ions through an aperture into the quadrupole filter, the quadrupole filter itself with control voltage supplies, an exit aperture, an ion detector and electronics, and a high-vacuum system. Picture taken from: http://www.files.chem.vt.edu/chem-ed/ms/graphics/quad-sch.gif

TOF = time of flight: A time-of-flight (TOF) instrument is one of the simplest mass analysers. It measures the *m*/*z* ratio of an ion by determining the time required for it to traverse the length of a drift region (flight tube). It operates in a pulsed mode so ions must be produced or extracted in pulses. An electric field accelerates all ions into a field-free drift region with a kinetic energy of qV, where q is the ion charge and V is the applied voltage. Since the ion kinetic energy is $0.5mv^2$, lighter ions have a higher velocity than heavier ions and reach the detector at the end of the drift region sooner. The reflectron is a series of rings or grids that act as an ion mirror. Some TOF mass analysers include an ion mirror at the end of the flight tube, which reflects ions back through the flight tube to a detector. In this way, the ion mirror serves to increase the length of the flight tube. The ion mirror also corrects for small energy differences among ions. Both of these factors contribute to an increase in mass resolution. One has to control the temperature of the room where the TOF MS instrument is located very well, as the length of the TOF tube increases or decreases with temperature. This would require recalibration every time when the temperature changes.

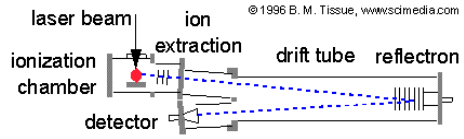


Figure: Time of Flight (TOF) mass analyser
Picture taken from: http://www.richmond.edu/~jbell2/maldi.jpg

Mass Detection

The analysis of proteins or peptides by MS can be divided into two general categories:

- Peptide mass mapping / peptide fingerprinting
- Amino acid "sequencing"

In peptide mass analysis or peptide mass fingerprinting, the masses of individual peptides in a mixture are measured and used to create a mass spectrum. In amino acid sequencing, a procedure known as tandem mass spectrometry, or MS/MS, is used to fragment a specific peptide into smaller peptides, which can then be used to deduce the amino acid sequence.

Peptide mass mapping

Mass spectrometers can determine the mass of a protein or peptide with a high degree of accuracy, but the intrinsic mass of a eukaryotic protein is not a uniquely identifying feature. However, the masses of the various peptides generated by fragmentation of an isolated protein with an enzyme of known cleavage specificity can uniquely identify a protein. This protein identification method is called peptide mass mapping or peptide mass fingerprinting. In this method, the masses of peptides obtained from the proteolytic digestion of an unknown protein are compared to the predicted masses of peptides from a theoretical digestion of proteins in a database. If enough peptides from the measured mass spectrum and the predicted mass spectrum overlap, a protein identification can be made. The success of this type of analysis is largely dependent on the mass accuracy of the mass spectrometer and to a lesser extent on the number of peptides identified from each protein species. The reliability of the identification of the protein depends on the size of the database searched (a smaller database, e.g. *E. coli*, provides a more easy identification of the protein of interest than when using a large database, e.g. human, mouse etc), the accuracy of the measured peptides and the number of peptides identified.

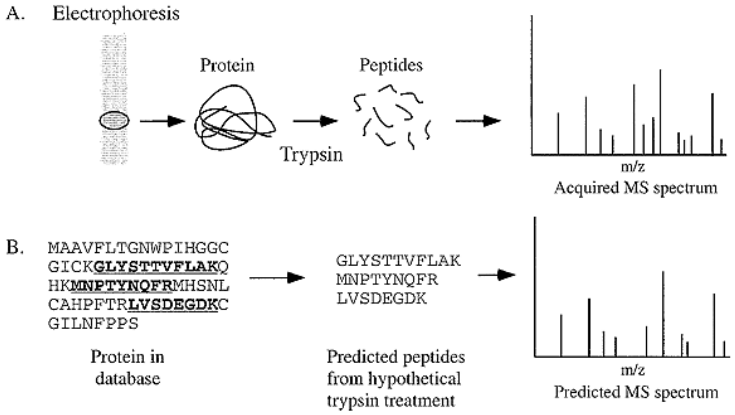


Figure: Strategy of protein identification by peptide mass fingerprinting (A) The unknown protein is excised from a gel and converted to peptides by the action of specific protease. The mass of the peptides produced is then measured in a mass spectrometer. (B) The mass spectrum of the unknown protein is searched against theoretical mass spectra produced by computer generated cleavage of proteins in the database.
Picture taken from: http://mmbr.asm.org/content/66/1/39/F9.small.gif

Identification of the protein measured via MS proceeds through the following sequence of steps:

1. Peptides are generated by digestion of the sample protein using sequence-specific cleavage enzymes (proteases). Residues at the carboxyl- or amino-terminus of the peptides, generated through cleavage with the proteases used, are considered fixed for the search. For example, the enzyme trypsin that is most frequently used for mass mapping generates arginine (R) or lysine (K) at the carboxyl-terminus of the peptides.
2. Peptide masses are measured as accurately as possible in a mass spectrometer. In modern mass spectrometers the accuracy is often 100 ppm (1 ppm = $1x10^{-6}$) or less. This implies that a peptide with a mass of 1000 Da is measured with a precision of 0.1 Da. An increase in mass accuracy will decrease the number of isobaric peptides (peptides with the same mass) for any given mass in a sequence database and therefore increase the stringency of the search.

3. The proteins in the database are "digested" in silico using the rules that apply to the proteolytic method used in the experiment to generate a list of theoretical masses that are compared to the set of measured masses.
4. An algorithm is used to compare the set of measured peptide masses against those sets of masses predicted for each protein in the database and to assign a score to each match that ranks the quality of the matches. Obviously, for a protein to be identified its sequence has to exist in the sequence database being used for comparison. Both protein and DNA sequence databases are equally suited. If DNA sequence databases are being used, the DNA sequences are translated into protein sequences prior to digestion. The approach is therefore best suited for genetically well-characterized organisms where either the entire genome is known or extensive protein or cDNA sequence exists.

Limitations of Mass Mapping

The principal advantage of peptide mass fingerprinting is speed, because the analysis and database search can be fully automated. However, there are also some limitations to the method.

- It is important to realise that peptide mass fingerprinting results in prediction of the amino acid composition of peptides but not the sequence itself! Protein identification by peptide mass mapping depends on the correlation of several peptide masses derived from the same protein with corresponding masses calculated from the theoretical digestion of that protein in the database. For this reason the method is not suited for
    - searches of EST databases, because ESTs only represent a portion of a gene's coding sequence. This portion may be not long enough to cover a sufficient number of peptides to identify the protein.
    - identification of proteins in complex mixtures, because it is not clear which peptides in the mixture belong to the same protein.

- Identification of a peptide by its mass tells you what amino acids are present in that peptide, but not the order of the amino acids. This problem is called peptide mass redundancy. For example, a peptide of 5 amino acids can have the same mass by simple rearrangement of its amino acids; e.g., peptide VAGSE has the same mass as AVGSE or AEVGS and so on. This problem can be overcome by using the masses of more peptides to obtain enough specificity for successful protein identification. However, this is not always possible. Note that mass redundancy occurs with greater frequency in large genomes.

- Peptide mass fingerprinting is effective only in the analysis of proteins from organisms whose genome is small (not so many isoforms), completely sequenced, and well annotated. If not, it is possible that peptide (or DNA) sequences are not present in the database preventing identification of the protein.

- Not all predicted peptides are detected; some peptides are missing. However, missing peptides is not a problem, since a relatively low number of peptide masses are sufficient for the identification of a protein.
    - MS may miss peptides, due to poor solubility, selective absorption, ion suppression, selective ionisation (some peptides do not ionise very well), very short or very large peptide length or other artefacts that cause sample loss or make specific peptides undetectable.
    - Missing peptides may represent alternative gene splicing. Further experimentation necessary.

- Some of the measured peptide masses are not present in the list of masses predicted from the protein. This may be due to:
    - changes in the expected peptide masses by:
        - post-translational modification (e.g. phosphorylation adds 80 u to an amino acid mass).
        - artefactual modifications arising from sample handling (such as oxidation of methionine).
        - post-translational processing (e.g. amino- or carboxyl-terminal processing).
    - contamination with other proteases.
    - the presence of more than one protein in the sample.
    - matching with a sequence homologue or a splice variant present in the database.
    - misidentification of the protein.
- Proteases never work perfectly. Some cleave at sites not expected, or miss cleavage sites due to the three dimensional structure of the protein or a modification that blocks access of the enzyme to the cleavage site.

Amino acid sequencing

The amino acid sequence of a single peptide is more precise than its mass for protein identification. For a small database, the masses of 3 or more peptides from a tryptic digest of a protein are necessary for the proper identification of a protein (peptide mass fingerprinting). For larger databases of course more peptides are needed for identification of the protein. However, in general not more than 10 peptides are needed to identify the protein under study, even for the most extensive database. In contrast, the amino acid sequence of a peptide can uniquely identify a protein. Of course, this is only possible if the amino acid sequence is known in the database. If the amino acid sequence is not present in the database, MS/MS can be used for protein identification.
From a MS/MS spectrum it is possible to identify the amino acid sequence of a peptide. Because this takes additional time (relative to peptide mass fingerprinting), it is not the first choice for high throughput protein identification. Rather, the much faster methods of peptide mass fingerprinting or peptide mass tag searching can be used first. If this method fails (e.g. not enough protein material, mass redundancy or no well-annotated database) or when more information of the peptides is needed, MS/MS can be used to identify the amino acid sequence of peptides. In theory it is possible to obtain the whole sequence of a protein using MS/MS methods. However this is hardly ever achieved, due to the large amount of protein needed and the fact that not all peptides are easily ionised.

The advantages of MS/MS are, that it is possible to

- do protein identification from complex mixtures of proteins, as only one peptide per protein is already enough for protein identification.
- compare the peptide sequence with DNA, protein as well as EST databases for protein identification.
- analyse proteins whose sequence is not known yet or
- analyse proteins from organisms that do not have well-annotated databases such as *Xenopus laevis*.

MS/MS

MS/MS consists of two parts. First, a mixture of charged molecules (indicated as arrows) is separated according to their m/z ratios to create a list of the most intense peaks. Then a single

molecule, the "parent" ion (indicated as the longer arrow), is selected and directed into a collision cell to generate "daughter" ions. The collision cell, or activation chamber, is filled with a gas (usually nitrogen or argon). This inert gas collides with the ionised molecule, and the vibrational energy of the gas causes the molecule to break into pieces. The newly generated fragments are then separated according to their m/z ratio, creating the MS/MS spectrum. For proteomics this means that a peptide is selected for identification of its amino acid sequence.

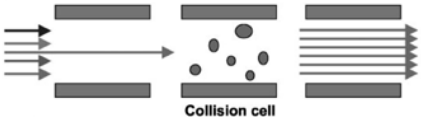


Figure: MS/MS. *Picture from Zhu et al. 2003.*

Peptides can undergo multiple types of fragmentation, therefore, nomenclature has been created to indicate what types of ions have been generated. If, after peptide bond cleavage, the positive charge of the amide bond remains on the ion coming from the N-terminal part of the peptide, it is designated a b-ion, whereas if the charge is remains on the ion coming from the C-terminal part of the peptide, it is called a y-ion.

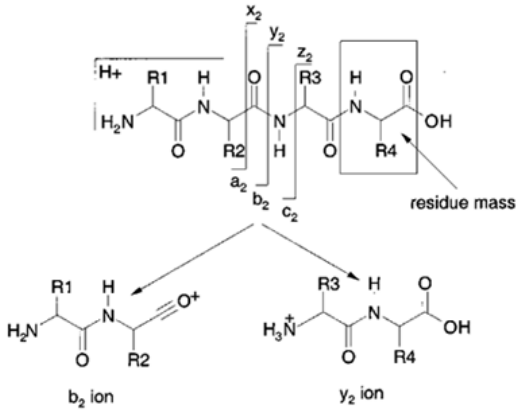


Figure: Nomenclature peptide fragmentation. *Picture from Aebersold and Goodlett, 2001.*

Since a protein can break in many places, more than one combination of b and y-ions will be produced. Therefore, subscripts are used to designate the specific amide-bond that was fragmented. b-ions are designated by a subscript that reflects the number of amino acid residues present on the fragment ion counted from the amino-terminus, whereas the subscript of y-ions indicates the number of amino acids present counting from the carboxy-terminus.

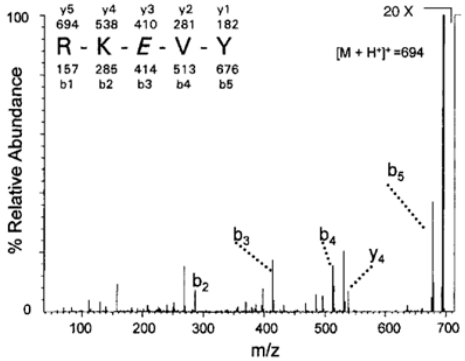


Figure: MS/MS spectrum of peptide RKEVY. *Picture from Aebersold and Goodlett, 2001.*

The difference in mass between adjacent b- or y-ions corresponds to that of an amino acid. This can be used to identify the amino acid and thus the peptide sequence. A difficulty is the identification of isoleucine and leucine, which are identical in mass and therefore indistinguishable.

If you know the amino acid sequence of the peptide, the fragment ion m/z values can easily be calculated using the (neutral) residue masses found in the Table. To calculate the masses of the b-ion series, 1 m/z unit (for 1.008 H) is added to the nominal mass of the first residue (b_1). To calculate the masses for the b2, b3 and following fragment ions, this process is continued by the addition of the nominal mass for the second, third, and following amino acid residues, respectively, until the final, C-terminal amino acid is included. To calculate the masses for the y-ion series, 19.022 m/z units (for H_3O) is added to the residue mass of the C-terminal amino acid. As for the b-ion series, this process is continued with the addition of the residue mass of the following amino acids (y_2, y_3, …, y_n).

Table: Amino acid reference data. The data in this table are for amino acid residues. To calculate the mass of a neutral peptide or protein, sum the residue masses plus the masses of the terminating groups (e.g. H at the N-terminus and H O at the C-terminus). *Taken from http://www.matrixscience.com/help/fragmentation_help.html* [3]

Residue	3-letter code	1-letter code	Formula	Mono-isotopic mass	Average mass
Alanine	Ala	A	C_3H_5NO	71.03712	71.08
Arginine	Arg	R	$C_6H_{12}N_4O$	156.10112	156.19
Asparagine	Asn	N	$C_4H_6N_2O_2$	114.04293	114.1
Aspartic acid	Asp	D	$C_4H_5NO_3$	115.02695	115.09
Cysteine	Cys	C	C_3H_5NOS	103.00919	103.14
Glutamic acid	Glu	E	$C_5H_7NO_3$	129.0426	129.12
Glutamine	Gln	Q	$C_5H_8N_2O_2$	128.05858	128.13
Glycine	Gly	G	C_2H_3NO	57.02147	57.05
Histidine	His	H	$C_6H_7N_3O$	137.05891	137.14
Isoleucine	Ile	I	$C_6H_{11}NO$	113.08407	113.16
Leucine	Leu	L	$C_6H_{11}NO$	113.08407	113.16
Lysine	Lys	K	$C_6H_{12}N_2O$	128.09497	128.17
Methionine	Met	M	C_5H_9NOS	131.04049	131.19
Phenylalanine	Phe	F	C_9H_9NO	147.06842	147.18
Proline	Pro	P	C_5H_7NO	97.05277	97.12
Serine	Ser	S	$C_3H_5NO_2$	87.03203	87.08
Threonine	Thr	T	$C_4H_7NO_2$	101.04768	101.1
Selenocysteine	SeC	U	C_3H_5NOSe	150.95364	150.03
Tryptophan	Trp	W	$C_{11}H_{10}N_2O$	186.07932	186.21
Tyrosine	Tyr	Y	$C_9H_9NO_2$	163.06333	163.18
Valine	Val	V	C_5H_9NO	99.06842	99.13

However, while it is relatively simple to calculate the m/z values of the b-and y-ion series from the peptide sequence, it is much less straightforward to read the amino acid sequence from the MS/MS spectrum as you can see in the spectrum above. This is because the rules of fragmentation are not completely understood. The types of fragment ions observed in an MS/MS spectrum depend on many factors including primary sequence, the amount of internal energy, how the energy was introduced, charge state, etc. When for instance, the N-terminus contains relatively a lot of basic amino acids (arginine, lysine), then the charge is most likely located on the N-terminus giving rise to only b-type ions.

Each peptide MS/MS spectrum consists of b- and y-ions, but both b- and y-type ions can also eliminate NH_3 (-7 u), or CO (-28 u), resulting in x- and a-ions or z- and c-ions respectively. These

ions can be observed in the mass spectrum as pairs of signals and can be used to as additional information to interpret the amino acid sequence.
In addition to fragmentation along the peptide backbone, cleavage can also occur along amino acid side chains, and this information can, for example, be used to distinguish isoleucine and leucine. Other examples are Gln, Lys and Arg that can lose the ammonia group (-17 u) and Ser, Thr, Asp and Glu that can lose water (-18 u). For the formulas to calculate the m/z values of these types of fragment ions have a look at
http://www.matrixscience.com/help/fragmentation_help.html.

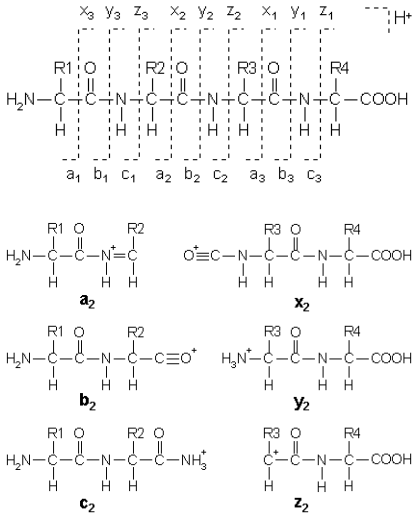


Figure: Nomenclature peptide fragmentation. *Picture from Matrix Science, Picture taken from: http://www.matrixscience.com/images/cleavages.gif*
http://www.matrixscience.com/images/abcxyz.gif

Because it is difficult to read the amino acid sequence from the MS/MS spectrum, it may take between half an hour and a couple of days to interpret the data by hand. Therefore, computer algorithms were developed that combine several sources of information to identify the peptide that was analysed by MS/MS. For example, the mass of the parent ion and the uninterpreted fragment ion pattern can be used as input for a search through the sequence information present in DNA, EST and protein databases, to identify the protein from which the peptide was derived. The algorithm SEQUEST first creates a list of peptide masses that correspond to the mass of the parent peptide. For each of these candidate peptides, the program calculates the masses of the expected fragment ions and generates a theoretical MS/MS spectrum for comparison. The program does not take the chemical properties of the amino acids present in the peptide into consideration, so all predicted fragment ions are equal in intensity (normally the fragment ions are different in intensity, see Figure). A large number of theoretical spectra are then compared with the observed spectrum and a score is given to each of them, which is based on a number of parameters such as the number of fragment ions predicted versus found. This will rank all possible candidates relative to each other, so that the top scoring peptide has the highest probability of being the peptide that was analysed by MS/MS.

Quantitative mass spectrometry

Using MS it is difficult to measure quantitative changes in protein abundance without direct visualisation of the proteins in gels. Peptides analysed in a mass spectrometer produce different specific signal intensities depending on their chemical composition, on the matrix in which they are present and on other poorly understood variables. In addition, samples prepared for mass spectrometry are handled separately and are therefore subject to different sample-handling errors. For these reasons, the intensity of a peptide ion signal does not accurately reflect the amount of

peptide in a sample; in other words, mass spectrometry is inherently not a quantitative technique. However, two peptides of identical chemical structure that differ in mass because they differ in isotopic composition are expected to generate identical specific signals in a mass spectrometer. ICAT makes use of this principle to compare the ratio of abundance of a peptide between two samples.

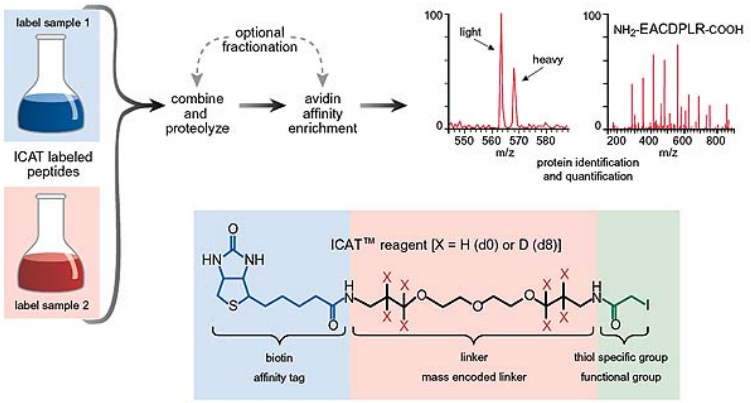


Figure: ICAT is an isotope-coded affinity tag reagent. The reagents consist of a protein reactive group that covalently attaches the reagent to reduced cysteine residues and an affinity group (biotin affinity tag) through which tagged peptides are selectively isolated. The chemical modifying group is linked to the affinity group through a mass-encoded linker, containing either 8 hydrogen atoms (d0) or 8 deuterium atoms (d8). The general scheme used for this reagent is shown: first, the protein samples are labelled separately with either light or heavy reagent, then the proteins are mixed and digested by enzymes; second, the labelled peptides are captured and then quantified and identified by LC-MS/MS. *Picture from Patterson and Aebersold, 2003.*

The ICAT approach involves labelling the cysteine residues in one sample with d0-ICAT reagent and the cysteine residues in a second sample with the d8-ICAT reagent. The samples are then combined. Alternatively, the cell lysate can be fractionated prior to reaction with the ICAT reagent. This can allow the enrichment of low-abundance proteins before the analysis begins. After enzymatic digestion of the combined samples, the ICAT-labelled peptides are isolated by affinity chromatography (using streptadivin to bind the biotin affinity tags) and analysed by LC-MS/MS.
Each cysteinyl peptide appears as a pair of signals, representing light and the heavy form respectively, which are separated by 8 Da (d0 vs. d8). The difference in peak heights between heavy and light peptide ions directly correlates with the difference in protein abundance in the cells. Thus, if a protein is present at a threefold higher level in one sample, this will be reflected in a threefold difference in peak heights. The MS/MS spectrum of either isotopic form of the peptide allows the protein to be identified. Thus in a single, automated operation this method identifies the proteins present in two related samples and determines the ratio of relative abundance.

The ICAT method works well for the differential analysis of many proteins in a complex mixture. The obvious limitation of the ICAT labelling approach is that a protein has to contain at least one cysteine residue to be detected. Therefore, many variations of this approach were developed. For example *in vivo* metabolic labelling of proteins with stable isotopes such as ^{15}N. Small organisms like *E. coli* or yeast can be grown on media enriched with ^{15}N to achieve >98% labelling. These organisms use the ^{15}N to make amino acids for protein synmthesis resulting in ^{15}N labelled proteins. Larger organisms like *C. elegans* or Drosophila can be fed with ^{15}N-labelled *E. coli* and yeast resulting in ~96% of the total amount of proteins that is labelled. By growing one batch of organisms, e.g. wildtype, on ^{14}N and another one, e.g. mutant A, on ^{15}N the relative abundance of individual proteins can be determined by mass spectrometry (of course after protein isolation).

For more information about this method read: *Krijgsveld et al. (2004) Metabolic labelling of C. elegans and D. melanogaster for quantitative proteomics. Nature Biotech. 21(8): 927-931.*

Protein microarrays

To obtain detailed information about a complex biological system, information on the state of many proteins is required. The analysis of the proteome of a cell (i.e. the quantification of all proteins and the determination of their post-translational modifications and how these are dependent on cell-state and environmental influences) is not possible without novel experimental approaches. High-throughput protein analysis methods allowing a fast, direct and quantitative detection are needed. Efforts are underway, therefore, to expand microarray technology beyond DNA chips and establish array-based approaches to characterise proteomes.

However, there are some differences between DNA and proteins. DNA is a very uniform and stable molecule, which binds its complementary targets by means of the well-defined base-pairing principle. Due to the complementary nature of the DNA molecule, interaction sequences (capture sequences) can easily be predicted from the primary DNA sequence of the target molecules. Efficient oligonucleotide synthesis or PCR-based approaches enable the fast and economic generation of a large variety of DNA capture molecules. This is different for proteins. There is no one-by-one interaction as observed for DNA base pairing. Proteins exhibit very diverse and individual tertiary molecular structures. Proteins are inherently unstable outside a narrow range of environmental conditions and usually they need to be captured in their native conformation. Their binding interaction takes place by different means such as electrostatic forces, hydrogen bonds and/or weak hydrophobic Van der Waals interactions, which are less specific and of lower and more variable affinity than those between Watson-Crick base–paired nucleic acids. In addition, individual proteins can even interact with different binding partners at the same time and in a synergistic way, forming functional protein complexes. At present, there is no way which would allow the prediction of high affinity protein capture molecules only on the basis of their primary amino acid sequence. Steady or dynamic post-translational modifications like glycosylation, phosphorylation, and acetylation must also be taken into consideration. Furthermore, proteins cannot be amplified before analysis. Therefore detection methods must be very sensitive. Proteins have no inherent properties that make them measurable at high sensitivity and the attachment of a detectable tag is prone to interfere with a protein's interactions. Therefore, the wide and dynamic range of protein abundances makes it hard to detect low-abundance proteins against a high-abundance background.

Table: Properties of DNA and proteins with respect to their application in microarray technology

Properties	**DNA**	**Protein**
Structure	Uniform Hydrophilic acidic backbone Stable	Hydrophobic and/or hydrophilic domains Fragile
Functional state	Denatured, no loss of activity >> can be stored dry	3D structure important for activity >> avoid denaturation
Interaction sites	1 by 1 interaction	Multiple active interaction sites
Interaction affinity	High	Dependent on individual protein: very low to high
Interaction specificity	High	Dependent on individual protein: very low to high
Activity prediction	Well defined Based on primary nucleotide Sequence	Not possible yet. Efforts are undertaken to predict models that are based on sequence homologies, structure, *etc.*
Amplification	Established (PCR)	Not available yet

Although there are difficulties, there is substantial benefit to be gained from using microarray technology with proteins. First, in principle thousands of proteins can be spotted on a single slide or similar support, allowing the identification and quantification of a large number of target proteins from a minute amount of sample. Second, hundreds or even thousands of copies of an array can be fabricated in parallel, enabling the same proteins to be probed repeatedly with many different molecules under many different conditions.
A unique advantage of protein function microarrays is that they can be used to study the interaction of proteins with other non-protein molecules, including nucleic acids, lipids, small organic compounds or DNA.

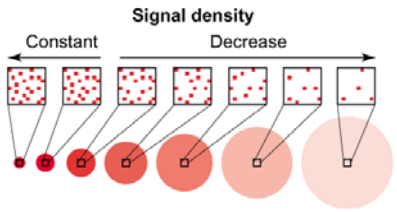


Figure: The highest signal intensities and optimal signal-to noise ratios can be achieved in small spots. Figure adapted from Templin et al. 2002.

With microarray techniques, high sensitivity can be achieved due to the fact that a system that uses a small amount of capture molecules and a small amount of sample can be more sensitive than a system that uses a hundred times more material. There are two reasons that explain this phenomenon. First, the binding reaction occurs at the highest possible target concentration. Second, the capture molecule-target complex is found only in the small area of the microspot resulting in a high local signal. Thus, the highest signal intensities and optimal signal-to noise ratios can be achieved in small spots.

Thus far, protein-microarray based interaction analysis has been described for the analysis of protein-protein, enzyme-substrate, protein-DNA, protein-oligosaccharide and protein-drug interactions. Low and high-density protein arrays were used to investigate the binding of DNA, RNA, small chemical ligands and proteins. Enzyme-substrate assays were performed for restriction enzymes, phosphatases, peroxidases and phosphokinases. All these applications have the potential to provide functional data on a proteome-wide scale.

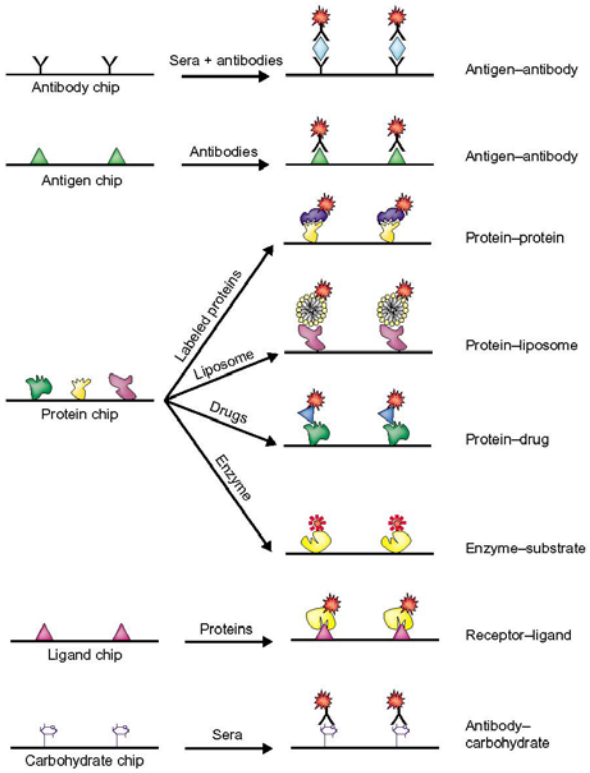


Figure: Different forms of protein microarrays. Figure from Zhu and Snyder, 2003.

Two types of protein arrays

There are two general types of protein microarrays, analytical arrays and functional arrays.

Analytical microarrays contain an ordered array of antibodies, antibody mimics, proteins or other protein-ligands like DNA, RNA, carbohydrates or small molecules. These capture agents have high affinity and specificity for a certain protein. Each agent captures its target protein from a complex protein mixture (such as serum or cell lysate), and the captured proteins are subsequently detected and quantified. In this way, you can measure the presence and concentrations of proteins in a complex mixture. Similar to the procedure in DNA microarray experiments, protein samples from two biological states to be compared are separately labelled with red or green fluorescent dyes, mixed, and incubated with the chips. Spots in red or green colour identify an excess of proteins from one state over the other. Monitoring protein expression on a large scale is a process that is sometimes called protein profiling.

The most common form of analytical arrays are antibodies/antibody mimic arrays in which antibodies (or similar reagents) that bind specific antigens are arrayed on a glass slide at high density. A lysate is passed over the array and the bound antigen is detected after washing. Detection is usually carried out by using labelled lysates or using a second antibody that recognises the antigen of interest. The biggest challenge is producing reagents that identify the protein of interest with high enough specificity in a high-throughput fashion. The most significant problem with antibody arrays is specificity. Proteins are often present in a very large dynamic range (10^6); thus, reagents that might have high affinity for one protein, but are low affinity for another will still exhibit detection of the lower affinity protein if it is present in a higher

concentration. Due to the high cross-reactivity of some antibodies, results obtained with antibody microarrays have to be verified and confirmed with standard methods.
To avoid the problem of cross-reactivity, it is possible to use sandwich assays, in which the first antibody is spotted on the array and then the antigen is detected with a second antibody that recognises a different part of the protein. This approach dramatically increases the specificity of the antigen detection, but requires that at least two high-quality antibodies exist for each antigen to be detected.

Regulated protein–protein interactions may complicate the interpretation of data arising from either antigen capture or sandwich assays. For example, a protein may be free in solution under one set of conditions and be part of a large, multi-protein complex under another set. Due to complex formation, the epitope recognised by the antibody may become obscured in the complex, resulting in an apparent decrease in signal on complex formation. Or, the presence of detection antibodies directed against other members of the complex will result in an apparent increase in signal. Thus, although the abundance of the protein does not actually change in going from the first set of conditions to the second, an antigen capture assay could report a significant increase or decrease of the protein. Knowledge of the biology would enable the investigator to circumvent these problems. Detection antibodies could be chosen that do not target protein–protein interfaces, and arrays could be subdivided physically to avoid the detection of associated proteins.

Functional protein microarrays contain sets of proteins or even an entire proteome to systematically measure or infer biochemical activities or other functional properties of proteins. For these chips native proteins or peptides are individually purified or synthesised using high-throughput approaches and arrayed onto a suitable surface. These chips are used to analyse protein activities, binding properties or post-translational modifications, for example kinase assay chips. With the proper detection method, functional protein microarrays can even be used to identify the substrates of enzymes of interest. Consequently, this class of chips is particularly useful in drug and drug-target identification and in building biological networks.

To analyse the biochemical activities of as many proteins as possible it is necessary to purify proteins in a high-throughput manner. No proteome is available yet in the form of individual purified proteins. But purifying proteins can easily be done by affinity purification of recombinant fusion-proteins (GST fusion or His-tag fusion; see tags section in Affinity chromatography + MS). In this way proteins can be purified using various host cells, including *E. coli,* yeast, insect cells and mammalian cells. However, remember that eukaryotic proteins expressed in a prokaryotic system are not properly post-translationally modified and that the attachment of a tag to a protein may interfere with the protein's interactions. A lot of the proteins on a big array have an unknown function, so you can never actually test if they are all still functionally.

How to generate data from a protein microarray?

Among many challenges, to build a viable protein chip a manufacturer needs to:

- Choose a surface chemistry that will allow immobilised proteins of diverse types to retain their secondary and tertiary structure and thus their biological activity
- Identify and isolate an capture agent (e.g. antibody) with which to capture the protein(s) of interest
- Devise a means of measuring the degree of protein binding, assuring both sensitivity and a suitable range of operation
- In some experiments, extract the detected protein from the chip and analyse it further.

Immobilisation:

The substratum for DNA arrays is typically amine or lysine coated glass, permitting adherence of the negatively charged DNA to the positively charged coating. Proteins can be attached to various kinds of surface via diffusion, adsorption/absorption, covalent cross-linking and affinity interaction. Except for affinity attachment, proteins are usually laid on the surface in a random fashion, which may alter the native conformation of proteins, reduce the activity or proteins, or make them inaccessible to binding molecules. However, when proteins are attached to the surface via their affinity tags, it is very likely that every protein molecule uniformly attaches to the surface and, therefore, proteins are more likely to remain in their native conformation, while the binding molecules have easier access to the active sites of proteins.

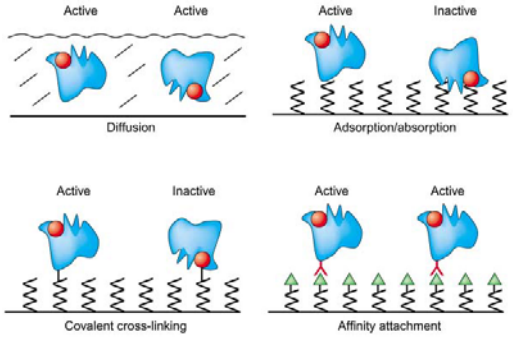


Figure: Attachment of proteins to the surface of protein microarrays. The active site of the protein is indicated by a red dot. Figure from Zhu and Snyder, 2003.

Thus, the requirements for an optimal substratum for protein arrays are

1) high binding capacity
2) it should not alter the protein structure
3) low background.

Capture agents:

A large number of highly specific capture molecules, like antibodies, antibody mimics, aptamers (nucleic acids that bind to proteins) or other ligands, which exhibit high affinity to their target molecules are a prerequisite for the establishment of protein microarrays used for the identification and quantification of target proteins.

However, antibodies cannot be manufactured with predictable affinity or specificity. Antibodies recognise the three-dimensional arrangement of charges, and the hydrophilic and/or hydrophobic properties of peptides or proteins. Such a physical landscape is not always unique to one amino acid sequence, resulting in non-specific binding of the antibody to proteins that were not used as antigen. In general, this cross-reactivity is more pronounced with polyclonal antibodies, which are a mixture of antibodies against the same antigen, than with monoclonal antibodies. Therefore, although antibodies are very useful, you have to keep in mind that binding is not necessarily proof that a product of a certain gene is present. A control may be to validate the specificity of an antibody by Western blot with a sample similar to that used on the array.

Detection:

Detection methods need to be accurate, sensitive and if possible avoid the need for labelling proteins. Detections methods are

- Chemiluminescence, which is sensitive but requires an enzymatic reaction.
- Fluorescence, which is sensitive but requires labelling of the protein
- Mass Spectrometry, which has low throughput but does not require labelling
- Surface Plasmon Resonance, which has low throughput but does not require labelling.

Fluorescence detection of binding event can be done by fluorescently labelling the proteins that will bind to the capturing agents on the array or labelling the secondary antibodies that will bind the bound proteins. However, labelling proteins reduces the quantitative accuracy of the assay, because the label can change the way the molecule binds to other molecules. Still fluorescence detection methods are generally the preferred detection method, because they are simple, safe, extremely sensitive and can have very high resolution.
Protein chips that capture proteins directly on a modified MALDI plate (SELDI protein arrays) can easily be analysed by MS. Mass spectrometry detection is best used for the detection and identification of captured proteins in at their post-translationally modified forms.

Quality control and array design

Multiple samples may be printed on each slide, representing before and after treatment samples or normal and diseased tissue lysates. Each sample is printed in duplicate or triplicate on the array, permitting statistical calculations for each sample. Samples may be printed in a miniature dilution curve on the array (4-6 spots per protein). This dilution curve ensures that the protein of interest is within the dynamic range of the assay based on the antibody sensitivity and affinity.

Accurate quantification with protein microarrays can be achieved by including positive and negative control spots and/or internal calibration spots. A positive control may be the incorporation of a (recombinant) protein and a reference lysate or peptide of known performance. The protein is serially diluted in the same manner as the test samples, providing a quantitative comparison of signal intensity for each antibody. The reference or control lysate typically consists of a pool of tissue or cell culture protein lysates of known origin and staining characteristics.

Background subtraction

Background intensity is determined by putting "empty" spots on the array layout. These are areas lacking sample, but the substratum will have been exposed to all blocking and detection reagents in the same manner as the experimental areas. The density of each background area is integrated in the same manner as the sample spot. Ideally the signal should be 2 standard deviations above background.

Normalising spot intensity

Total protein concentration variability exists between samples on the array due to the variability in the number of cells obtained per sample. This variability is controlled for each spot on the array by normalisation of the total protein per spot. Spot intensities on the array are reduced to a single value corresponding to the amount of protein in the sample.
The assumption is made that intensity above background is proportional to protein concentration. With this assumption made, the intensity versus log dilution plot should appear as an exponential decay curve. Curve fitting techniques may be applied for determining overall curve fit, background intensity and decay constants.

Data generation from protein microarrays

Each array is scanned, spot intensity analysed, data is normalised, and a standardised, single data value is generated for each sample on the array. This single data point may then be used for comparison to every other spot on the array. Each sample appears in several dilutions on the array. A binding score for each sample is determined with the following formula, where P1 represents the neat spot and P2, P3, *etc.* represent the dilution spots, and b is a bias term set to 50:

Score = (P12 P3)/[(P2 1 P3)/2 1 b]

This score attempts to reconstruct the specific binding (total minus non-specific) and expresses the result relative to the non-specific binding in that assay. The bias term increases the statistical readability of the score by preventing very small denominator values.
Empirically, the score shows stability across different arrays, and is largely resistant to variations in overall slide intensity due to variations in staining efficiency, *etc.* If trends are observed across the slide a corrected score may be calculated by subtracting the local-average score for left, middle, and right portions of each array from the score as follows:

Corrected score = score – local average score

Standardisation of score

For some antibodies, the spread of the corrected score values for the "empty" (background) spots may overlap that of the real samples, while in others, the real sample scores may be markedly above those of the empty spots. These empty spots may serve as an indicator of precision for each antibody stained array. To combine scores from different antibodies, the standard deviation of the empty spots may be used to standardise the corrected score as follows:

Standardised score = corrected score - [average(corrected score)/SD(empty corrected scores)]

The results of the each antibody labelled array in standardised score units is combined into a single data matrix. The maximum standard score for each array can be interpreted as a signal-to-noise ratio. This matrix can then be used for further analysis of the data, using for example clustering techniques.

The future

Further developments and optimisation of array production and assay performance combined with high-throughput generation of protein targets and ligands will extend the number of applications of protein microarrays dramatically. Proteomic research and diagnostic applications will be the two major fields addressed by protein microarray technologies. In medical research, protein microarrays will accelerate immune diagnostics significantly by analysing in parallel all relevant diagnostic parameters of interest. The reduction of sample volume is of great importance for all applications in which only minimal amounts of samples are available. One example might be the analysis of multiple tumour markers from a minimum amount of biopsy material. Furthermore, new possibilities for patient monitoring during disease treatment and therapy will be developed based on this emerging technology. Microarray-based technology beyond DNA chips will accelerate basic research in the area of protein-protein interactions and will allow protein profiling from limited numbers of proteins up to high-density array-based proteomic approaches. Protein and peptide arrays will be used to analyse enzyme-substrate specificity and for measurement of enzyme activity on different kinds of substrates in a highly parallel fashion. The whole field of protein microarray technology shows a dynamic development driven by the increasing genomic information. New technologies such as automated protein expression and purification systems, used for the generation of capture molecules and the need for analysis of whole 'proteomes' will be a driving force for fast developments within the field of protein microarray technology.

Structural proteomics

For many proteins the sequence is known, but amino acid sequence alone is insufficient to reveal what the proteins do, how they perform their roles, or with which proteins they interact. A basic rule of biology is that form meets function, which means that if you know the shape of a

molecule, you can understand how it works. Many times, a protein's role is not fully understood until its 3D shape is known.
Protein structural information is collected by the Protein Data Bank (PDB), an international repository for 3D structural files. Although an ever increasing number structures are deposited annually in the PDB (in 2004 almost 5000 3D structures will be deposited), the number of unique, new 3D structures/folds that are deposited is declining.
Roughly no function can be assigned to about one-third of the sequences in organisms for which the genomes have been sequenced. However, when a protein with unknown structure has 25-30% or more sequence identity to a structure in the PDB, the structure of the unknown protein usually can be modelled accurately. This process is called "homology modelling". If the sequence identity is less than 25-30%, modelling the unknown sequence can lead to significant errors, even though the type of fold may be identified.
Nowadays, several research efforts focus on the 3D structure of proteins (structural genomics) with the ultimate goal to obtain 3D structures for all proteins in a proteome. By determining the structure of at least one member of all sequence families sharing 30% or more sequence identity it should be possible to reach this goal. When there is a 3D structure from one member of a family, the structure of the other family members can be modelled with significant accuracy using homology modelling. It is estimated that when the structural database reaches the number of 50.000 3D structures (2007?) 80% of all sequences can be modelled reliably using homology modelling. The resulting 20% of the sequences are difficult to model. Most of them are related to membrane proteins, or proteins that are heavily glycolisated.

How to get a structure

In order to obtain detailed information about a protein's structure at the atomic level, you need to be able to purify the protein in large amounts, crystallize the protein, and then use either X-ray crystallography or nuclear magnetic resonance (NMR) to determine the precise location of every atom in the protein. The rate-limiting step for structural proteomics is the production of samples and crystals.

•Crystallography by X-ray diffraction
–most reliable technique to date
–depending on proteins that want to crystallize

•Nuclear Magnetic Resonance
–although magnets become stronger, only relatively small structures or domains can be solved (up to 35 kDa).
–no need to make crystals
–yields distance information
–relies on distance geometry algorithms to convert distance information to 3D-model

•Mass Spectrometry
–can be used for elucidating structural features such as disulfide-bond, post translational modifications, protein-protein interaction, antigen epitopes, etc.

Molecular modelling helped by experimental data

Much experimental data can aid the structure prediction process. Some of these are:
• Disulphide bonds, which provide tight restraints on the location of cysteines in space
• Spectroscopic data and secondary structure prediction, which can give you and idea of the secondary structure content of your protein
• Site directed mutagenesis studies, which can give insights as to residues involved in active or binding sites

• Knowledge of proteolytic cleavage sites, post-translational modifications, such as phosphorylation or glycosylation can suggest residues that must be accessible

Remember to keep all of the available data in mind when doing predictive work. Always ask yourself whether a prediction agrees with the results of experiments. If not, then it may be necessary to modify what you've done.

Protein – protein interactions

Although proteins are actively involved in various biological activities, they must interact with other molecules to fulfil their roles. For instance, enzymes, receptors, and transcription factors have to bind their substrates, ligands and target DNA elements respectively, to execute their functions. Thus the identification of binding partners is crucial to understanding the function of a protein.
Clues to the function of an unknown protein can be obtained by investigating its interaction with other proteins whose functions are already known. Thus, if the function of one protein is known, then the function of its binding partner is likely to be related. This concept has been termed "guilt by association"

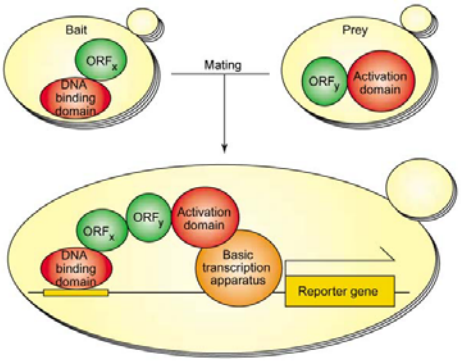
Several methods exist for the identification of protein-protein interactions: yeast two-hybrid, affinity chromatography combined with MS, protein arrays and in silico methods.
A major difference between the yeast two-hybrid strategy and the affinity chromatography approach is that the yeast two-hybrid system is only used to detect pairwise interactions between proteins. In contrast, an affinity chromatography approach allows subunits consisting of many proteins to be isolated and identified.
Both yeast two-hybrid and affinity chromatography/mass spectrometry techniques aim to detect physical binding between proteins, whereas *in silico* methods seek to predict functional associations. However, in many cases such functional associations do take the form of physical binding.

Each technique produces a unique distribution of interactions with respect to functional categories of interacting proteins. These differences in coverage suggest that the methods have specific strengths and weaknesses. The data sets based on purified complexes, for example, predict relatively few interactions for proteins involved in transport and sensing (possibly because these are enriched in transmembrane proteins, which are more difficult to purify). Similarly, interactions detected by the yeast two-hybrid technology largely fail to cover certain categories; for example, proteins involved in translation are found comparatively less often than by other methods. For these reasons it is not surprising that some interactions that are found in one system are missed in another and vice versa. Therefore, only a combination of these approaches will eventually lead to the discovery of all protein-protein interactions in a given cell or organism. Currently almost 100,000 interactions between yeast proteins are available from the different high-throughput methods (the exact number depends on filtering criteria). Of these, only a surprisingly small number is supported by more than one method. There are three possible explanations for this: the methods may not have reached saturation; many of the methods may produce a significant fraction of false positives; and some methods may have difficulties for certain types of interactions, resulting in complementarities between the methods.

Yeast two-hybrid

The yeast two-hybrid takes advantage of the finding that many eukaryotic transcription factors can be divided into two functionally distinct domains that mediate DNA binding and transcriptional activation. In the classical yeast two-hybrid approach, a "bait" is constructed by

fusing a protein X to the DNA-binding domain (BD) derived from a transcription factor and a "prey" is constructed by fusing a protein Y to the activation domain (AD) of a transcription factor. The bait and prey fusions are expressed in yeast, either they are co-expressed in the same yeast or the bait and prey are brought in the same yeast via mating. In yeast, the interaction of proteins X and Y leads to the reconstitution of a functional transcription factor. Reconstitution of the transcription factor is measured by assaying the activity of reporter genes. Commonly, the *lacZ* gene encoding bacterial β-galactosidase is used as reporter gene as β-galactosidase enzyme activity can be easily measured using a colorimetric assay. The plasmids encoding the fusion proteins contain auxotrophic markers (for example *HIS3* and *LEU2*) that can be selected for by growing the yeast on plates with medium lacking histidine or leucine.



Currently, two approaches are being used to generate comprehensive protein interaction maps. In the so called "matrix approach" or "array approach" (Fig. a), a set of open reading frames (ORFs) is amplified using the polymerase chain reaction (PCR), cloned as both bait and prey constructs (*i.e.* as fusion to a BD and as a fusion to an AD), and then introduced into isogenic reporter strains of opposite mating type. The reporter strain expressing a BD fusion is then mated with an array of yeast clones each expressing a different AD fusion. Practically, this task is carried out by robots which transfer aliquots from a lawn of cells expressing one BD fusion to arrays of cells each expressing a different AD fusion. This procedure is repeated for each strain expressing a BD fusion, until all BD fusions have been mated with the entire AD array. Positive interactions are selected through the ability of diploid yeast colonies containing an interacting fusion pair to grow on selective media. In order to sort out the false positives arising from such approach, the experiments are performed in duplicate and only interactors found in both experiments are considered to be true positives. An advantage of the matrix approach is that it rapidly becomes clear which locations produce false positive interactions, providing reassurance that the system is working properly; if a particular AD fusion in the array interacts with all BD fusions, it most likely represents a false positive and should thus be discarded. On the other hand, the matrix approach also has certain disadvantages. The use of full-length ORFs as BD and AD fusions may prevent the identification of certain interactions due to problems such as proper folding of full-length proteins, degradation or toxicity.

A second approach in genome-wide yeast two-hybrid analysis is the so called "exhaustive library screening approach", in which BD fusions are screened against complex libraries containing AD fusions of full-length ORFs or ORF fragments (Fig. b). As opposed to the matrix approach, this method does not separate the different AD fusions on an array. Instead, the library is divided into pools, and each yeast strain expressing a BD fusion is mated with a library pool. Then, diploid cells containing an interacting protein pair are selected. The library screening approach is more sensitive than the matrix approach since it uses not only full-length ORFs, but also random fragments of ORFs. Often, a protein-protein interaction can be detected using fragments of the proteins in question, but not the full-length proteins. For instance, a protein may not fold properly

when expressed in yeast, or it may become degraded. The use of fragments often overcomes these problems. On the other hand, library screens are much more time consuming and expensive than matrix screens since they require the analysis of larger numbers of clones. In addition, the library plasmids encoding AD fusions have to be isolated and sequenced from all selected diploids in order to identify the interacting proteins.

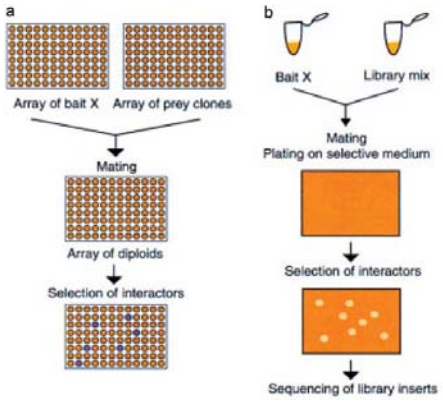


Figure: a) High-throughput yeast two-hybrid using the matrix approach. A matrix (or array) of prey clones is created by dispensing one yeast clone expressing a given AD-Y fusion protein into each well of a multiwell plate. Using a robot, the array of prey clones is then transferred to a multiwell plate containing yeast that express one BD-X fusion and prey and bait clones are allowed to mate. Those diploids where BD-X and a particular AD-Y interact are selected based on expression of a reporter gene, such as β-galactosidase (producing a blue colour). **b)** In the exhaustive library screening approach one BD-fused bait X is screened against an entire library and positives are selected based on their ability to grow on selection plates. As opposed to the matrix approach, where each prey can be identified by its position in the array, diploids that have survived selection in the library screening need to be picked up and the library plasmids encoding the interacting prey have to be isolated and sequenced in order to identify the interacting protein. Libraries can be made either from random genomic or cDNA fragments or from full-length ORFs that are cloned separately and then pooled.
Picture taken from: http://ppi.fli-leibniz.de/PPI_PDF_free/auerbach2002.pdf

False negatives

- The ORFs should be cloned into the plasmids in the right reading frame.
- The use of PCR to amplify the yeast ORFs may introduce mutations that abolish interactions.
- Proteins may fold improperly or be degraded.
- Interactions may be dependent on the presence of accessory proteins or other molecules not present in yeast
- Interactions may be dependent on previous activation or post-translation modifications of the proteins that do not take place in yeast.
- The presence of the AD-domain or the BD-domain in the fusion-protein may interfere with establishment of interactions.
- The yeast two-hybrid system is more sensitive than affinity chromatography, but still transient or weak protein interactions may be missed.
- Interactions in the yeast two-hybrid systems have to take place in the nucleus. Proteins that possess hydrophobic transmembrane domains will be unable to reach the nucleus. Libraries may contain fragments encoding partial proteins or protein domains. When these fragments encode the intracellular part of membrane proteins interactions with these proteins still can be found, reducing the amount of false negatives. However, when these fragments encode domains, the interactions found might normally not take place as they are inhibited by the presence of the rest of the molecular structure, resulting in false positives.

False positives

- Some proteins may be inherently susceptible to non-specific binding interactions (i.e. they are “sticky”)
- Transcription factors and other acid proteins may activate the reporter genes by themselves (auto-activation)

Yeast two-hybrid is famous for its high amount of false positives. To remove as much false positives as possible, the stringency of the selection is important. The stringency can be increased by using multiple reporter genes and by repeating the experiment. If interactions are found several times in different screens they have a higher chance of being biologically significant. Yeast two-hybrid interactions are found between fusion proteins in the nucleus of yeast, so their biological relevance is not guaranteed. Therefore the interactions should be verified by other means.

Affinity chromatography and Mass Spectrometry

A large part of the protein-protein interactions identified thus far is based upon yeast two-hybrid approaches. However, yeast two-hybrid can only find direct interactions between two proteins. Thus, the two-hybrid approach is not useful for the identification of all members from a larger protein complex. A suitable approach to analyse protein complexes is by combining affinity chromatography with mass spectrometry.

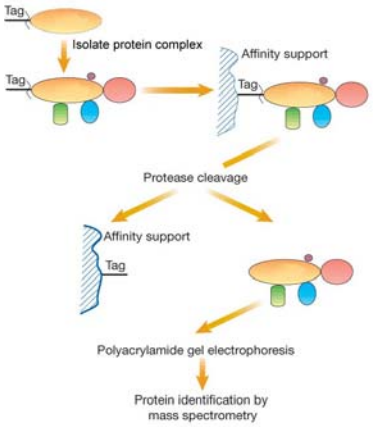


Figure: Affinity chromatography combined with mass spectrometry to isolate protein complexes and identify interacting proteins. The protein of interest (protein X) is expressed as a fusion protein with a cleavable affinity tag to identify interacting proteins. The affinity tag is able to bind to agarose beads. A protease (e.g. thrombin) is used to cleave between the tag and protein X, which results in elution of all proteins that are specifically bound to protein X. The eluted proteins are resolved by one- or two-dimensional gel electrophoresis. The bands or spots corresponding to proteins specifically bound to the tagged proteins (compared to the control sample where only the tag is used for complex isolation) are excised and analysed by mass spectrometry.
http://www.nature.com/nature/journal/v405/n6788/images/405837ae.2.jpg

Affinity chromatography is a technique in which affinity tags are used to chemically immobilize a ligand, such as a protein, to a matrix on a column. Each affinity tag used for purification of fused proteins has its unique features and often tags or domains are used in concert as extra purification steps.
Glutathione *S*-transferase binds tightly to glutathione-agarose and is eluted with glutathione, allowing very high levels of purification in one step; however, it is known to dimerise and this may interfere with protein function. His6 binds immobilized metal-ion columns and is eluted with imidazole; this tag is very simple, but it results in more modest purification than some other tags.

Calmodulin-binding peptide binds calmodulin-agarose columns and is eluted with EGTA; purification is efficient, but proteins with required divalent metal ions may be affected.
An extremely useful variation of these commonly used tags uses very-high-affinity purification tags coupled with protease cleavage sites to remove the purification tag and simultaneously elute the proteins. An example is the generation of fusion proteins in which the target protein is fused to a TAP-tag. TAP stands for Tandem Affinity Purification. The TAP-tag consists of a tobacco etch virus (TEV) protease site linker and a protein A domain. The protein A domain binds tightly to immunoglobulin-g columns, following which the target protein is released using treatment with the highly specific TEV protease. This results in extremely efficient purification of proteins. Additional fusion tags used for other purposes include maltose-binding protein, which increases the solubility of fused proteins; epitope tags such as haemagglutinin, myc and FLAG tags, for immunoprecipitation and detection.

The success rate with the TAP-tag approach in yeast was quite high. This is an encouraging observation. There are however some problems with this method as well. Most importantly, the modified gene (with its TAP tag) needs to be expressed and folded correctly. Additional problems with this method are:

<u>False negatives</u>

- The homologous or heterologous fusion protein that is introduced into the host cells (yeast, *E. coli*) must be localised properly. Failure of the protein to behave in its native condition could explain why some previously known interactions were not observed.
- The affinity tag must not interfere with the function of the bait protein. In some cases, the tags are large (e.g. 20 kD) relative to the average size of a protein (about 50 kD).
- Transient protein interactions may be missed
- Some protein complexes require highly specific physiological conditions in which to form and thus may be missed.
- There may be bias against hydrophobic proteins (which are more difficult to purify than soluble proteins) and low molecular-weight proteins below 15 kD

<u>False positives</u>

- Some proteins may be inherently susceptible to non-specific binding interactions (i.e. they are “sticky”)
- Proteins that are (partly) denatured may non-specifically bind to other proteins.

For biochemical analysis of proteins, their expression in a homologous system is ideal, because the proteins are in their natural environment, are subject to native modifications and can interact wit their natural partners. This has been possible for proteins from yeast and from bacteria such as *E. coli*, but heterologous expression is usually used for proteins from other organisms. In most cases expression is attempted in *E. coli*. However, the *E. coli* bacterium is a prokaryote, so it might miss cofactors, folding mechanisms or the proper posttranslational modifications necessary to produce an active protein. An alternative is the use of (eukaryotic) insect cell, which are able to do modifications that are usually similar to the ones in mammalian cells.

Furthermore, an endogenous promoter will often maintain the stoichiometry of interacting proteins and minimize artificial over-expression, whereas inducible expression allows the purification of proteins that are normally not expressed under laboratory conditions, and can be used in species where genomic integration is not feasible, such as the human.

In silico predictions of protein-protein interactions

Protein-protein interactions are not limited to direct physical binding. Proteins may also interact indirectly – by sharing a substrate in a metabolic pathway, by regulating each other transcriptionally, or by participating in larger multi-protein assemblies. From the Transcriptomics module we saw that it is possible to use gene co-expression to find functional associations between genes and their gene products. In addition, now that multiple genomes are sequenced, it is possible to use genome sequence information to screen whole genomes for the prediction of functional protein-protein associations. These kinds of functional associations can be predicted from genomic associations between the genes that encode the proteins: groups of genes that are functionally related tend to have similar expression profiles, tend to show similar species coverage (phylogenetic co-occurrence profile), are often located in close proximity on the genome (conserved gene neighbourhood; only in prokaryotes), and tend to be involved in gene-fusion events (Rosetta stone approach). The STRING database contains predictions of functional association from these methods. The ArrayProspector contains predictions of functional association based on co-expression of the gene on microarray experiments.

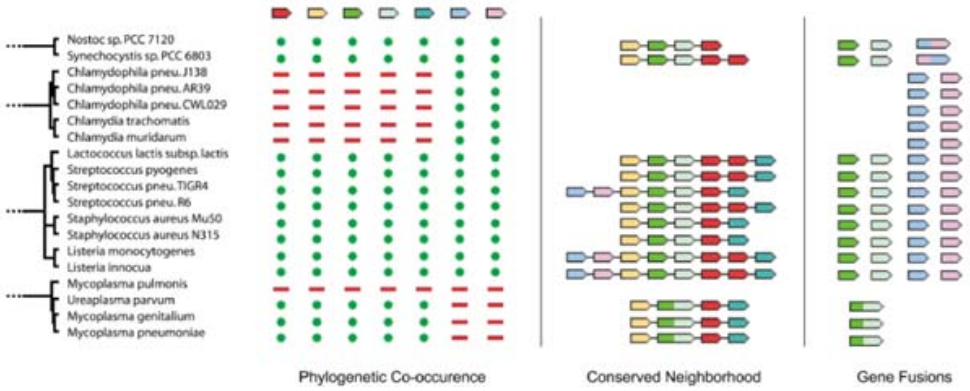


Figure: Three types of genomic evidence for a functional association between proteins. Left: phylogentic co-occurrence. Middle: Conserved gene neighbourhood. Right: Gene fusions.
Picture taken from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC165481/bin/gkg034f1.jpg

Thus are three types of evidence for a functional association between proteins that can be found by looking at the genome sequence.

- Gene neighbourhood
- Gene fusion / Rosetta stone approach
- Gene co-occurrence / phylogenetic co-occurrence profiles

Each of these methods is statistical in nature. You can compute a probability associated with each pair-wise relationship. For example, you can compute the probability of observing a certain similarity between two phylogenetic profiles, by measuring how often this degree of similarity would be found if you randomly assign gene families to organisms.

The benefits of *in silico* techniques to find protein-protein interactions are that they are fast and inexpensive and that coverage expands as more genomes are sequenced. However, *in silico* techniques will only give predictions of interactions. To know for certain if an interaction takes place *in vivo* verification of the interaction by other means has to take place.

Gene co-expression

Remember that correlations between expression profiles do not necessarily imply co-regulation, and co-regulation does not always indicate functional interaction. By taking evolution into account, it is possible to increase the reliability of co-expression data for function prediction. The encoded proteins of conserved co-expressed gene pairs are highly likely to be part of the same pathway not only between pairs of orthologues in two species, but also between parallel duplicated gene pairs in one species (paralogues). In other words, if a pair of genes is co-expressed, often it is also the case that the paralogues or homologues of this gene pair are co-expressed, although not necessarily in the same way. The conserved co-expression indicates that it is highly likely that the gene products are part of the same pathway.

Orthologues: genes in different species that evolved from a common ancestor by speciation.
Paralogues: genes related by a duplication in the genome.

Normally orthologues retain the same function in the course of evolution. Orthologues that are widely distributed (in a wide range of species) perform the core biological functions shared by all organisms. They do not include the proteins unique to the biology of a particular organism. Paralogues is a different story. The most likely fate of a duplicated gene is gene loss, because it has been observed that redundant genes are removed from genomes. If paralogues coexist during evolution they therefore have divergent functionalities or complementary functions. Either their proteins become involved in novel biochemical pathways or the paralogues have distinct expression domains. The latter means that the original function e.g. was expressed in the roots and leaves of a plant but that after duplication one copy is expressed exclusively in the roots while the other copy becomes responsible for expression in the leaves, Both copies become complementary because their expression domain is different (can be caused by mutations in the regulatory element).
Paralogues can be identified by a genome self comparison and orthologues can be determined as the best reciprocal hit in a pairwise genome comparison between two distinct genomes (BLAST or Clustal W). The best reciprocal hit means that two proteins, X in proteome A and Y in proteome B, are predicted true orthologues if reciprocal searches of proteome A with Y and proteome B with X each produce the highest scoring match with the other protein. This definition is not full proof in identifying orthologues, but works well when the two species are not too distant phylogentically. At larger phylogenetic distances gene duplications might have occurred with subsequent divergence of gene functions. In such case only a many to many relationship will adequately describe orthologues and detection of the highest similarity will not result in the identification of the complete set of orthologues. These problems were overcome by the definition of clusters of orthologues.

The COG database (http://www.ncbi.nlm.nih.gov/COG/) is a database where by mutually comparing the proteomes of several completely sequences prokaryotes clusters of orthologues were identified. Each COG consists of individual orthologous genes or orthologous groups of paralogues from three of more phylogenetic lineages.

Gene neighbourhood

Two species that have recently diverged from a common ancestor might be expected to share a similar set of genes and also similar chromosomes with the genes positioned in the same order along the chromosomes. Over evolutionary time, the sequence of each pair of genes will slowly diverge as the species diverge and other changes such as duplication and gene loss change the gene content. The number of rearrangements in a given period of evolutionary history may vary

significantly from one organism to the next. But the general assumption is that genes falling in the same functional category or belonging to the same pathway have been shown to cluster together on the chromosomes of both organisms, i.e. their gene context tends to be conserved.

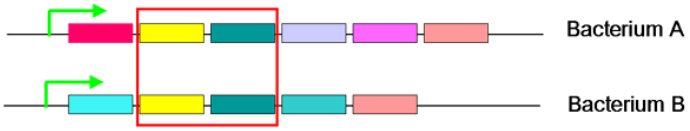


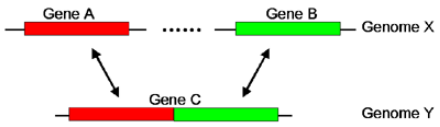
Figure: In prokaryotes, the presence of two genes together in an operon can be used as evidence for a functional association between the products of these genes. In this example the yellow and green genes, indicated by the square, are in bacterium A as well as in bacterium B in each other's neighbourhood.

Put the other way around: Genomic associations between genes reflect what selection regards as functionally interacting proteins. Thus, the ordering of related genes in genome sequences can be used as a basis for the identification of interacting proteins.
The strength of the genomic association correlates with the strength of the functional association. This method is only applicable to prokaryotes, because in prokaryotic genomes genes are grouped in operons resulting in co-regulation of these genes. Genes that frequently occur in the same operon in a diverse set of species are more likely to interact than genes that occur together in an operon in only two species.

STRING is an example of a web-service where you can search for local gene context in prokaryotic genomes to find functionally related proteins: http://www.bork.embl-heidelberg.de/STRING/

Gene fusion

The gene fusion approach is based on the Rosetta Stone. Some pairs of interacting proteins are encoded by two distinct genes in one genome that have fused into a single gene in another genome. These proteins can be defined as composite proteins and their individual domains as component proteins. The underlying assumption is that if a composite protein is uniquely similar to two component proteins in another species, which may not necessarily be encoded by neighbouring genes, the component proteins are likely to interact. This domain function analysis has been called the Rosetta Stone approach. The Rosetta Stone approach makes the prediction that protein pairs generated from gene fusions have related biological functions. For example, they may function in the same protein complex, pathway, or biological process.



An organism may want to fuse two genes, encoding biologically related proteins, to overcome the energy costs necessary to increase the concentrations of the two proteins in a local environment to an effective concentration so that they can interact.
Identification of gene-fusion events based on sequence comparison is possible because there must be selective pressure for certain genes to be fused over the course of evolution. In this way, the approach can also predict possible protein-protein interactions.

Gene co-occurrence / phylogenetic profiles

Prediction of protein interactions or functional relatedness based on gene co-occurrence is based on the assumption that proteins that function together in a biochemical pathway should evolve in a correlated fashion. By cataloguing all the proteins that are expressed from several genomes, you can determine the pattern of the presence and absence or protein families across organisms. These patterns are called phylogenetic profiles.

If proteins A and B have a related function they should be found together in a large proportion of genomes, whereas if proteins A and B do not have a related function they will be found to have a random association in genomes.

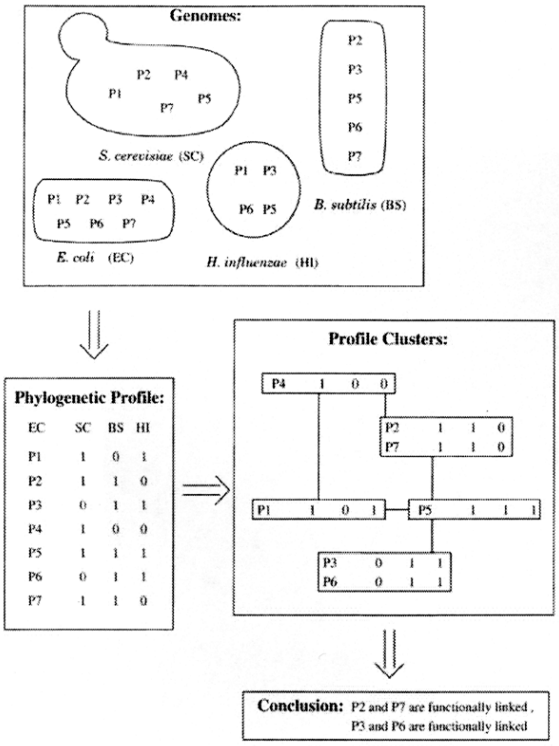


Figure: Interacting proteins have a tendency to be either present or absent together from fully sequenced genomes, that is, to have a similar 'phylogenetic profile'. Picture adapted from Pellegrini, 2001.

Homology modelling of 3D interactions

Genome-scale interaction discovery approaches, like the two-hybrid system or affinity purification methods, have uncovered large parts of the protein-protein interaction network. However these methods only find out what interacts with what, but not how they interact.

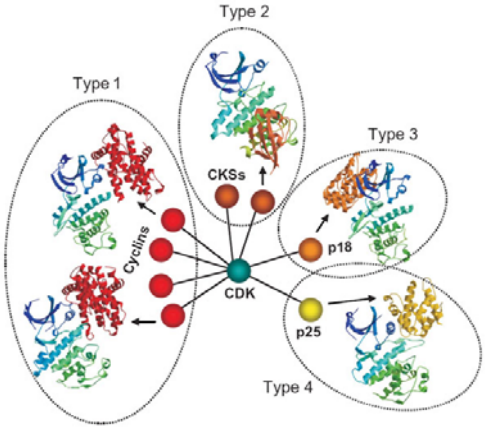


Figure: CDK interacts with 8 different proteins. However, one can distinguish only 4 different "interaction types". *Picture taken from Aloy and Russell, 2004*

These details can sometimes come from similarities to interacting proteins of known three-dimensional structure. Based on protein homology at the 3D level it was found out that there are, roughly, only 1000 different protein folds in nature. Nature is remarkably restricted to these few types of protein folds to perform a great diversity of functions, even though other folds may also be possible. Among the known 3D structures are also many protein-protein interactions and protein complexes. Like homology modelling for single proteins it is also possible to do homology modelling for protein-protein interactions. Because when the sequence of the proteins is homologous, their interaction interface should also be homologous. Actually, when the protein pair you are interested in shows >25% sequence identity to a protein pair with a known 3D structure it is in 90% of the cases possible to model the interaction between the two proteins.

Like protein "folds" there are also protein "interaction types". A single interaction type groups together all "pairs" of proteins that interact in a way similar to that observed in the three dimensional structure of a complex. At the moment (end 2004) there are about 1800 interaction types known. It is expected that in nature there exist between roughly 6000 and 10000 interaction types.

Post-translational modifications

In the field of proteomics, the study of different post-translational modifications in proteins has become important, largely because the modifications of proteins cannot be visualized by transcriptomics. Transcriptomics only gives you information about the transcriptional levels of genes.

Post translation modification of proteins plays a very important role in regulating protein activity and sometimes even protein function. The identification of the type of modification and its location can often provide crucial information for understanding the function or regulation of a given protein in biological pathways. So far, more than 200 different modifications have been reported, many of which are known to control signalling pathways and cellular processes. Most likely, all proteins will turn out to be post-translationally modified.

Table: Some common and important post-translational modifications. A more comprehensive list of post-translational modifications and their Δmass values can be found at the Unimod database: *http://www.unimod.org*.

Modification type	Mono-isotopic Δmass(Da)	Average Δmass(Da)	Function and notes
Phosphorylation pTyr, pSer, pThr	+79.66331	+79.9799	Reversible, activation/inactivation of enzyme activity, modulation of molecular interactions, signalling
Acetylation	+42.010565	+42.0367	Protein stability, protection of N terminus, regulation of protein-DNA interactions (histones)
Methylation	+14.015650	+14.0266	Regulation of gene expression
Farnesylaltion Myristoylation Palmitoylation	+204.187801 +210.198366 +238.229666	+204.3511 +210.3446 +238.4088	Fatty acid modification. Cellular localisation and targeting signals, membrane tethering, mediator of protein-protein interactions
Glycosylation N-linked O-linked	 >800 >800		Excreted proteins, cell-cell recognition/signalling O-GlcNAc, reversible, regulatory functions
Glycosylphos-phatidylinositol (GPI) anchor	>1000		Membrane tethering of enzymes and receptors, mainly to outer leaflet of plasma membrane
Hydroxylation	+15.994915	+15.9994	Protein stability and protein-ligand interactions
Sulfation (sTyr)	+79.956815	+80.0632	Modulator of protein-protein and receptor-ligand interactions
Pyroglutamic acid	-17.026549	-17.0365	Protein stability, blocked N-terminus
Ubiquitination Gly-gly	>1000 114.042927	 114.042927	Destruction signal. After tryptic digestion, ubiquitination site is modified with the Gly-Gly dipeptide

In principle, the methods used for protein identification (MS in combination with 2D-PAGE or chromatography) are also applicable for the analysis of post-translational modifications. However, the detection and identification of proteins that are post-translationally modified is significantly more complex than simple protein identification. Localizing the modification on the protein adds another layer of complexity.

- While proteins can be identified by the sequence or the mass spectrum of a single peptide using MS or MS/MS, the identification of post-translational modifications requires the isolation and analysis of the specific peptide that contains the modified residue(s).
- The detection of modified peptides requires a method that has high sensitivity or some form of fractionation must be applied to enrich the sample with the modified protein, because:
  - usually only a small fraction of a given protein is modified.
  - a protein may have modifications at multiple sites (giving rise to differentially modified forms of the same protein).
  - the modification may alter the cleavage pattern of the protein, because some cleavage sites may not be accessible for the protease.
  - modified peptides often do not ionise as well as unmodified peptides.
- The bond between the modification and the peptide is frequently labile. It may therefore be difficult to find conditions that maintain the peptide in its modified state during sample work-up and ionisation.

Many strategies have been developed to analyse protein modifications, but most of them focus on only a specific type of modification, such as protein phosphorylation. Many of the general approaches and specific methods discussed in the context of protein phosphorylation are however directly or with minor adaptations applicable to the analysis of different types of modifications. Phosphorylation is regulated by two counteracting enzyme systems, kinases and phosphatases that catalyse protein phosphorylation and dephosphorylation, respectively. There are assumed to be hundreds of protein kinases/phosphatases differing in their substrate specificities, kinetic properties, tissue distribution, and association with regulatory pathways. The most common type of protein phosphorylation involves the formation of phosphate ester bonds with the hydroxyl side chains of serine, threonine, and tyrosine. It is thought that at least 30% of all proteins contain covalently bound phosphate, which implies that in fact most cellular processes are regulated by phosphorylation.
Most biochemical techniques can only say whether a protein is modified, but not where it is modified. It is important to know where a protein is modified, because the same modification at different sites may have different consequences. The use of mass spectrometry methods to analyse peptides now offers the best method to characterise protein modifications.

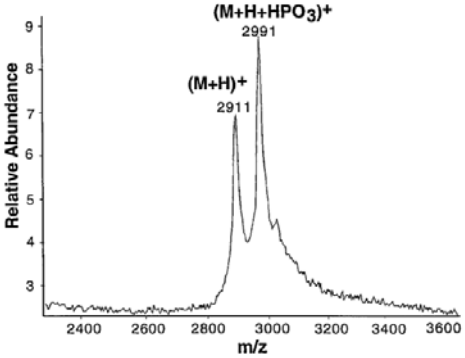


Figure: MS spectrum of peptide "YNSTGSPPGKPPSTODEHINSGDTPAVR" in a phosphorylated and unphosphorylated form. The peaks in the spectrum differ in mass by 80 Da, corresponding to the addition of HPO_3 to the hydroxyl group on the side chain of Ser. The peptide contains 4 serine residues, thus the site of phosphorylation cannot be assigned based on the molecular mass data. Picture from Aebersold and Goodlett, 2001.

Typically in phosphorylation studies, the amino acid sequence of the protein investigated is known. Therefore, phosphopeptides derived from the protein can in principle, be detected by a net mass differential of 80 Da that occurs when phosphate is added to serine, threonine or tyrosine ($H_3PO_4 - H_2O$ = ((3*1+31+4*16)-(2*1+16)) = 80). Thus a peptide mass map of the proteolytically fragmented phosphoprotein can potentially identify the phosphorylated peptide by comparison to the known protein sequence. However, with this method you still do not know which residue is phosphorylated.
It takes only about two seconds to acquire an MS/MS spectrum, so in 30 minutes roughly 800 MS/MS spectra can be collected. Therefore, different methods were developed to characterise the sites of phosphorylation in a protein. The sites of phosphorylation can be determined for a single isolated phosphorylated protein or for proteins present in a complex mixture. The actual determination of the phosphorylated residue(s) generally consists of the following steps:

1. Purification or enrichment of phospho-protein(s)
    - 2D gel electrophoresis + in vivo or in vitro labelling of protein(s) with 32P or detection of phospho-proteins by Western blotting using pTyr, pThr and/or pSer specific antibodies.
    - Immunoprecipitation using pTyr, pThr and/or pSer specific antibodies

    - IMAC: immobilised metal affinity chromatography.
2. Enzymatic or chemical cleavage of the (phospho-)proteins into peptides
3. Instead of isolating phosphorylated proteins (step 1), it is also possible to start with cleaving all proteins from a cell lysate and then enrich the sample with phosphorylated peptides using for example IMAC.
4. MS/MS
    - Identify phospho-peptide(s) by precursor ion scanning or neutral loss scanning
    - Determine location of phosphorylated residue(s) (product ion scanning)

MS/MS and the detection of post-translational modifications

Precursor ion scanning

In the syllabus we read that there are many different types of MS ionisation sources and mass analysers, each with their specific advantages and disadvantages. Depending on the application, different MS instruments can be used. For MS/MS analysis several key factors are important. These are:

- selectivity
- accuracy and resolution
- measurement speed

Below we will explain MS/MS based on a triple quadrupole MS/MS apparatus. The triple quad has the advantage of high selectivity, but the mass that is selected is not measured very accurately. Moreover, the sensitivity of a triple quad MS is modest and the resolution is limited. In contrast, a quadrupole-TOF (Q-TOF) instrument can measure the mass of a molecule with high accuracy and high resolution, it has a good sensitivity and a very good m/z range. The best instrument for high accuracy and resolution is a FT-MS (also known as FT-ICR MS). This instrument uses a superconducting magnet to trap the ions and measures the orbiting frequency (time-dependent) of the ions. The time-dependent signal can be Fourier-transformed in order to reveal the m/z values. The FT-MS is expensive, needs a real expert to operate, but provides the scientist with the best MS spectra possible. The third point, the measurement speed has to do with the price per sample. The longer it takes to analyse a sample the higher the costs. Ion trap-MS and MALDI-TOF-TOF measurements do not take much time, when performing MS/MS fragmentation. The Quadrupole-tof systems are quite slow in this respect and hence much more expensive to use. Therefore, for each experiment, a choice has to be made between accuracy and selectivity compared to the speed of the measurements to be done.
Precursor ion scanning uses ESI-MS/MS (in principle in negative ion mode) to identify phospho-peptides. In this method, peptides are negatively charged by electrospray ionisation (ESI) and their mass is measured in the first quadrupole mass spectrometer (Quad 1) that is allowed to scan ions over a wide range of m/z ratios. Every peptide is then broken into pieces in the collision cell (Quad 2) and the resulting fragment ions pass through the third quadrupole (Quad 3). Quad 3 is maintained at a fixed m/z ratio, so only specific ions are allowed to pass. In the case of phospho-peptide analysis this is usually 79 m/z (i.e. loss of the phosphate group, PO_3^-). Consequently, a filtered mass spectrum is obtained that only shows ions that produce a 79 m/z ion upon fragmentation, i.e. only phospho-peptides. A selected phospho-peptide is usually called precursor ion or parent ion. The 79 m/z ion used to identify the precursor ion is often called the daughter ion.

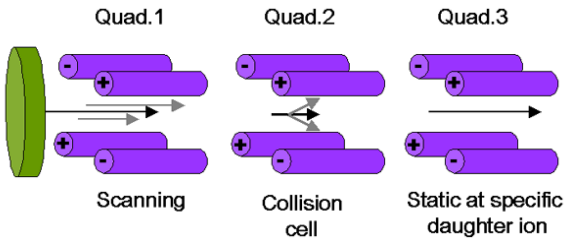


Figure: Ion precursor scanning for the detection of phospho-peptides. The first quadrupole is scanning the complete m/z-value spectrum. Each ion is then fragmented in a collision cell (quadrupole 2) before it goes to quadrupole 3. Quadrupole 3 is selecting those ions that have a specific m/z value. In this way phospho-peptides can be detected from a mixture of peptides.

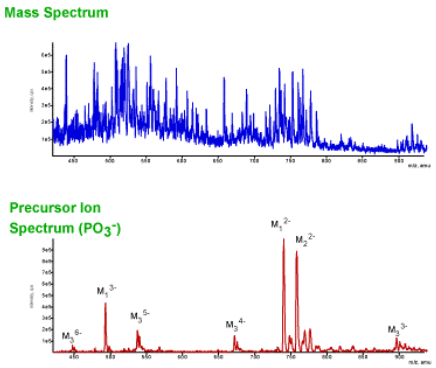


Figure: Detection of phosphorylated peptides from a peptide mixture. Mass spectrum derived from ESI-MS/MS with the quadrupole in the 'scan' mode. Precursor ion spectrum as a result from ESI-MS/MS with the first quadrupole in the scanning mode and the second quadrupole in the static mode at a m/z value of 79. In the mass spectrum three different precursor ions can be found (M1 – M3). Picture from http://www.mann.embl-heidelberg.de/GroupPages/PageLink/activities/Phosphorylation3.html.

In the case where the protein sequence is known, the peptide molecular weights obtained are sufficient to identify the sequences that are phosphorylated, although the specific location of the phosphorylated residue is not known. Unfortunately the negative ion mode MS/MS spectra generally produce insufficient fragment ions for sequence elucidation. Therefore, once phospho-peptides are identified, the complete peptide mixture can be sprayed again under acidic conditions (positive ion mode) and selected phospho-peptides can be sequenced by conventional MS/MS methods to determine the location of the phosphorylated residues. As the sample has to be injected twice, a higher quantity of the sample is necessary for this procedure.

Alternative MS phosphate scanning methods

Precursor ion scanning can also be used to identify phospho-tyrosine-containing peptides by scanning in positive ion mode for the appearance of the phospho-tyrosine immonium ions (m/z = 216.043) (see Exercise 2). Because phospho-tyrosine containing peptides are identified in positive ion mode, MS/MS sequencing of the parent ion can be performed simultaneously. Owing to the inherent lability of the alkyl phopshoesters of serine and threonine, the corresponding immonium ions are formed with a very low yield, so that the sensitivity of such experiments is low. Therefore, phospho-serine and phospho-threonine containing peptides are usually detected using other thechniques.

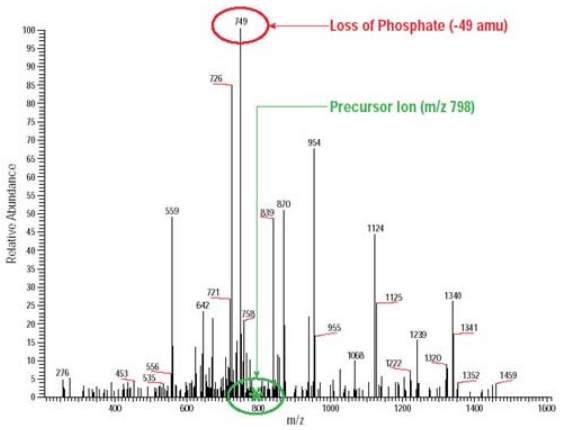


Figure: Due to ß-elimination, phosphoserine and phosphothreonine residues may lose a phosphate group. In this example, the loss of 49 corresponds to the difference between the precursor ion mass of a doubly charged phosphorylated peptide (2*49=98) and its dephosphorylated variant, as can be observed in the MS spectrum.

Phospho-peptides can also be identified using "neutral loss scanning". This method uses electrospray ionisation (ESI) in the positive ion mode to induce the (neutral) loss of H_3PO_4 from phospho-serine or phospho-threonine-containing peptides due to a process known as ß-elimination. Only phospho-threonine and phospho-serine may undergo neutral loss of H_3PO_4 (3*1+31+4*16=98 Da) by ß-elimination. Phospho-tyrosine does not show neutral loss of H_3PO_4. Phospho-tyrosines will not easily lose their phosphate group due to stability of the ß-protons in the benzene ring, which minimise ß-elimination. The value of the neutral loss depends on the charge state of the precursor ion.

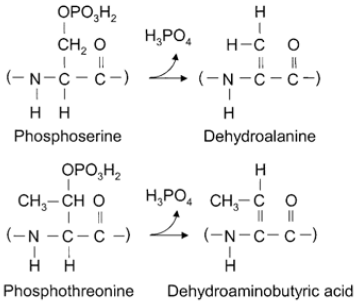


Figure: ß-elimination of phospho-serine and phospho-threonine. During collision-induced-dissociation to fragment peptide ions, phospho-serine and phospho-threonine undergo ß-elimination to dehydroalanine and dehydroaminobutryic acid respectively. Such reactions also occur in solution under alkaline conditions.
First and second picture taken from: http://216.74.34.73/m/37/Article/Thermo_Ar304_Fig2_graph.jpg

With neutral loss scanning, ions are scanned in parallel in two mass analysers (after quadrupole 1 and quadrupole 3) with an offset in m/z value (see figure below). The offset in the m/z values measured by the two mass analysers at any given time is constant and corresponds to the loss of phosphate during fragmentation of the parent ion in the collision cell (quad 2). Thus, the first mass analyser measures the peptide with the phosphate group (quad 1), and the second mass analyser measures the peptide without phopshate group (quad 3). Depending on the charge of the phospho-peptide ion, $[M+H]^+$, $[M + 2H]^{2+}$or $[M+3H]^{3+}$, the quadrupoles require an offset value of 98 m/z, 49 m/z or 24.5 m/z, respectively. In this way only ions that lose 98 Da upon fragmentation will reach the detector of the second mass analyser after the third quadrupole. Thus the third quadrupole acts as a mass filter looking for ions one neutral side chain less in mass. In the final mass spectrum you will only see those parent ions that proved that they possessed a neutral side chain of 98 Da. This is different from the MS spectrum shown above!! The parent ions that show the loss of 98 Da in the second mass analyser can be selected for immediate sequencing.

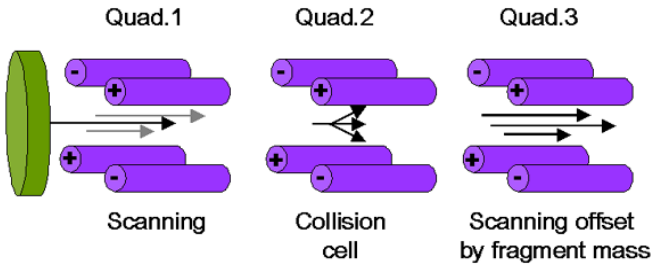


Figure: Neutral loss scanning for the detection of phospho-peptides. Each ion is fragmented in a collision cell before it goes to the third quadrupole. The third quadrupole is also scanning the complete spectrum but with an offset value corresponding to the loss of a phosphate group (M-H_3PO_4. In this way only peptides that show the neutral loss of a phosphate group are detected from a mixture of peptides.

The neutral loss scanning method is not the most popular method among Mass spectrometrists, due to problems with false negative results. The measurement should be quite accurate otherwise two molecules derived from the same peptide that differ in a mass close to the mass of a neutral loss of H_3PO_4 are also detected as a positive. An advantage of the method is that it is carried out in positive ion mode. In order to identify the location of the phosphorylated peptide, MS/MS spectra can be acquired in the same experiment using the detection of a neutral loss of phosphate as a trigger to initiate MS/MS.

Product ion scanning

Often the information obtained by precursor ion scanning, neutral ion scanning or in-source fragmentation scanning is not sufficient for identification of the phosphorylated residue in a phospho-peptide. In fact, these methods are designed to distinguish phospho-peptides from non-phospho-peptides and to potentially indicate the phospho-peptide mass rather than to produce sequence information. Consequently these methods can only successfully identify a phosphorylated residue if the peptide sequence is known and contains only one residue that might be phosphorylated. If there is more than one possible amino acid residue present in the peptide that can be phosphorylated, then it is necessary to acquire MS/MS spectra. In the MS/MS spectrum phosphorylated residues can be identified due to their increase in mass. The masses of serine, threonine and tyrosine residues are 87, 101 and 163 Da, respectively. Phosphorylation adds a phosphogroup of 80 Da to these residues resulting in masses of 167, 181 and 223 Da for serine, threonine and tyrosine respectively.

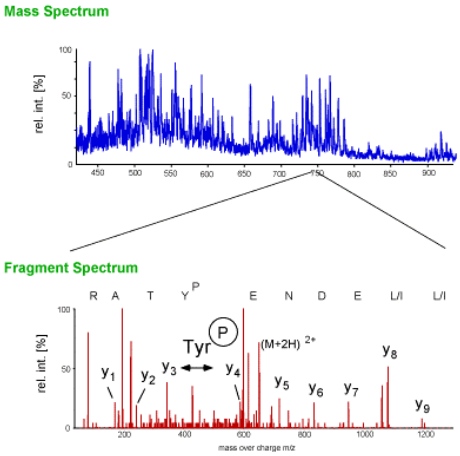


Figure: Mass spectrum of a peptide mixture containing phospho-peptides and MS/MS spectrum of a selected phopo-peptide. Picture from http://www.mann.embl-heidelberg.de/GroupPages/PageLink/activities/Phosphorylation3.html.

IMAC

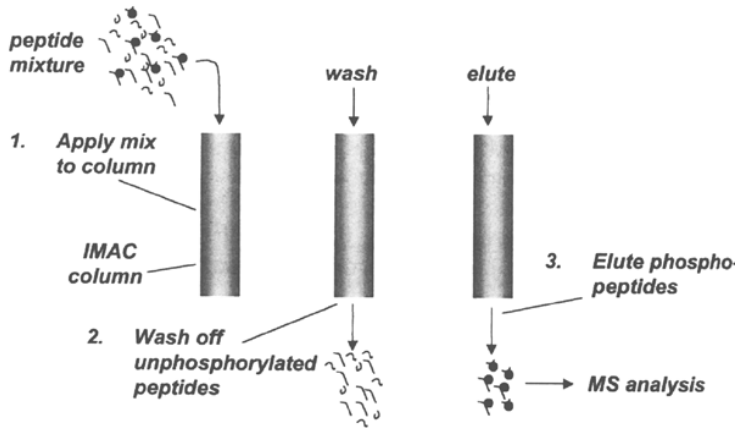


Figure: Use of immobilised metal affinity chromatography (IMAC) to isolate phosphopeptides from a protein digest. Picture taken from Aloy and Russel (2004)

The IMAC procedure uses a column with positively charged metal ions (like Fe^{3+} or Ga^{3+}) that will bind phosphorylated proteins with a higher affinity than non-phosphorylated proteins due to the negative charge of the phosphate group.

Literature

Proteomics

- Patterson and Aebersold **2003** Proteomics: the first decade and beyond. Nature Genetics Supplement 33:311-323.
- Zhu et al. **2003** Proteomics. Annual Reviews in Biochemistry 72: 783-812.
- Pevsner, J. (2009). Bioinformatics and Funcional Genomics. John Wiley and Sons Ltd. ISBN 9780470085851

MS

- Aebersold and Mann **2003** Mass spectrometry-based proteomics. Nature 422: 198-207.
- Aebersold and Goodlett **2001** Mass spectrometry in proteomics. Chemical Reviews 101(2): 269-295.
- Mann and Jensen **2003** Proteomic analysis of post-translational modifications. Nature Biotechnology 21:255-261.
- Steen et al. **2003** Phosphotyrosine mapping in Bcr/Abl oncoprotein using phosphotyrosine-specific immonium ion scanning. Molecular and cellular proteomics 2(3): 138-145.

Protein microarrays

- Espina et al., **2003** Protein microarrays: molecular profiling technologies for clinical specimens. Proteomics 3: 2091-2100.
- Templin et al., **2003** Protein microarrays: promising tools for proteomic research. Proteomics 3: 2155-2166.
- Zhu and Snyder, **2003** Protein chip technology. Current Opinion in Chemical Biology 7: 55-63.

Protein-protein interactions

- Aloy and Russell **2004** Ten thousand interactions for the molecular biologist. Nature Biotechnology 22(10): 1317-1321.
- Auerbach et al. **2002** The post-genomic era of interactive proteomics: facts and perspectives. Proteomics 2: 611-623.
- Gavin et al. **2002** Functional organisation of the yeast proteome by systematic analysis of protein complexes. Nature 415: 141-147.
- Ito et al. **2001** Exploring the protein interactome using comprehensive two-hybrid projects. Trends in Biotechnology 19(10): S23-S27.
- Pellegrini et al. **2001** Computational method to assing microbial genes to pathways. Journal of Cellular Biochemistry 37(Suppl):106-109.

Pictures from internet are retrieved on 24-04-2012

Useful links

Protein sequence databases

- **GenPept (NCBI)** – translated ORFs from DNA Data Bank of Japan (DDBJ), EMBL, GenBank and the Nucleotide Sequence Database. Contains minimal annotation and redundancy: ftp://ncbi.nlm.nih.gov/genbank/genpept.fsa.Z (Uncompress and use setdb to make a blastable database.)

- **Entrez Protein (NCBI)** – translated ORFs from DDBJ/EMBL/GenBank + sequences from SwissProt/PIR/PDB/RefSeq. Contains additional information extracted from curated databases such as SiwwProt and PIR. Database is redundant: http://www.ncbi.nlm.nih.gov/

- **RefSeq** – The Reference Sequence collection aims to provide a comprehensive, integrated, non-redundant set of sequences, including genomic DNA, transcript (RNA), and protein products, for major research organisms. Partly manually curated but most records automatically generated: http://www.ncbi.nlm.nih.gov/RefSeq

- **PIR-PSD** – Protein Information Resource Protein Sequence Database. Comprehensive, non-redundant protein sequence data, organised by superfamily and family, and annotated with functional, structural, bibliographic and genetic data. Integrated with other databases: http://pir.georgetown.edu/

- **Swiss-Prot** – High quality, universal, non-redundant, curated protein sequence database. Highly integrated with other databases. http://www.ebi.ac.uk/swissprot/index.html / http://www.expasy.org/sprot/

- **TrEMBL** – Computer annotated entries derived from translation of all coding sequences in the DDBJ/EMBL/GenBank nucleotide sequence database that are not yet included in Swiss-Prot. Enhancement of information content using InterPro. TrEMBL is non-redundant and follows the Swiss-Prot format and conventions.http://www.expasy.org/sprot/

- **UniProt** – Combination of Swiss-Prot, TrEMBL and PIR-PSD databases into a single resource. The database consists of two parts, fully manually annotated records called Swiss-Prot and a part that is computationally analysed and awaits manual curation called TrEMBL: http://www.uniprot.org

  - **UniParc** – UniProt Archive. Most comprehensive publicly accessible non-redundant protein sequence collection. Sequences from SwissProt, TrEMBL, PIR-PSD, EMBL, Enseml, International protein index (IPI, http://www.ebi.ac.uk/IPI), PDB, RefSeq, FlyBase, WormBase and the patent offices in Europe, US and Japan: http://www.pir.uniprot.org/database/archive.shtml

  - **UniProt-NREF / UniRef** – Comprehensive, non-redundant sequence collection clustered by sequence identity and taxonomy with source attribution. Identical sequences and sub-fragments from the same source organism (species) are presented as a single NREF entry with accession numbers of all the merged UniProt entries, the protein sequence, taxonomy, bibliography, links and close sequence neighbours from the same organism: http://www.expasy.uniprot.org/database/nref.shtml

2D-electrophoresis databases

- **Human 2D-PAGE databases** for proteome analysis in health and disease: http://proteomics.cancer.dk/

- **SWISS-2DPAGE** - Two-dimensional polyacrylamide gel electrophoresis database: http://au.expasy.org/ch2d/

- **2-DE Gel Protein Databases** at Harefield: http://www.harefield.nthames.nhs.uk/nhli/protein

- **The Human Myocardial Two-Dimensional Electrophoresis Protein Database**: http://userpage.chemie.fu-berlin.de/~pleiss/dhzb.html

- Partial List of Web **2D Electrophoretic Gel Databases**: http://www-lecb.ncifcrf.gov/EP/table2Ddatabases.html

Mass Spectrometry

- **Introduction to Mass Spectrometry** http://www.astbury.leeds.ac.uk/Facil/MStut/mstutorial.htm or http://masspec.scripps.edu/information/intro/index.html or http://www.asms.org/whatisms/index.html or http://www.matrixscience.com/help_index.html

- **Expasy** – List of protein identification and characterization tools using MS data: http://www.expasy.org/tools/

- **PeptIdent** – tool that allows the identification of proteins using pI, Mw and peptide mass fingerprinting data. Experimentally measured, user-specified peptide masses are compared with the

theoretical peptides calculated for all proteins in the Swiss-Prot/TrEMBL databases. A species (or group of species) can also be specified for the search: http://www.expasy.org/tools/peptident.html

- **MultiIdent** – tool that allows the identification of proteins using pI, MW, amino acid composition, sequence tag and peptide mass fingerprinting data. One or more species and a Swiss-Prot keyword can also be specified for the search: http://www.expasy.org/tools/multiident/
- **Aldente** – http://www.expasy.org/tools/aldente/
- **ProFound** – Peptide Mapping, enables protein identification from simple protein mixtures: http://prowl.rockefeller.edu/
- **Protein prospector / MS-FIT** – Proteomics tools for mining sequence databases in conjunction with Mass Spectrometry experiments: http://prospector.ucsf.edu/
- **MASCOT** – search engine that uses mass spectrometry data to identify proteins from primary sequence databases: http://www.matrixscience.com/search_form_select.html
- **Mass-Search** – Searching SwissProt or EMBL by protein mass after digestion: http://cbrg.inf.ethz.ch/Server/MassSearch.html
- **MOWSE** – will search the owl protein sequence database with protein fragment information, and return the protein(s) which most likely correspond to your peptide-data: http://srs.hgmp.mrc.ac.uk/cgi-bin/mowse
- **PeptideSearch** – Protein identification by peptide mapping or peptide sequencing: http://www.mann.embl-heidelberg.de/GroupPages/PageLink/peptidesearchpage.html
- **PeptideMapper** – the UMIST protein search tool, which uses Mass Spec. data produced by the digestion of a protein to identify a match to a protein from a database: http://wolf.bms.umist.ac.uk/mapper/
- **PepSea** – Protein identification by peptide mapping or peptide sequencing from Protana, Denmark: http://www.unb.br/cbsp/paginiciais/pepseaseqtag.htm

Protein interaction

- **ArrayProspector** - A web resource of functional associations inferred from microarray expression data: http://www.bork.embl.de/ArrayProspector/
- **STRING** – Search Tool for the Retrieval of Interacting Genes/Proteins: http://www.bork.embl-heidelberg.de/STRING/
- **PathBLAST** – searches the yeast protein-protein interaction network to identify all protein interaction pathways that align with a pathway query: http://www.pathblast.org
- **BIND** – Biomolecular interaction Network database. Database of interactions involving protein, DNA and small molecules. Large list of other links within: http://www.bind.ca
- **DIP** – Database of interacting proteins: http://dip.doe-mbi.ucla.edu
- **GRID** – Database of genetic and physical interactions http://biodata.mshri.on.ca/grid/servlet/Index.
- **Predictome** – visualizing the predicted functional associations among genes and proteins in many different organisms. Associations, or gene links, are created using a variety of techniques, both experimental (yeast two-hybrid, immuno-coprecipitation, correlated expression) and computational (gene fusion, chromosomal proximity, gene co-evolution): http://predictome.bu.edu/

- **BRITE** – Biomolecular Relations in Information Transmission and Expression database of the Kyoto Encyclopaedia of Genes and Genomes: http://www.genome.ad.jp/brite/

Protein structure

- **Delta Mass** - A Database of Protein Post Translational Modifications http://www.abrf.org/index.cfm/dm.home

- **Structure prediction guide**: http://speedy.embl-heidelberg.de/gtsp/index.html

- **Structure Determination of Proteins with NMR Spectroscopy**: http://www.cryst.bbk.ac.uk/PPS2/projects/schirra/html/home.htm

- **Crystallography course**: http://www-structure.llnl.gov/Xray/101index.html

- **Homology modelling course**: http://www.cmbi.kun.nl/gvteach/hommod/index.shtml

- **COGs** – phylogenetic classification of proteins encoded in complete genomes: http://www.ncbi.nlm.nih.gov/COG/

- **PDB** – Protein databank; home of 3D protein structure data: http://www.rcsb.org/pdb

- **ModBase** - Database of 3D protein models calculated by comparative modelling: http://alto.compbio.ucsf.edu/modbase-cgi/index.cgi

- **EBI-MSD** – European Bioinformatics Institute Macromolecular Structure Database European collaborative arm for PDB deposition with additional macromolecular analysis: http://msd.ebi.ac.uk

- **PDBsum** – A pre-calculated analysis of every protein of known structure currently available: http://www.biochem.uci.ac.uk/bsmpdbsum

- **DALI** – All by all protein structure comparison database: http://www.ebi.ac.uk/dali

- **CATH** – Semi automated classification of protein structure correlated to function. Proteins are clustered at four major levels: Class, Architecture, Topology and Homologous superfamily: http://www.biochem.ucl.ac.uk/bsm/cath/

- **SCOP** – Hand-curated classification of protein structure correlated to function: http://scop.mrc-lmb.cam.ac.uk/scop

De Hogeschool van Arnhem en Nijmegen (HAN) omvat vier faculteiten:

- Faculteit Gezondheid, Gedrag en Maatschappij
- Faculteit Educatie
- Faculteit Economie en Management
- Faculteit Techniek

20132014-IAS-HB-9215-H-539
2000000005348