Development of Complex Curricula for Molecular Bionics and Infobionics Programs within a consortial* framework**

Consortium leader

PETER PAZMANY CATHOLIC UNIVERSITY

Consortium members

SEMMELWEIS UNIVERSITY, DIALOG CAMPUS PUBLISHER

The Project has been realised with the support of the European Union and has been co-financed by the European Social Fund ***

**Molekuláris bionika és Infobionika Szakok tananyagának komplex fejlesztése konzorciumi keretben

***A projekt az Európai Unió támogatásával, az Európai Szociális Alap társfinanszírozásával valósul meg.

INTRODUCTION TO BIOINFORMATICS

(BEVEZETÉS A BIOINFORMATIKÁBA)

CHAPTER 6

Principles of proteomics

(A proteomika alapjai)

Péter Gál

Proteomics

Nucleic acids (DNA and RNA) are the information-carrier-molecules of the cells.

Proteins are the actual functional molecules of the cells. They are responsible for the running of the complex biochemical reaction network of the cells in interaction with each other and with a diverse spectrum of other molecules and ions.

The true understanding of a living system cannot be achieved without the direct study of proteins.

Proteome is the entire set of proteins in a cell (or tissue, or organism) produced by the genome. The proteome also includes the post-translationally modified forms of the polypeptides. The branch of science that deals with the proteomes is proteomics.

The rate of synthesis of different proteins in an organism vary among different tissues and different cells under different physiological states.

Methods are available for analysis of transcription patterns of genes (e.g. DNA microarray).

There are several reasons however why direct analysis of protein turnover is necessary.

1.) The transcriptome may not accurately represent the proteome either qualitatively or quantitatively.

The abundance of a given mRNA may not reflect the abundance of the protein it encodes. Protein synthesis is regulated not only transcriptionally but also post-transcriptionally (e.g. mRNA stability, rate of translation) and protein level is also controlled by degradation. Moreover, some mRNAs will never be translated (especially the alternatively spliced molecules).

2.) Protein diversity is generated post-translationally.

After translation many proteins are covalently modified (e.g. glycosylation, phosphorylation, proteolysis, etc). The actual levels of the post-translationally modified protein forms cannot be predicted from the level of the corresponding transcripts. Proteomical methods are necessary to measure the abundance of the modified forms of the same gene product.

3.) Different compartment of the cells, tissues and body may contain different amounts of the same protein. Most trafficking of gene products occurs at the protein level and the local protein concentration cannot be inferred from the transcription level.

4.) Some biological samples contain only proteins.

The extracellular space of most organism does not contain nucleic acids. For example: serum, cerebrospinal fluid, gastric juice and urine consist of proteins only.

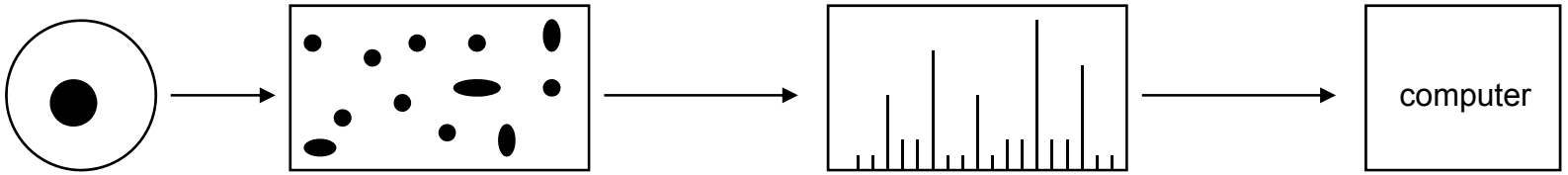
The protein levels and compositions of such fluids can change under different physiological or pathological conditions.

The emergence or disappearance of a certain protein in a body fluid may be a sign of disease. These proteins are useful biomarkers of the disease.

The proteomical analysis of such body fluid can be useful diagnostic tools.

Data collection for expression proteomics

Proteomical analysis of a sample means the separation of complex protein mixtures, the identification of individual components and their systematic quantitative analysis.



Sample

e.g. cell, tissue

2D gel

Mass spectrometry

Databases

Sample colection: disruption of biological samples (cells, tissues)

↓

Sample separation: typically by two-dimensional gel electrophoresis

↓

Identification and quantitation of the individual proteins (Mass spectrometry)

↓

Creating databases: catalog of the proteins in different biological samples

Two-dimensional gel electrophoresis (2DGE)

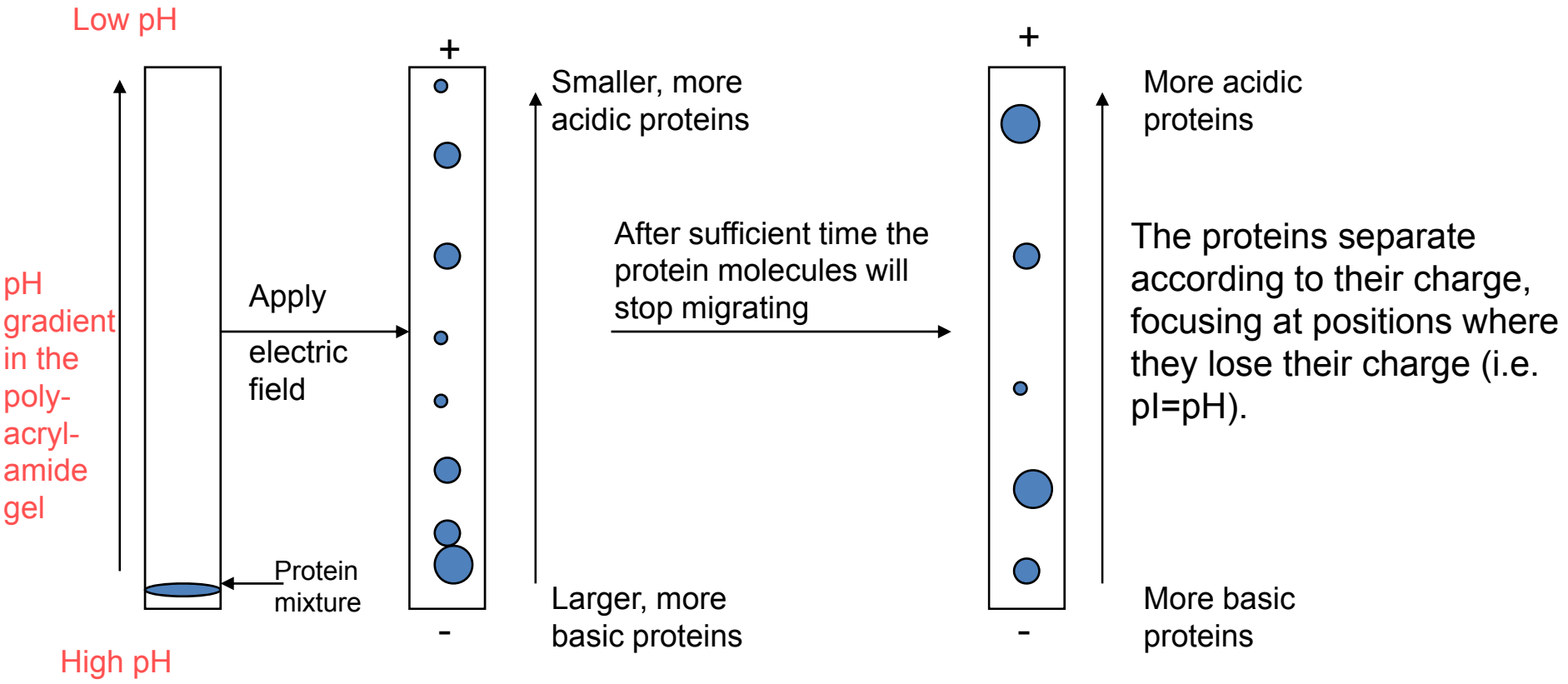
First dimension: isoelectric focusing (IEF) → separation according to net charge

Second dimension: sodium dodecylsulfate polyacrylamide gel electrophoresis (SDS-PAGE) → separation according to molecular mass

The two separation methods are applied one after the other in orthogonal dimensions.

Other separation techniques such as capillary electrophoresis (CE) are also used in proteomics, which can be followed by chromatographic methods (e.g. size exclusion chromatography, reversed phase HPLC) as the second dimension.

The principle of isoelectric focusing



Protein separation according to mass by means of SDS-PAGE

The second dimension in a two-dimensional gel electrophoresis experiment is usually sodium dodecylsulfate polyacrylamide gel electrophoresis (SDS-PAGE).

The proteins are denatured in the detergent (SDS) and their original charge are masked by the excess negative charge of the bound SDS molecules.

Consequently all protein SDS complexes have essentially the same charge density.

Since all the molecules have the same mass/charge ratio, the gel separates them according to their size (i.e. mass).

Two-dimensional electrophoresis

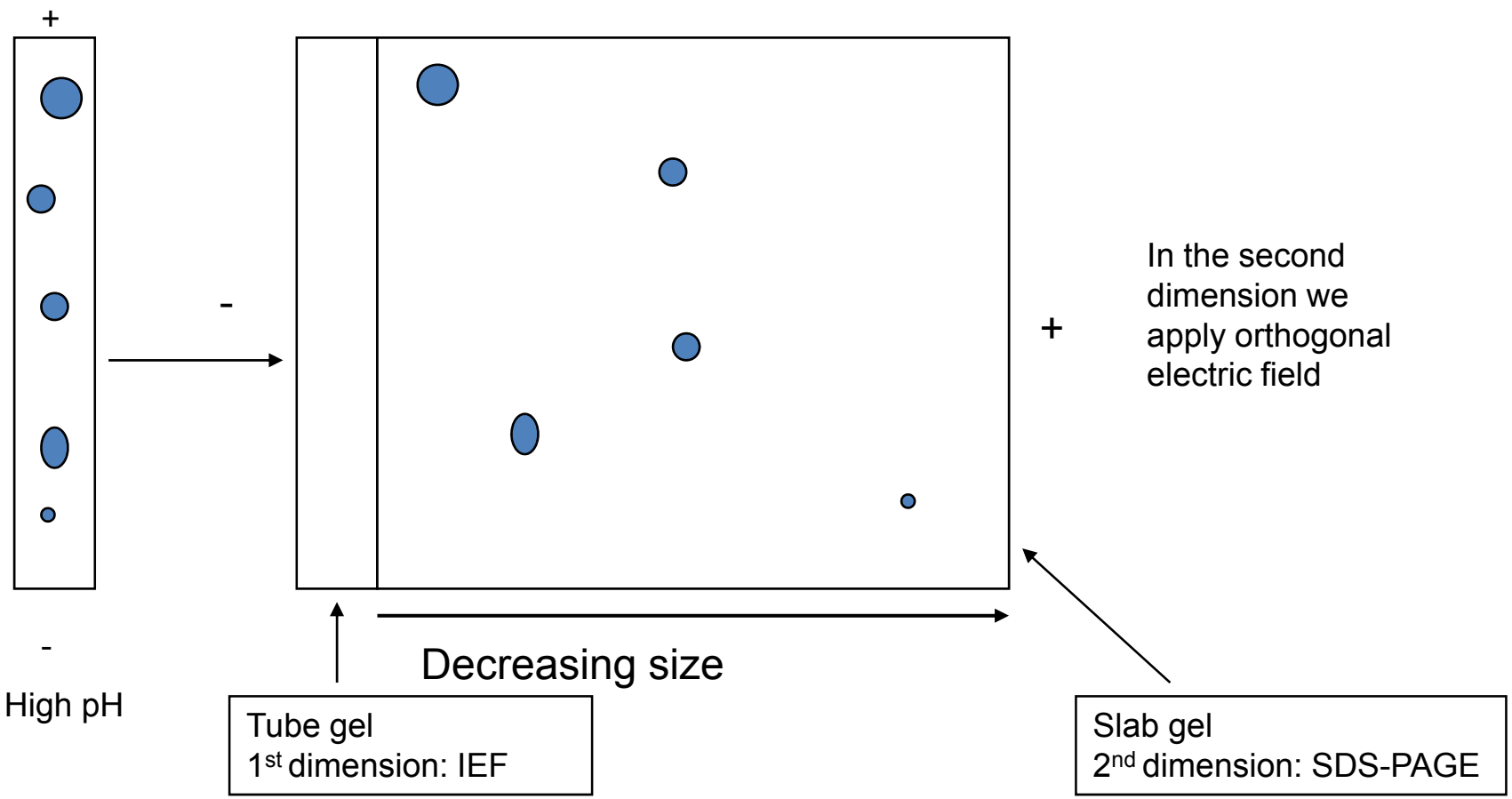
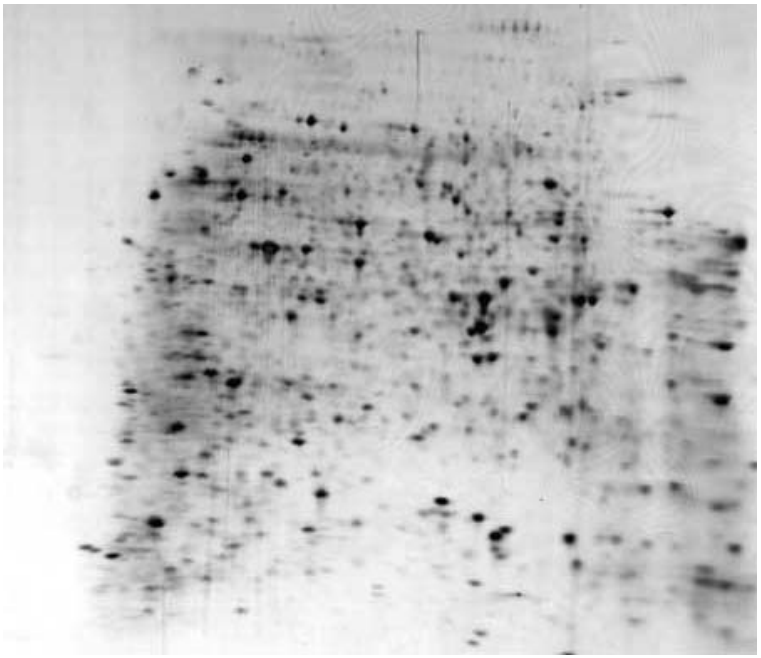


Image of a two-dimensional gel



Each spot represents a protein. The gel is usually stained with fluorescent dye and scanned with a laser scanner. The imiges are strored in data bases. The spots can be excised in order to identify the protein by mass spectrometry.

Protein sequencing

Sequencing of protein is more cumbersome than DNA sequencing.

Polypeptide chains are the polymeric association of 20 different kinds of amino acid subunits, while DNA contains only four types of nucleotides.

Moreover, DNA can easily be replicated *in vitro* using DNA polymerase enzymes, which forms the basis of all the high throughput DNA sequencing methods.

Proteins cannot be „replicated”, like the DNA, therefore only the degradative sequencing strategies work.

The main methods for analysing of protein sequences:

1.) Determination of the amino acid composition by complete hydrolysis

2.) Determination of the amino terminal residue of the protein

3.) Determination of the sequence by chemical (Edman) degradation

4.) Identification of the protein by mass spectrometry (MS)

5.) Determination of the sequence by tandem mass spectrometry

1.) Proteins can be hydrolyzed completely by boiling in concentrated (6M) hydrochloric acid. The resulting free amino acids can be labeled, separated and analyzed qualitatively as well as quantitatively. If we have the amino acid composition we can search in the protein sequence databases for protein with the same amino acid composition. It is a slow and laborious way of protein identification.

2.) Identification of the amino terminal residue of a protein is often the first step in the protein sequencing procedure. The amino terminal amino acid can be labeled by 1-fluoro-2,4-dinitrobenzene (FDNB) (Sanger method) or dansyl chloride. After hydrolysis of the labeled protein the amino terminal residue can be identified.

3.) The Edman degradation is a chemical method for protein sequencing. In this procedure amino acids form the N-terminus are removed selectively and progressively. The removed amino acids can then be identified (e.g. chromatography). The peptide is reacted with phenylisothiocyanate followed by mild acidic hydrolysis that results in the cleavage of the peptid bond connecting the labeled (N-terminal) amino acid to the rest of the protein. The liberated, labeled amino acid can be identified. The rest of the peptide chain remains intact and the new amino terminus can be exposed to the same procedure. This procedure is repeated until the entire sequence is determined.

The Edman degradation procedure can be automated.

The Edman sequenator can determine the sequence of 10 amino acids in about 24 hours, however longer sequences (30-40 amino acids) require several days.

The upper limit of contiguous amino acids that can be cleaved and determined from a protein is about 50, since the efficiency of the degradations is less than 100%.

Larger proteins can be degraded into smaller fragments (e.g. proteolysis) before Edman sequencing.

Edman degradation is the most convenient method for determining the N-terminal sequence of a protein.

It is also extremely sensitive: 0.5-1 pmol of pure protein is enough.

4.) **Mass spectrometry**

The principle of protein identification with mass spectrometry:

Measuring the molecular mass of the protein with high precision.

It identifies the protein unambiguously.

It is suitable for rapid (few minutes) sequencing of short stretches of polypeptides (20-30 aa).

Low quantity of sample is required (0.5-1 pmol).

6-8-long amino acid sequence is usually sufficient to identify a protein unambiguously.

Knowing the short protein sequence we can identify and clone the corresponding gene.

The protein molecule should be brought into the gas phase which is followed by ionization in the vacuum. The protein molecule-ion is introduced into an electric and/or magnetic field. Its path through the field (e.g. time of flight TOF) is a function of its mass-to-charge ratio (m/z). This measured property of the ionized protein molecule can be used to deduce its mass (M) with very high precision.

Problem: Introducing a biological macromolecule, especially a protein, into the gas phase is not an easy task. The standard methods that were used in the case of small molecules caused the rapid decomposition of the macromolecules.

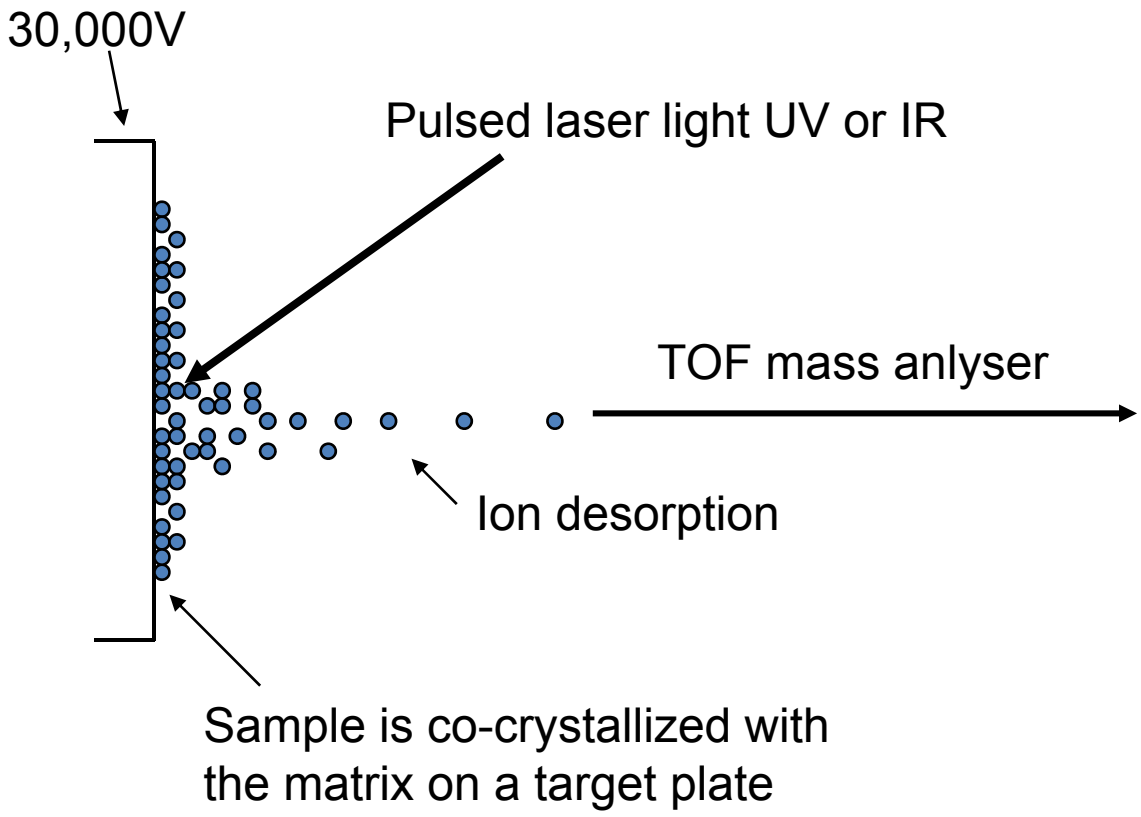
There are two so-called soft-ionization methods that achieve the ionization of peptides without significant fragmentation:

1.) MALDI (matrix-assisted laser desorption/ionization)

The peptide is mixed with a large excess of a matrix compound that can absorb energy from the laser light and emit it in form of heat. A short pulse of laser light causes the rapid ionization and sublimation of the peptide into the vacuum system.

MALDI is used predominantly for the analysis of peptide mixtures, such as the peptides derived form a single spot from a 2D gel. Besides of peptides MALDI can be used to measure the mass of a wide range of macromolecules.

MALDI MS

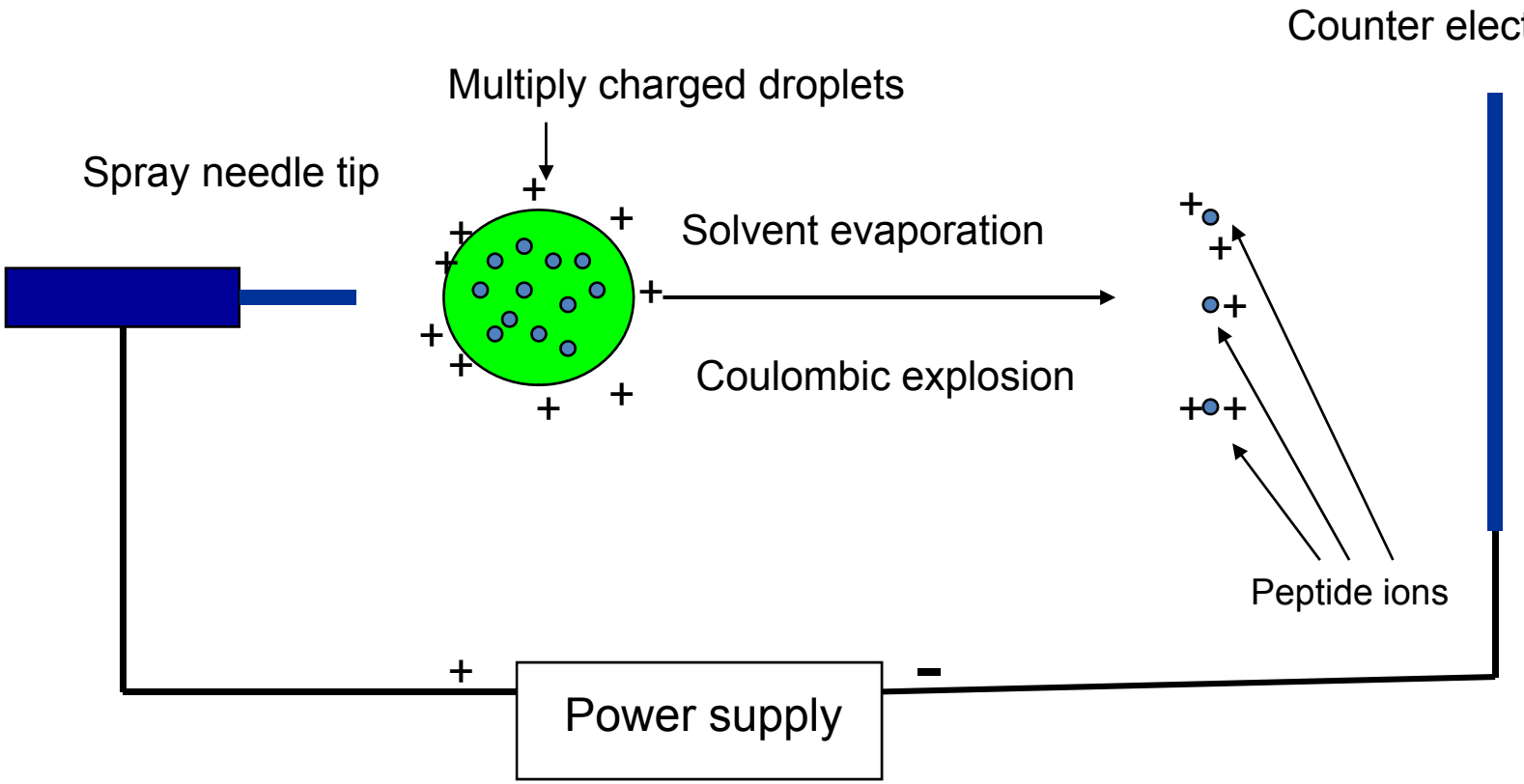

30,000V
Pulsed laser light UV or IR
TOF mass anlyser
Ion desorption
Sample is co-crystallized with
the matrix on a target plate

2.) ESI (electrospray ionization)

The protein solution is passed through a charged needle that is kept at a high electrical potential, dispersing the solution into a fine mist of charged microdroplets. The solvent rapidly evaporates and the resulting multiply charged macromolecular ions are thus introduced nondestructively into the gas phase. The charge usually comes form the absorbed protons. The m/z of the charged molecules can be analysed in the vacuum chamber.

Whereas MALDI-MS is used to analyse simple peptide mixtures, ESI-MS is more suited to the analysis of complex samples.

ESI-MS


Counter electrode
Multiply charged droplets
Spray needle tip
Solvent evaporation
Coulombic explosion
Peptide ions
+
Power supply
-

Molecular mass determination with ESI MS

Since a protein acquires variable numbers of protons, and thus positive charges, the spectrum will contain a series of successive peaks differing by a charge of 1 and a mass of 1 (e.g. 1 proton).

$(m/z)_2 = (M+n_2X)/n_2$

$(m/z)_1 = (M+(n_2+1)X)/n_2+1$

n_2: number of charges

X: mass of the added group (proton in this case)

Two unknowns (M, n_2), two equations. We can solve it for M and n_2 at any two neighboring peaks.

5.) Protein sequencing by tandem MS (MS/MS)

The protein to be sequenced is first treated with a protease or chemical reagent in order to cleave it into shorter fragments.

The mixture is then injected into the tandem MS device.

In the first MS the peptide mixture is sorted so that only one type of peptide molecule is selected for further analysis.

The selected peptide is further fragmented in the collision cell by colliding the molecules with a stream of inert gas such as helium or argon, a process known as collision-induced dissociation. The peptides will brake mainly at the peptide-bonds. The procedure is designed to fragment most of the peptide molecules, but one molecule will suffer only one breakage.

This process results in two series of ions depending on localization of the charge:

B-series: the charge remains on the N-terminal fragment

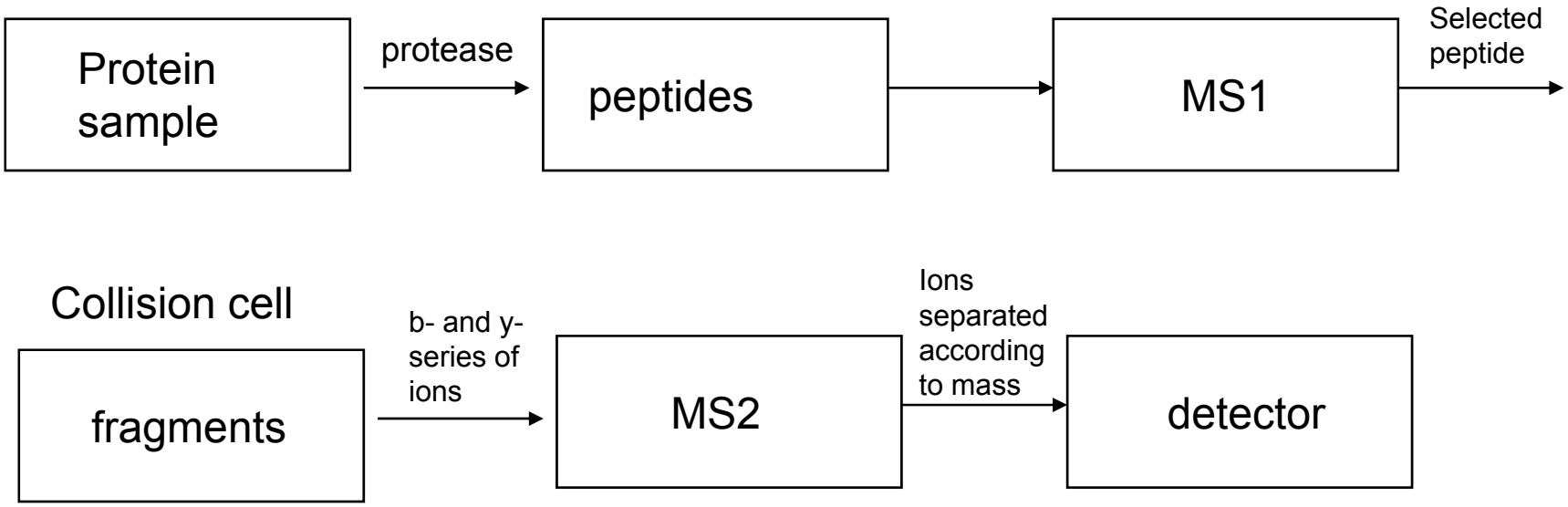
Y-series: the charge remains on the C-terminal fragment

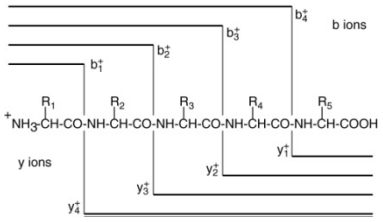
In both cases we have molecule ions with contiguous and intact amino acids.

In the second MS device we can arrange the elements of the b- or y-series according to increasing mass.

The difference in mass between consecutive ions in either series should correspond to the masses of individual amino acids. Using this information we can derive the sequence of the peptide.

Protein sequencing by means of tandem MS (MS/MS)





The successive fragments differ from each other only in one amino acid that was lost in each case. The sequence can be read.

To interpret the MS/MS spectrum we have to unambiguously identify the members of the b- and/or y-series.

At the sites of the breakage there are no intact carboxyl and amino groups. The only intact α-amino and α-carboxyl groups can be found at the very ends of the peptide fragments. The members of the b- and/or y-series can be assigned by the slight differences between the intact N- and C-termini.

Another useful approach is to divide the sample into two aliquots, attach a specific mass label to either the N- or C-terminus of the intact peptide in one of the aliquots, and then compare the mass spectra to identify the modified and unmodified forms.

One example for end labeling is the methyl esterification of the C-terminus of a peptide. This treatment adds 14 units to the original mass of the peptide. The members of the y-series can then be distinguished form other ions by comparison of the mass spectra of the of the treated and untreated samples.

Another labeling strategy is to incorporate a heavy isotope (^{18}O) into the C-terminus. If we perform the proteolytic fragmentation of the original protein in a buffer containing ^{18}O-water, trypsin incorporates a heavy oxygen atom into the carboxyl group of the newly generated C-terminus of each peptide. Thus each y-series fragment will carry two extra mass units.

The *de novo* „ladder" protein sequencing by mass spectrometry is a quick and reliable method to generate short sequences for protein identification.

Since the method is based on mass determination there are a few limitations:

1.) Leucine and isoleucine have identical masses (113), therefore mass determination cannot distinguish between them.

2.) Two adjacent glycine residues (57) may be mistaken for a single asparagine residue (114).

3.) In the case of glutamine and lysine there is only a slight difference in mass (128.13 vs. 128.17).

The ambiguities in the MS/MS generated protein sequences can be resolved by chemical (Edman) sequencing or by cloning and sequencing the corresponding gene.

The tandem MS sequencing alone cannot yield complete sequence information. Edman degradation and mass spectrometry complement each other in protein sequencing.

Tandem MS protein sequencing however is ideal for proteome research aimed at cataloging the hundreds of cellular proteins separated on the 2D gels.

Thousands of protein sequences are available in databases accessible through the Internet. The comparison of a newly obtained sequence with this large bank of sorted sequences can offer insights into its three dimensional structure, function, cellular localization, etc.

Protein interaction databases

Proteins are the executive molecules of the cell. Most of the functions of a cell are carried out by proteins. Most proteins exert their function as a part of larger complexes rather than working in isolation. Protein interactions lie at the heart of most biological processes.

Protein-protein interactions results in the formation of transient or stable multi-subunit complexes.

Investigation of protein interactions can help in the functional annotation of uncharacterized, hypothetical proteins.

An understanding of the nature of protein complexes is a step towards the elucidation of molecular pathways such as signaling cascades and regulatory networks.

In addition to protein-protein interactions other type of interactions such as protein-nucleic acid interactions are also important.

Proteins also interact with small molecules, which act as ligands, substrates, cofactors or allosteric regulators.

We use methods for detecting protein interactions that do not reveal the precise chemical nature of the interactions but simply report that such interactions take place.

There are many different techniques to detect protein-protein interactions but the major high throughput technique is the yeast two-hybrid system.

Protein chips are also emerging as useful tools for the characterization of protein interactions.

Molecular interaction data can be used for the construction of interaction maps of the entire proteome. These maps are graphs showing proteins or protein complexes as nodes and interactions as the links between them.

There are four types of methods for studying protein interactions:

1.) Genetic methods

2.) Affinity methods

3.) Molecular and atomic methods

4.) Library-based methods

1.) Genetic methods

In genetically amenable species such as the yeast *Saccharomyces cerevisiae*, the fruit fly *Drosophila melanogaster*, the nematode *Caenorhabditis elegans*, the mouse *Mus musculus*, and the model plant *Arabidopsis thaliana* protein-protein interactions can be inferred from genetic analysis.

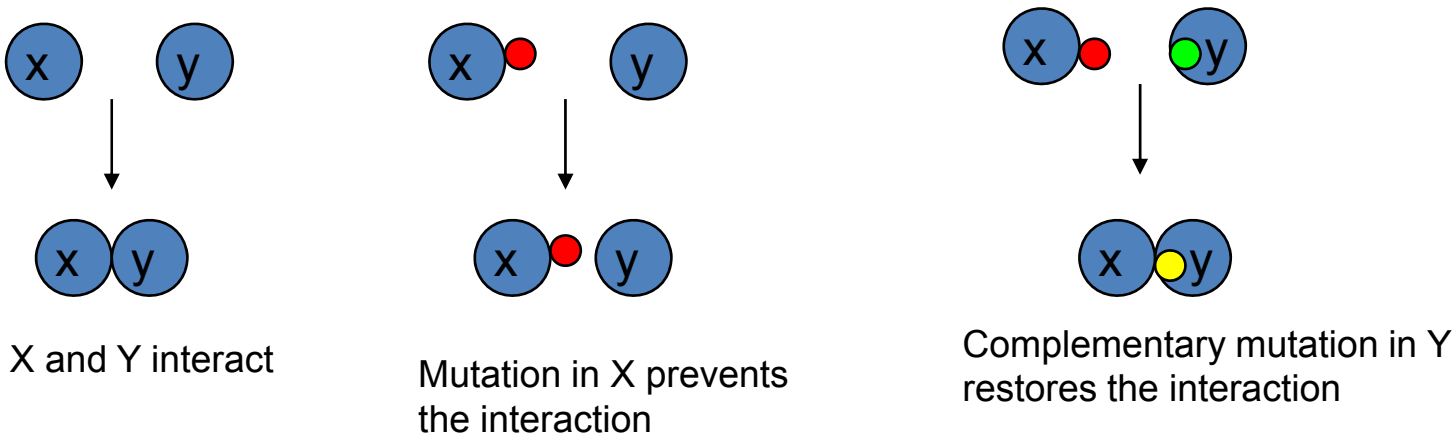
The basis of the genetic methods is that interactions between two given proteins can be studied by looking at the behavior of mutations in their corresponding genes.

One method is screening for **suppressor mutants**, i.e. secondary mutations that correct the phenotype of a primary mutation.

Suppressor mutations

A primary mutation in gene X causes a conformational change in protein X that prevents its interaction with protein Y therefore causing a loss of function.

The suppressor mutation in gene Y introduces a complementary change in protein Y that restores the interaction and thus rescues the mutant phenotype.



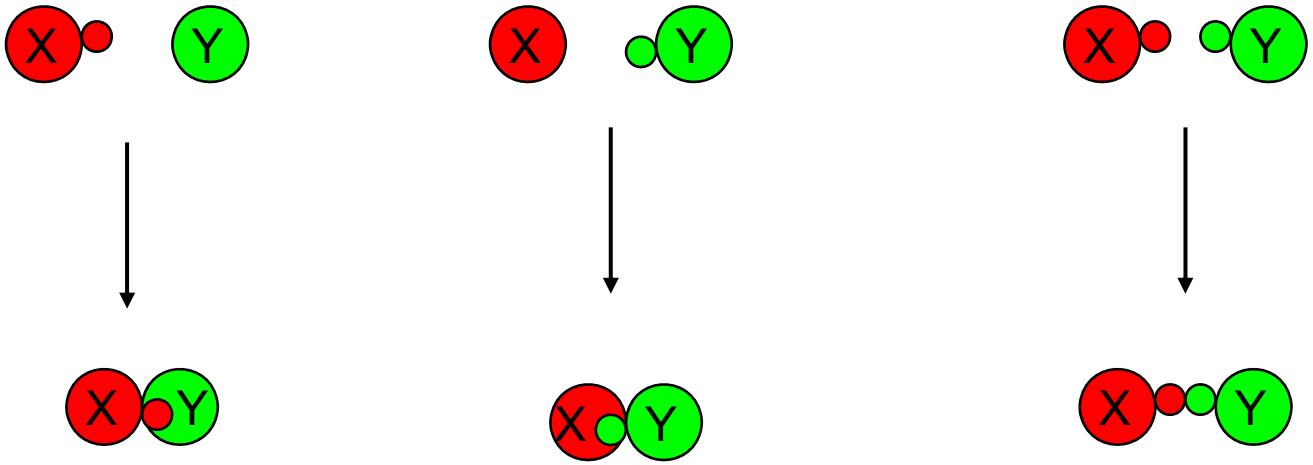
Another genetic test for protein interactions is the **synthetic lethal screen.**

A single mutation in protein X or Y is not lethal if only one mutant protein is present in a cell (organism).

However if both mutations are present in the same individual the interaction between X and Y is disrupted and a lethal phenotype is observed.

Example: Synthetic genetic array (SGA) system for yeast: A mutation in one yeast gene can be crossed to a set of 5000 viable deletion mutants. In this way we can identify all the proteins involved in the same pathway or complex.

Synthetic lethal effect



Mutation in X can be tolerated → interaction with Y

Mutation in Y can be tolerated → interaction with X

Mutation in both X and Y cannot be tolerated → no interaction between X and Y →lethal phenotype

In the **dominant negative** approach a nonfunctional mutant form of the protein is introduced into the cell (e.g. mRNA injection, transient or stable expression of recombinant protein) that quashes the activity of any normally functioning version of the protein in the same cell.

The formation of multi-subunit complexes can be demonstrated by this approach.

Although the genetic methods provide valuable information about protein interactions they do not provide definitive proof.

Candidate protein interactions must be confirmed at the biochemical level experimentally.

Genetic and genomic methods infer protein interactions but do not demonstrate them directly. Biochemical methods are needed to prove them.

In the **affinity chromatography** method protein X is immobilized on a matrix such as a Sepharose column.

A complex mixture of proteins (e.g. cell lysate) is passed through the column under controlled conditions (i.e. pH, ionic strength, temperature). Most of the proteins in the mixture will pass through the column but those that interact with protein X will be retained.

After washing the column with the binding (usually low-salt) buffer the bound proteins can be eluted by a buffer having different composition (e.g. high-salt, low pH, detergent, etc.).

The eluted interacting proteins can then be identified by mass spectrometry or immunoblotting.

The bait protein can be immobilized on the column by different techniques.

Using recombinant DNA technology we can put N- or C-terminal tags to the protein through which we can bind it to the matrix. Frequently used affinity tags are glutatione-S-transferase (GST), chitin-binding domain, maltose-binding protein, etc. The bait protein is recombinantly expressed as a fusion with the tag protein.

The recombinant affinity tags (e.g. His-tag) can be directly used for protein purification.

GST-pulldown is a popular example for detection of protein-protein interaction by affinity based method. The GST-fusion bait protein is immobilized on gluthatione-coated Sepharose beads. Theoretically any protein can be expressed as GST-fusion protein. A control experiment with immobilized GST is necessary to weed out those proteins that interact with GST itself.

Another methods to detect binary protein-protein interactions are co-immunoprecipitation and chemical cross-linking.

Protein chips (or protein microarrays) are suitable for large-scale analysis of protein interactions.

Protein chips are manufactured in a similar manner as the spotted DNA microarrays, that is, by the robotic spotting of small amounts of liquid onto a solid miniature platform, such as a glass slide.

The most common type of analytical protein chip is the antibody array. The chip is flooded with the complex mixture of proteins, allowing the antibodies to capture any antigens that are present, and then washed to remove unbound proteins. The proteins can be labeled with fluorescence dye and detected in a laser scanning device.

Library-based methods for the analysis of protein interactions

The yeast two-hybrid system (Y2H)

The basic principle of the system is the assembly of an active transcription factor from two fusion proteins and the detection of this assembly by the activation of a marker gene.

The bait protein is expressed as a fusion (a hybrid) with the DNA binding domain of a transcription factor. This construct is unable to activate the transcription of the marker gene alone. This bait fusion is expressed in one haploid yeast strain.

Another haploid yeast strain is used to create an expression library in which each clone is expressed as a fusion protein with the transactivation domain of the transcription factor. This construct also is unable to activate the transcription of the marker gene alone.

The third component of the system is a reporter gene that is activated specifically by the two-hybrid transcription factor.

The two strains of yeast are then mated to yield a diploid strain expressing both the hybrid bait protein and one candidate hybrid prey protein.

In those cells where the bait and prey do not interact, the transcription factor remains unassembled and the marker gene remains silent.

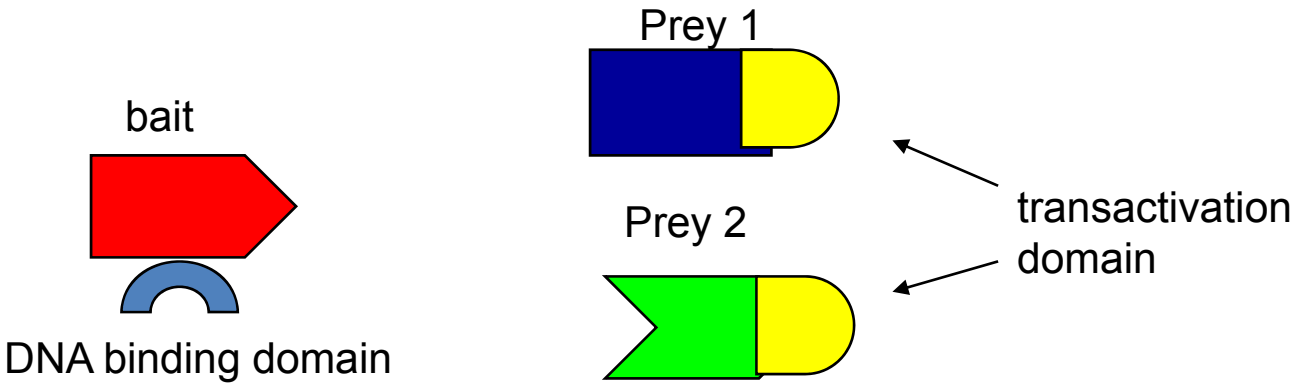
In those cells where the bait interacts with the prey the transcription factor is assembled and the reporter gene activated, allowing the cells to be isolated and the cDNA sequence of the prey protein determined.

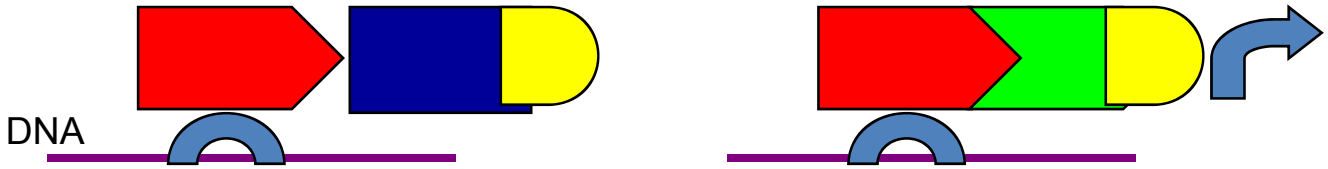
The Y2H system was the first technology to facilitate global protein interaction analysis. Tens of thousands of interactions can be screened in a single experiment.

Several very comprehensive large-scale studies have been made: testing the entire genome of some viruses, bacteria, yeast, fruit fly, nematode worm.

The interactome: the sum of all binary interactions between the proteins of a cell or organism.

Principle of the yeast two-hybrid system





Prey 1: no interaction→no transcription

Prey 2: interaction→transcription of the reporter gene

Limitations of the Y2H system

The Y2H system is the only available *in vivo* technology for the high-throughput systematic analysis of binary protein-protein interactions, but there is a relatively high level of false positives and false negatives.

False positives: the reporter gene is activated in the absence of any specific interaction between the bait and prey. (Possible reasons: the prey is sticky (i.e. it makes nonspecifical interactions), or it can activate the transcription without interacting with the bait (autoactivation).)

False negatives: the reporter gene is not activated even if the bait and prey do normally interact. (Possible reasons: in the fusion protein the prey has a non-native conformation, or the interacting surface is not accessible due to the fusion.)

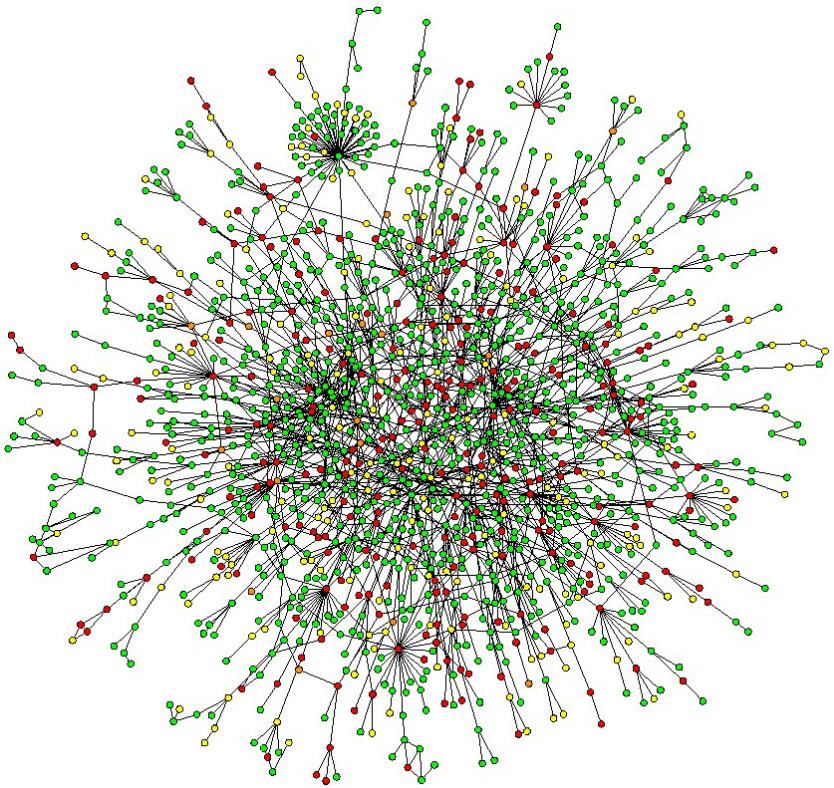
Proteins that physically interact with each other may be involved in the same molecular pathway or network, or may form part of a multi-subunit complex.

One role of bioinformatics is to provide protein interaction databases that allow interaction data to be stored, queried, assessed for confidence and used for pathway reconstruction.

The ultimate challenge for bioinformatics in the field of protein interaction technology is the reconstruction of the interactome, i.e. the sum of all protein interactions in the cell.

The simplest way to represent protein interactions is a graph with proteins as nodes and interactions as links.

Protein interaction network of yeast



For a small number of proteins, interaction maps are very useful.

However, larger numbers of proteins yield graphs of incredible complexity.

Simplification can be achieved by clustering functionally similar proteins, that is, allocating proteins to functional categories (e.g. DNA replication, membrane transport, etc.).

A functional interaction map shows a network of the basic cellular functions.

The interaction networks can be used in systems biology to describe the metabolic and regulatory interactions in the cell.

The role of bioinformatics in drug development

Drug is a molecule which triggers a biological effect.

The biological effect is triggered through the interaction with a target molecule in the body. The target molecule is usually a protein (in most cases an enzyme or receptor) to which the drug can bind.

The biological effect of a drug can be beneficial or harmful.

The pharmaceutical industry aims to develop drugs with specific beneficial effects to treat human diseases.

A drug target can be endogenous (a human protein whose (improper) activity contributes to the pathogenesis of the disease) or it can be a protein produced by an infectious agent (e.g. pathogenic bacterium).

A drug can either stimulate or block the activity of the target protein in order to trigger the beneficial effect.

The major phases of drug development:

Preclinical phase:

1.) Target identification

2.) Target validation (disease models)

3.) Lead discovery

4.) Lead optimization

5.) Animal studies

Clinical phase:

Phase I

Phase II

Phase III

Phase IV

1.) Target identification

The first step in drug discovery is to select a suitable target through which the patho-physiological process can be modulated. Nearly all major areas of bioinformatics can be applied at this stage. Structural- and functional genomics, proteomics, systems biology can help in finding the proper target protein. Genome annotation (gene finding by computer), analysing of global expression data (e.g. DNA microarrays), analysis of protein interaction data are of great importance.

The association between genomic mutations and the disease (e.g. SNP patterns) can also provide useful hints.

2.) Target validation

Extensive experimental testing of the therapeutic potential of the target molecule. It must be clearly demonstrated that the target contributes to a human disease. This process includes „wet laboratory" experiments (e.g. enzymatic measurements) and creation and application animal disease models.

3.) Lead discovery

Lead discovery means search for compounds that have some of the desired biological effects. High throughput screening (HTS) is a method to test large compound libraries. The initial compound libraries are usually too large, therefore screening of a smaller, focused libraries is beneficial. *In silico* library focusing can be a task for bioinformaticians.

4.) Lead optimization

Lead optimization involves testing of chemically modified forms of the lead compound in order to find candidate drugs with better therapeutic profile. Optimization includes increasing the specificity, improvement of synthesis and formulation, etc.

5.) Animal studies

The safety and tolerance levels of the candidate drugs are first tested in animal experiments.

Clinical phase

Clinical trials of the candidate drugs are performed in four phases.

The purpose of the clinical trials is to determine safety and tolerance levels of humans, and to study how the drug is metabolized.

Clinical trials:

Phase I: involves healthy human volunteers to test the safety and tolerance levels in human.

Phase II: involves small number of patients in order to check the safety and efficacy of the candidate drug and to select the dose regimen.

Phase III: involves large number of patients to monitor the effect of the candidate drug.

Phase IV: The long term monitoring of the potential adverse side effects.

In the recent years development of genomics, proteomics and systems biology have revolutionized drug discovery, especially target identification and validation. The sequencing and annotation of the human genome yielded thousands of potential new targets. It must be noted however that only a small fraction of these potential targets are considered by pharmaceutical companies mainly because of economical reasons. The cost of the development of a new drug can currently be estimated somewhere in the region of 500 million to more than two billion dollars. It is therefore not surprising that pharmaceutical companies focus primarily on drugs that treat major diseases with a large market potential. Developing drugs for rare diseases („orphan drugs") is usually unprofitable. Recently several drug regulation authorities (e.g. US Food and Drug Administration, European Medicines Agency) have initiated special programs to encourage development of orphan drugs.

Bioinformatics is also important to model the interaction between the target protein and the potential drug molecules. **Rational drug design** uses protein structural data to predict the type of ligands that will interact with a given target, and thus form the basis of lead discovery.

Personalized medicine is the concept to treat the patient with drugs tailored to his/her genotype. While a drug can be very effective in treating a patient, in other patients it may show only little benefit or even it may exert severe side effects. Such individual variations may reflect differences between the individual's genomes resulting in the differences in the structure and/or metabolism of the target protein. **Pharmacogenomics** is the field of science that deals with the relationship between genomic variations and drug response patterns. Individual variations in a particular genome can be screened by DNA microarray (SNP profiling) or by direct sequencing (next generation sequencing). Individual differences in drug response patterns are responsible, in part, for the high failure rates of new drug candidates at the clinical trials. Application of personalized medicine may increase the success rate: i.e. more new drugs can reach the market.

Useful bioinformatics sites on the WWW

1.) NCBI ≡ National Center for Biotechnology Information / www.ncbi.nlm.nih.gov

One of the best starting points for studying bioinformatics. A resource of public databases (with its own data retrieval tool: Entrez), bioinformatics tools and applications. Link to many useful sites and resources for bioinformatics software. An excellent science primer.

2.) EBI ≡ European Bioinformatics Institute / www.ebi.ac.uk

The EMBL European Bioinformatics Institute outstation. A resource for biological databases and software, much of which has excellent tutorial support.

3.) EMBL ≡ European Molecular Biology Laboratory, Heidelberg / www.embl-heidelberg.de

4.) Sanger Institute / www.sanger.ac.uk

An institute of genomics supported by the Wellcome Trust.

5.) ExPASy ≡ Expert Protein Analysis System / www.expasy.ch

The proteomics server of the Swiss Institute of Bioinformatics. Annotated protein sequence databank with many useful softwares.

6.) PDB ≡ Protein Data Bank / www.rcsb.org

The protein structure databank.

7.) Nucleic Acid Research database issue / www3.oup.co.uk/nar/database/c/

The periodical NAR publishes annually a database issue compiling the most important and reliable biological databases. The 2010 issue contains 1230 databases.