

Proteomics: Tools and applications

“The most beautiful thing we can experience is the mysterious. It is the source of all true art and all science. He to whom this emotion is a stranger, who can no longer pause to wonder and stand rapt in awe, is as good as dead: his eyes are closed.”

(A. Einstein)

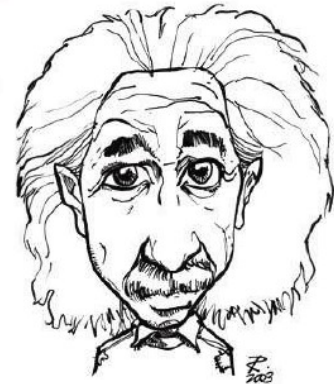


Table of contents

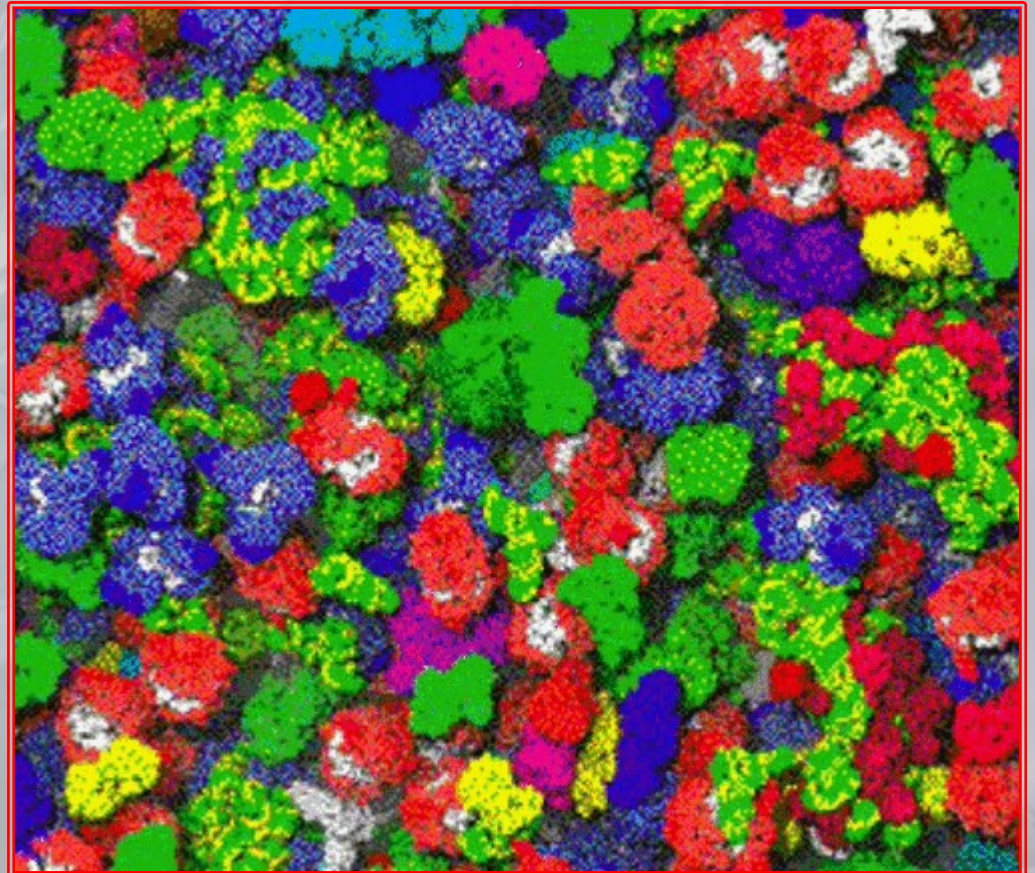
- ✦ From the genome to the proteome
- ✦ Classification of proteins
- ✦ Experimental techniques
- ✦ Inhibitor and drug design
- ✦ Screening of ligands
- ✦ X-ray solved crystal structures
- ✦ NMR structures
- ✦ Protein–protein interactions
- ✦ Empirical methods and predictive techniques
- ✦ Post–translational modification prediction

Introduction – 1

- ✦ While the genome collects the whole genetic material of an organism, the **proteome** represents the set of its proteins
- ✦ The nature of genes – their simple chemical composition and their ability to be used as templates to make exact copies of themselves – made them relatively easy to study and analyze with automatic methods
- ✦ The nature of proteins – with their twenty elementary components, the complex chemical changes they subdue, together with their replication inability – made them much more difficult to be analyzed
- ✦ **Proteomics**: Science that allows an in-depth study of the proteome, the complete kit of proteins expressed in a cell or in a tissue (Wilkins et al., 1996)

Introduction – 2

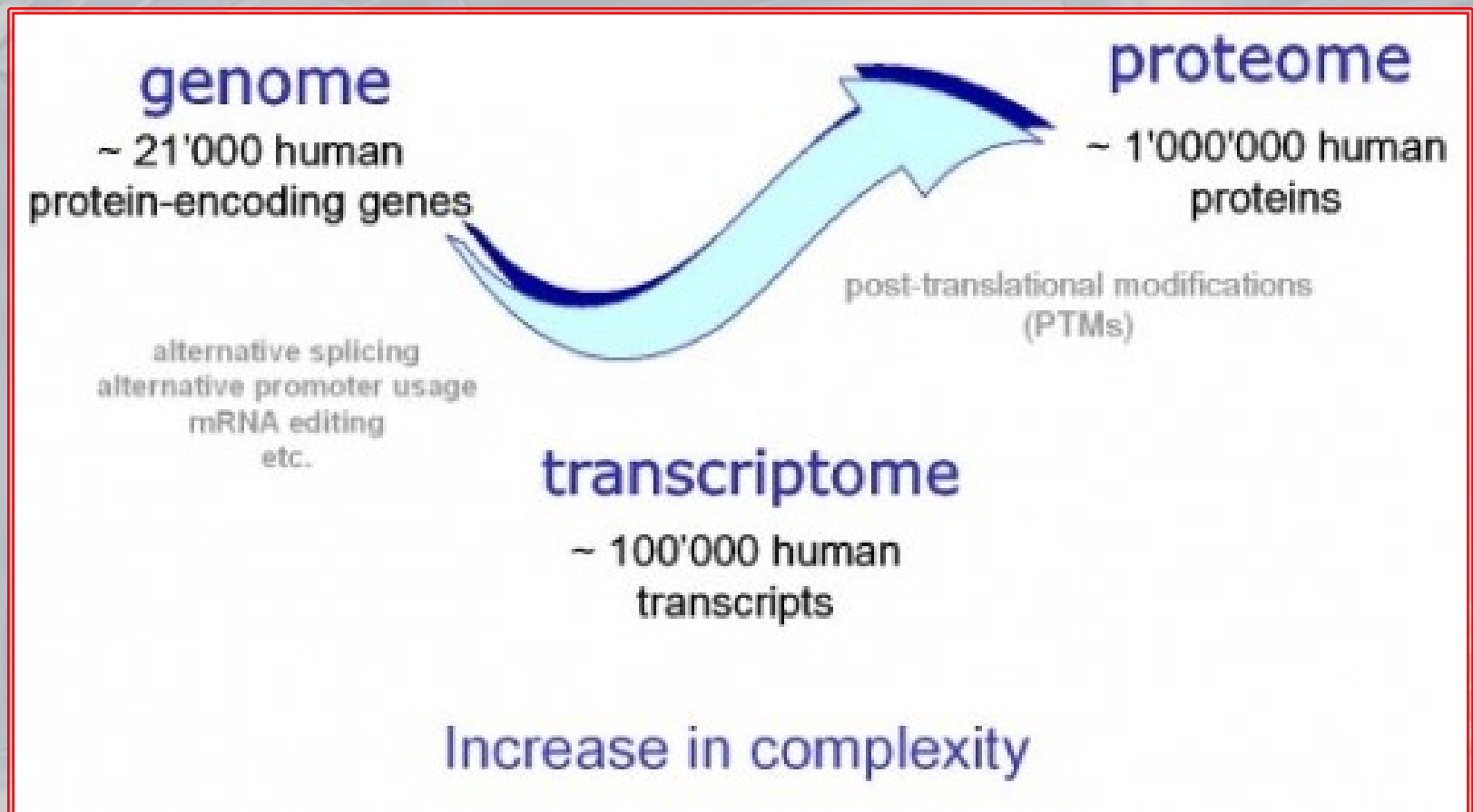
- ✦ Proteins are the agents that, within the cell, “do what it is needed to do”
- ✦ Proteins, like all cellular components, have a certain turn-over; their average life can vary from a few minutes to many days



Introduction – 3

- ✦ One of the major findings, of the new post-genomic era, is that the old paradigm according to which a gene codes for a single protein is no longer valid, at least in eucaryotic organisms
- ✦ In fact, due to alternative splicing and post-translational modifications (proteolytic cleavage, glycosylation, phosphorylation) of proteins, more than one proteome may correspond to a single genome
- ✦ The genome of a living being, even when completely sequenced, does not allow to fully understand all the biological functions that characterize an organism, which depend on multiple factors, including regulatory and metabolic pathways of proteins

Introduction – 4



Introduction – 5

- ✦ Therefore, **proteomics** appears complementary to genomics, and essential for the understanding of biological mechanisms
- ✦ Proteomics allows the study of proteins, both in the form just translated from genes and in isoforms (different proteins coming from the same gene due to alternative splicing), or after any post-translational change, which may occur in the cell after translation
- ✦ The study of isoforms and of post-translational modifications allows to understand interaction mechanisms between proteins: these mechanisms will affect their activity and function

Introduction – 6

- ✦ *The long-term challenge of proteomics is enormous: to define the identities, quantities, structures and functions of complete complements of proteins, and to characterize how these properties vary in different cellular contexts (M. Tyers & M. Mann, 2003)*
- ✦ **Proteomics aims**
 - Genomics integrated strategies
 - Study of post-translational modifications
 - Identification of novel protein targets for drugs
 - Analysis of tumor tissues
 - Comparison between normal and diseased tissues
 - Comparison between diseased and pharmacologically treated tissues
 - ...

Introduction – 7

- ✦ The current proteomic studies are mainly focused on two principal areas:
 - Functional proteomics
 - Expression proteomics
- ✦ Functional proteomics aims at defining the biological function of proteins, the role of which is still unknown, and to identify *in vivo* protein–protein interactions, in order to describe cellular mechanisms at the molecular level
- ✦ Expression proteomics is focused on the qualitative and quantitative study of the expression profiles of different proteins; the expression of proteins can in fact be altered by changes in cellular conditions (different growth conditions, stress, in presence of cellular diseases, etc.)
 - The different protein profiles in a tissue, the absence, the presence or some different quantity levels, are potential biomarkers of physiological and/or pathological situations

From genome to proteome – 1

- ✦ Despite the ability to generate staggering amounts of data, techniques for gene expression analysis provide little information about proteins that are present within a cell, and even less about what is their function and how it is carried out
 - The longevity of an mRNA and that of the protein it encodes are usually very different and the correlation between the relative abundance of an mRNA and of its corresponding protein, within each cell, is usually less than 0.5
 - Many proteins, after translation, undergo extensive biochemical changes, in very different ways
 - ✗ These modifications, almost invariably, alter the protein activity and manifest themselves in different forms, depending on particular tissues and circumstances

From genome to proteome – 2

- Many proteins are not functionally relevant until they are assembled into larger complexes or if they are not transported in appropriate locations within or outside the cell
 - ✗ The amino acid sequence can only offer some indications on the purpose of such interactions and on the final destination of the protein
- ➡ Difficulty in deducing the population of the proteins of a cell and their individual role, even aggravated by the limited availability of proteins that can be directly analyzed

From genome to proteome – 3

- ✦ Proteins require much more accurate manipulations compared to DNA, because their tertiary structure can easily be altered when they come into contact with an inappropriate surface or environment
- ✦ Moreover:
 - The ability of nucleic acids to specifically hybridize with other nucleotide sequences make the DNA identification a relatively simple task
 - The identification of proteins is much more difficult and requires complicated analysis of mass spectrometry and advanced software tools, or the generation of specific antibodies

From genome to proteome – 4

- ✦ However, the potentialities arising from the knowledge of the proteome of an organism are enormous, for example for:
 - Augmenting the efficiency of genetically engineered organisms
 - Understanding the molecular basis of some diseases
 - Designing new targeted drugs

From genome to proteome – 5

✦ Genome vs Proteome

- The caterpillar and the butterfly are genetically identical, but possess very different proteome and phenotype, as well as the tadpole and the frog!



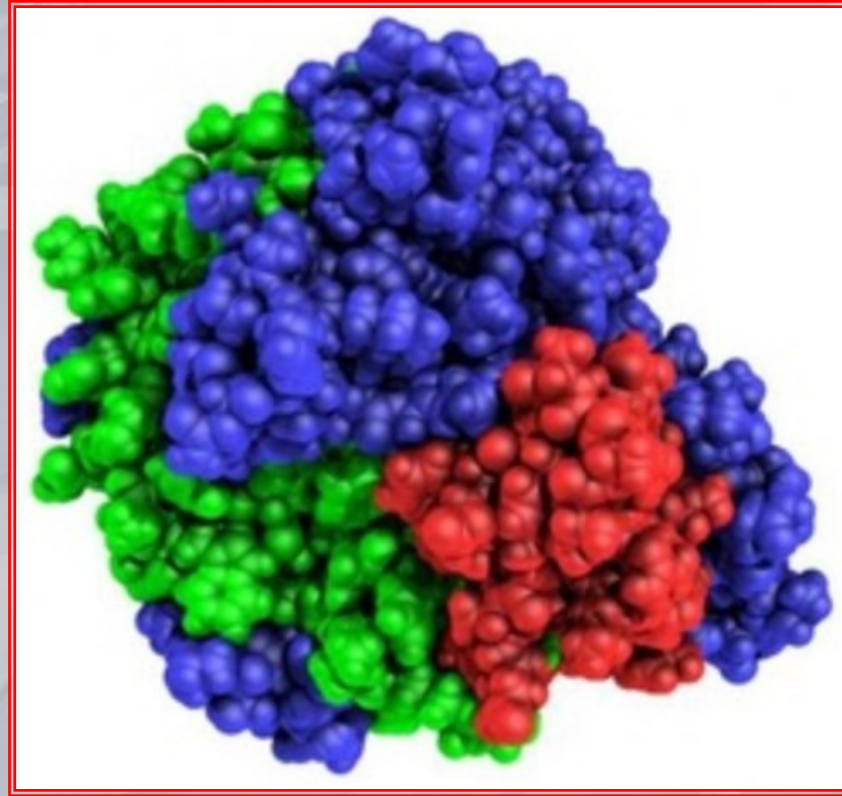
Classification of proteins

- ✦ Proteomic data indexing and cataloging are very difficult tasks, given the wide variety of different proteins, which are useful to the cell to carry out its tasks
- ✦ Several systemic methods proposed: the oldest one, due to the International Enzyme Commission, assigns each enzyme to one of six different categories based on its function
- ✦ Alternatively, classification methods based on the evolutionary history and on structural similarities, with about a thousand families of homologous proteins

Enzymes: An introduction – 1

- ✦ Enzymes are mainly globular proteins, that is, proteins with a round shape tertiary structure
- ✦ They act as catalysts, that is they speed up biological reactions
- ✦ They are highly specialized, i.e. they can act only on one specific type of substance
- ✦ Our body is a complex factory in which many aspects must be regulated, and enzymes are responsible for such a regulation

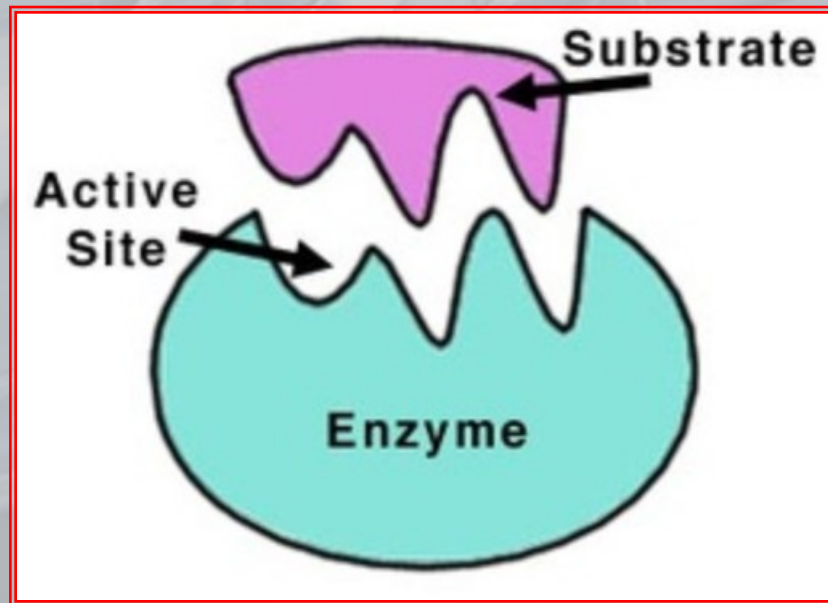
Enzymes: An introduction – 2



- ✦ The manufacture of complex tissues, such as skin and blood, are controlled by enzymes as well as the breaking down of chemicals to provide energy to the body

Enzymes: An introduction – 3

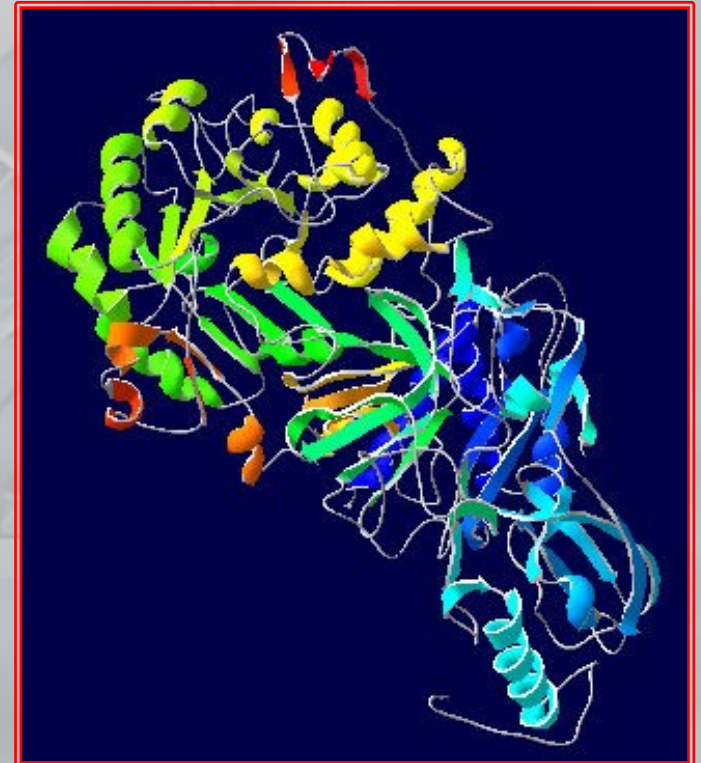
- ✦ Enzymes are larger than the **substrates**, i.e. molecules on which they act; they convert substrates into different molecules, known as **products**
- ✦ Enzymes bind substrates in the so-called **active site**, which is made by a cleft or pocket in the enzyme itself



A representation of an enzyme, its substrate and the active site

Enzyme functions – Example

- ✦ In all vertebrates, except for birds and some reptiles, urea is used to eliminate the nitrogenous products of metabolism from the body
- ✦ The enzyme urease, which is found in airborne bacteria, catalyzes urea molecules; one urease enzyme can catalyze 30,000 urea molecules per second
- ✦ Without the urease, the urea, to decompose by itself (in carbon dioxide and ammonia), would take about 3 million years
- ✦ In general, enzymes are highly specific, meaning that they will catalyze only reactions involving those molecules which can fit inside their active site



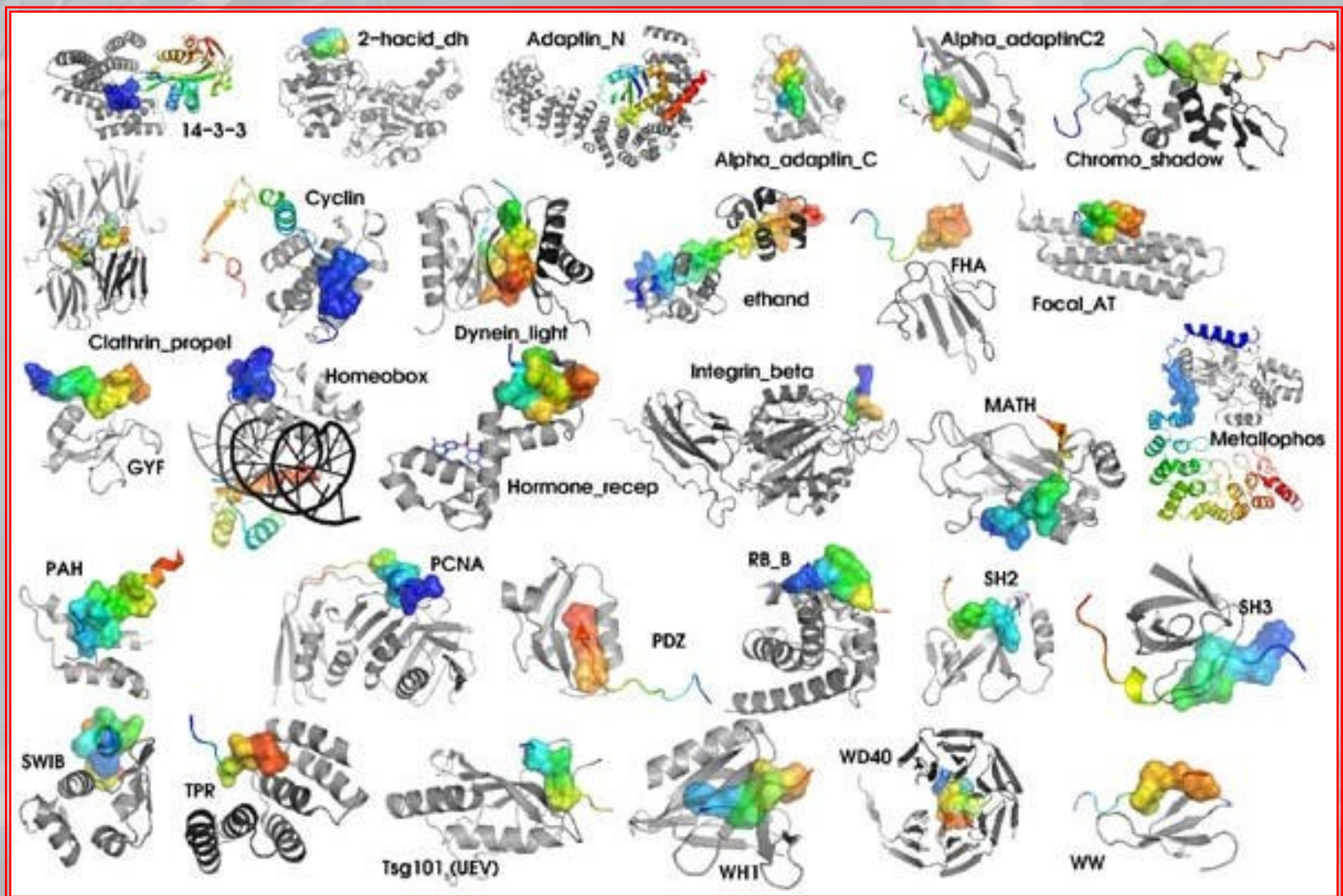
Environmental factors

- ✦ There are several factors which can affect the function of an enzyme
 - **Temperature:** Higher temperatures speed up reactions; however, the temperature should not exceed some limits otherwise the enzyme structure is damaged
 - **Ph:** Excessive acidity or basicity will cause the denaturation of the enzyme structure; in fact, the active site has an electric charge and, if the environment becomes too acid or basic, the charge may change
 - **Enzyme concentration:** Increasing the concentration of enzymes, with ample substrates and cofactors, increases linearly the rate of reaction
 - **Substrate concentration:** Increasing concentration of substrates with a fixed amount of enzyme and with ample cofactors, increases linearly the rate of reaction

Enzyme nomenclature – 1

- ✦ The rapid growth, during the '50s of the past century, of the number of known enzymes made it necessary to establish conventions regarding their nomenclature
- ✦ Before the *International Enzyme Commission* founding, in 1955, it was not unusual that a single enzyme was known by different names, or that the same name was assigned to different enzymes
- ✦ Also, some names did not give any indication of the nature of chemical reactions catalyzed by the related enzyme

Enzyme nomenclature – 2



Enzyme nomenclature – 3

- ✦ In 1965, a systematic approach was suggested to classify enzymes into six main classes, on the basis of the general types of reactions they catalyze
- ✦ Each enzyme is assigned a numerical code, where the first number refers to the main class, the second and the third numbers correspond to specific subclasses, and the final number is the serial number of the enzyme in its subclass
- ✦ Most enzyme names end in “ase” (some exceptions are pepsin, rennin, and trypsin)

Enzyme nomenclature – 4

No.	Class	Type of reaction catalysed	Examples
1.	Oxidoreductases	Transfer of electrons (hydride ions or H atoms)	Dehydrogenases, oxidases
2.	Transferases	Group transfer reactions	Transaminase, kinases
3.	Hydrolases	Hydrolysis reactions (transfer of functional groups to water)	Estrases, digestive enzymes
4.	Lyases	Addition of groups to double bonds or formation of double bonds by removal of groups	Phospho hexo isomerase, fumarase
5.	Isomerases	Transfer of groups within molecules to yield isomeric forms	Decarboxylases, aldolases
6.	Ligases	Formation of C–C, C–S, C–O, and C–N bonds by condensation reactions coupled to ATP cleavage	Citric acid synthetase

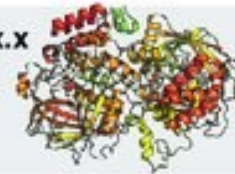
Enzyme nomenclature – 5

Redox



Oxidoreductases EC 1.x.x.x

Dehydrogenases
Hydrogenases
Oxidases/Oxygenases
Hydroxylases



Catalase

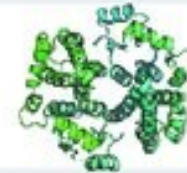
Single replacement



Transferases EC 2.x.x.x

Acyltransferases
Aminotransferases
Phosphotransferases

Glutathione
S-transferase



Double replacement/acid-base



Hydrolases EC 3.x.x.x

Esterases
Lipases
Phosphatases
Peptidases



6-Phosphogluconolactonase

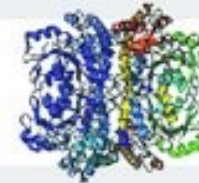
Decomposition



Lyases EC 4.x.x.x

Decarboxylases
Aldolases
Synthases

Cystathionine
gamma-lyase



Isomerisation



Isomerases 5.x.x.x

Razemases
Mutases



Triosephosphate
isomerase

Synthesis



Ligases 6.x.x.x

Synthetases
Carboxylases

Tryptophanyl-tRNA
synthetase



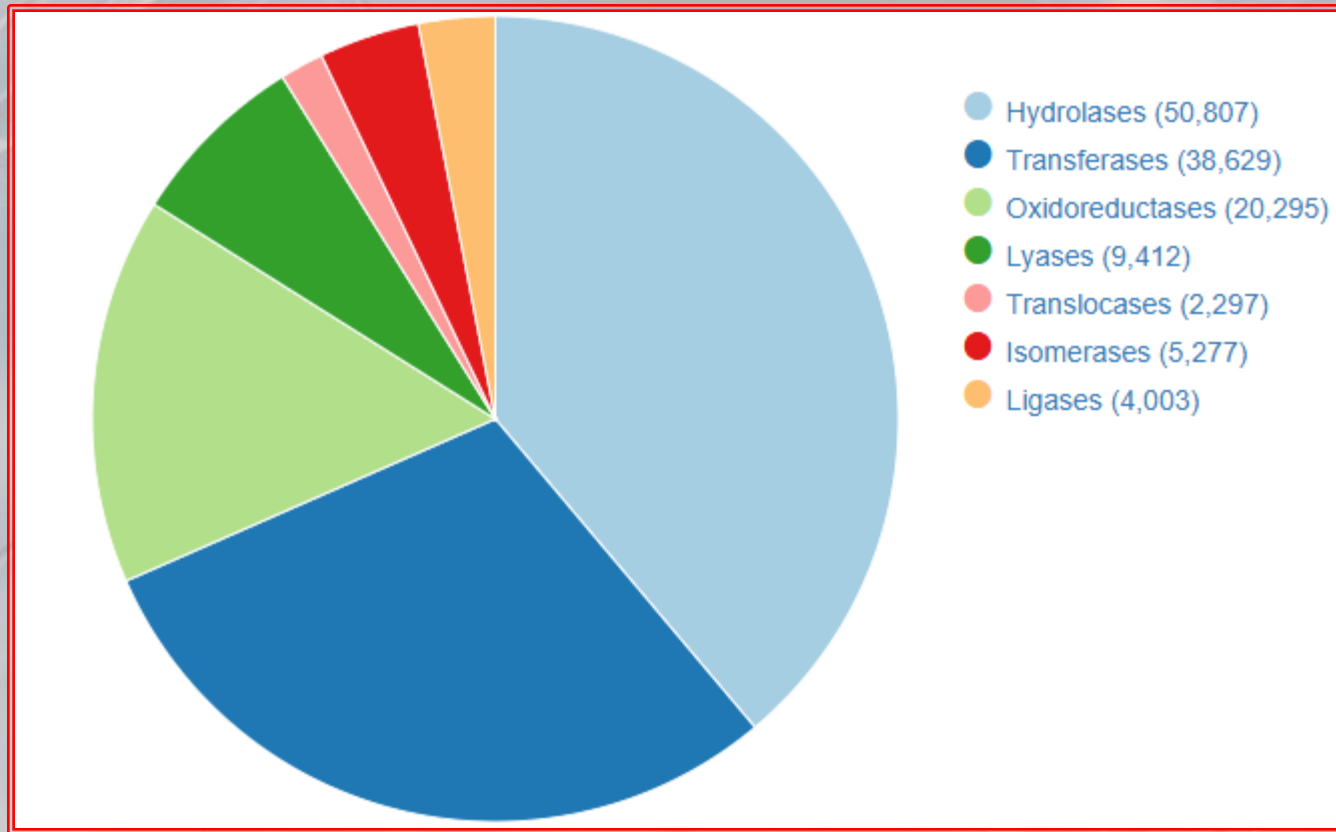
Current Opinion in Biotechnology

Enzyme nomenclature – 6

• Examples

- The **alcohol dehydrogenase** is identified as 1.1.1.1 — main class: oxidoreductase; class: activity on CH–OH group of the donor; subclass: with NAD or NADP (molecules that allow the oxide–reduction) as acceptor; it is the first of the original 269 enzymes present in this category (now they are 20295)
- The **RNA polymerase DNA–dependent** is identified by the number 2.7.7.6 — main class: transferase; class: transfer of phosphate groups; subclass: nucleotidil–transferase; it is the sixth of the original 60 enzymes found in this category (now they are 38629)

Enzyme nomenclature – 6



- More recently, a seventh class has been added, that of translocases, enzymes that move molecules and ions across membranes

Families and superfamilies – 1

- The amino acid sequence similarity, among the many thousands of proteins for which it is available, suggests that all the proteins existing today may derive from about 1000 original proteins
- It is unclear, however, if the limited number of existing proteins is dictated more by physical constraints on the polypeptide folding or by the sufficient variety of structural and chemical properties that they possess, or maybe by a combination of both factors

Families and superfamilies – 2

- ✦ One of the strongest arguments in favor of the hypothesis of evolution comes from a study published in 1991 by Dorit *et al.*, in which they theorized that exons themselves closely correspond to the protein functional domains, whereas all the proteins derive from various arrangements of the about 7000 available exons
- ✦ However, regardless of the basis of similarity, sequence alignment methods and similarity search in databases are often used to explore the possible relationships between different protein families
 - ▶ Useful to predict the protein structure, which seems to underlie to evolutionary constraints stronger than the amino acid sequence

Families and superfamilies – 3

- ✦ By definition, proteins that have a sequence identity greater than 50% are members of the same **family**
- ✦ Similarly, **superfamilies** are groups of protein families related with a lower, but still detectable, level of sequence similarity (30%)
 - ⇒ they share a common origin, even if more far in time
- ✦ All the proteins can be further grouped into **folds** and **classes**, on the basis of the predominant characteristics of their secondary structure: membrane proteins, mainly α proteins, mainly β proteins, α/β structures, $\alpha+\beta$ structures, etc.

Families and superfamilies – 4

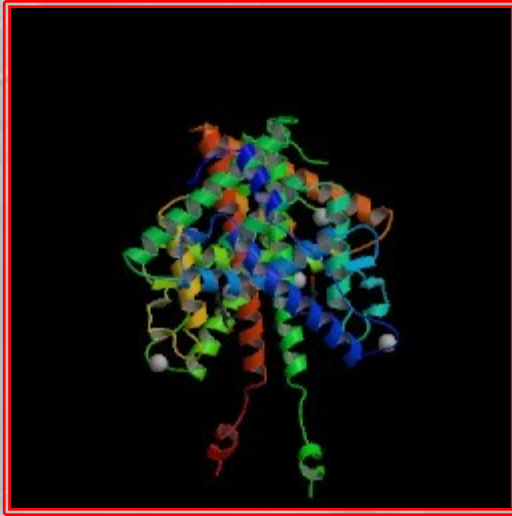
The four major classes of proteins

all α	composed mainly of α -helices (example: the four-helical bundle)
all β	composed mainly of β -sheets (example: antibodies and TCRs)
α/β	composed of α -helices and β -sheets that alternate along the chain
$\alpha + \beta$	composed of α -helices and β -strands that tend to segregate

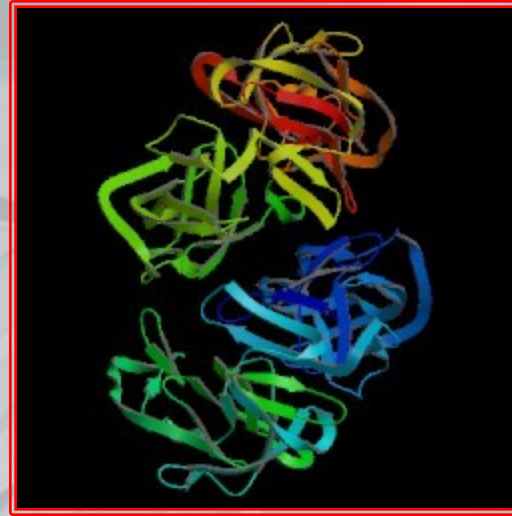
- ✦ Several hierarchical databases have been constructed, which group proteins according to particular features:
 - **SCOP** – Structural Classification Of Proteins
 - **CATH** – Class, Architecture, Topology and Homologous superfamilies
 - **FSSP** – Fold classification based on Structure–Structure alignment of Proteins

Families and superfamilies – 5

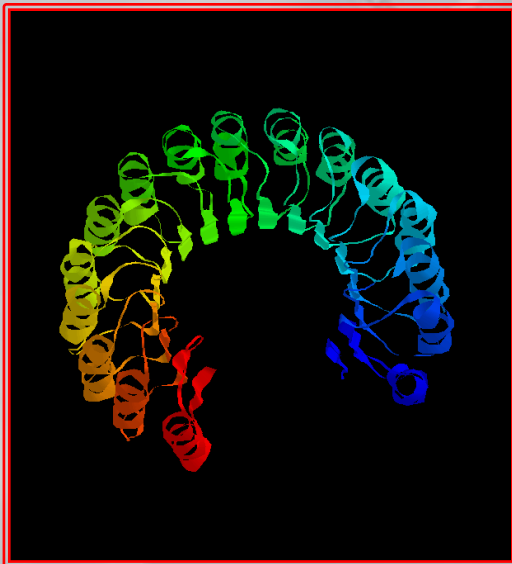
All α protein



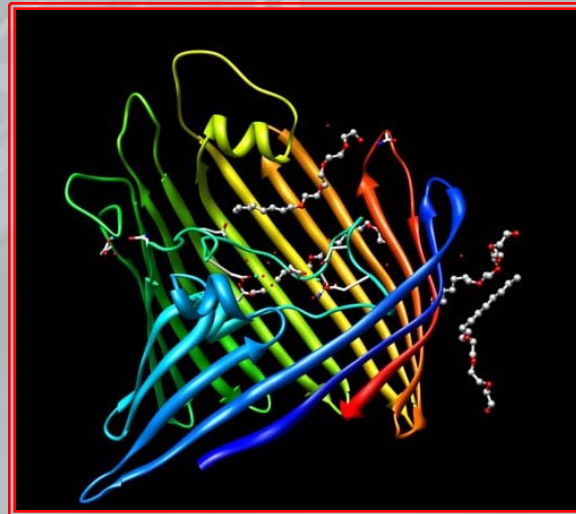
All β protein



α/β protein



$\alpha+\beta$ protein



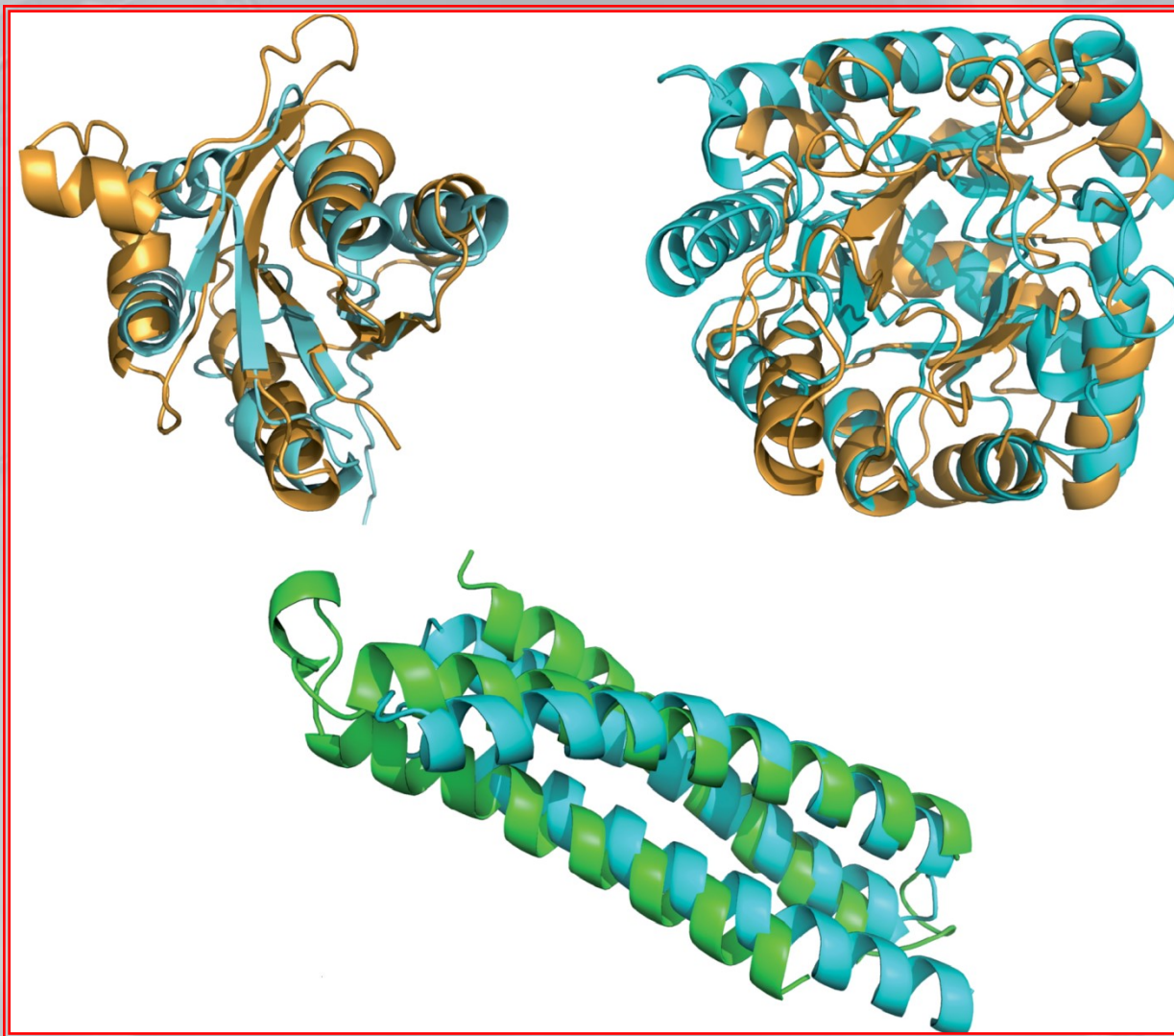
Folds – 1

- ✦ While protein families show clear evolutionary relationships and protein superfamilies have probable evolutionary relationships, proteins are said to share a common **fold** if they have the same secondary structure with the same kind of 3D arrangement and with the same topological connections
- ✦ The term fold is used as a synonym of **structural motif**, even if it generally refers to combinations of more extensive secondary structures – in some cases, a fold involves a half of the total protein structure
- ✦ Different proteins sharing the same fold often have peripheral elements of secondary structure and turn regions that differ both in size and conformation

Folds – 2

- ✦ The number of folds is limited: Nature has re-used the same folding types again and again for performing totally new functions
- ✦ Proteins belonging to the same fold category may also not have a common evolutionary origin, but may be the result of an exon shuffling, in which proteins with new functions are created through the process of recombination of exons corresponding to functional domains of existing genes
- ✦ Alternatively, structural similarities can arise only from physical and chemical characteristics of the proteins, which favor certain arrangements and certain chain topologies

Folds – 3















Example: SCOPe database – 1

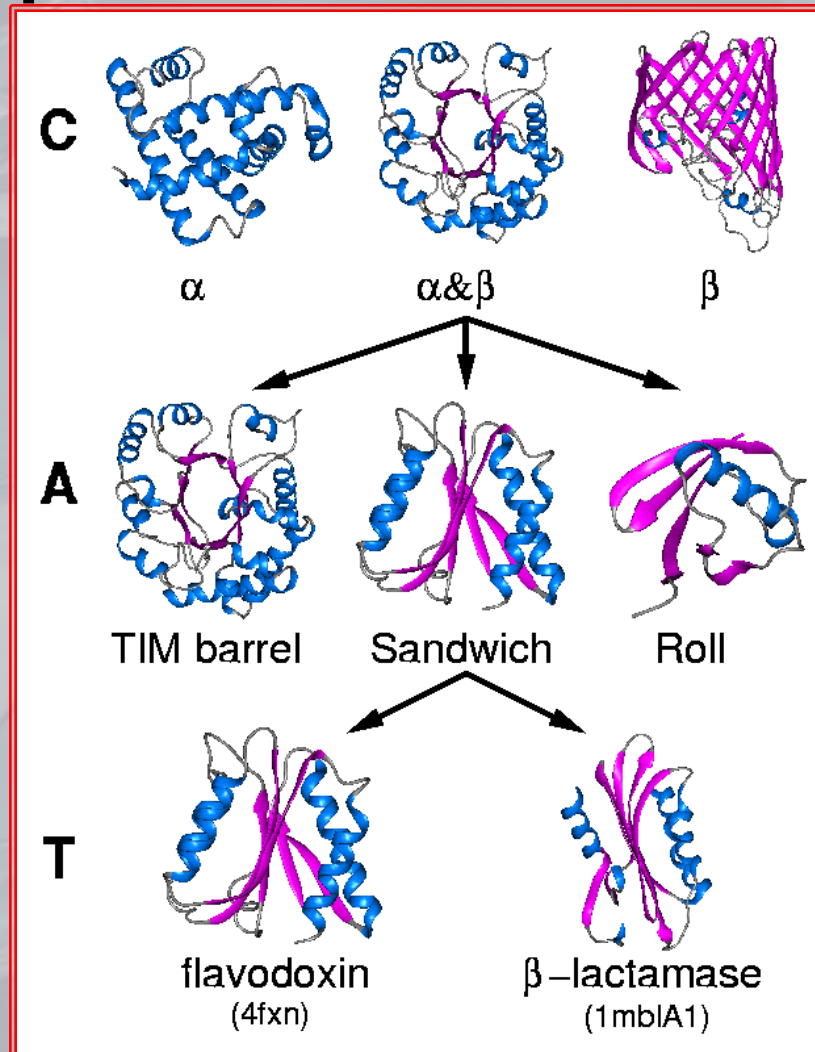
- ✦ **Class:** Proteins with similar secondary structure composition (ex.: all α -proteins, membrane and cell-surface proteins), but with different overall tertiary structures and evolutionary origins
- ✦ **Fold:** Different superfamilies having same major secondary structures in same arrangement and with analogous topological connections; similar tertiary structure, but not necessarily evolutionary relatedness
- ✦ **Superfamily:** Different families whose structural and functional features suggest common evolutionary origin
- ✦ **Family:** Evolutionary related proteins with a significant sequence identity

Example: SCOPe database – 2

Classes in SCOPe 2.08:

1.  a: All alpha proteins [46456] (290 folds)
2.  b: All beta proteins [48724] (180 folds)
3.  c: Alpha and beta proteins (a/b) [51349] (148 folds)
4.  d: Alpha and beta proteins (a+b) [53931] (396 folds)
5.  e: Multi-domain proteins (alpha and beta) [56572] (74 folds)
6.  f: Membrane and cell surface proteins and peptides [56835] (69 folds)
7.  g: Small proteins [56992] (100 folds)
8.  h: Coiled coil proteins [57942] (7 folds)
9.  i: Low resolution protein structures [58117] (25 folds)
10.  j: Peptides [58231] (151 folds)
11.  k: Designed proteins [58788] (44 folds)
12.  l: Artifacts [310555] (1 fold)

Example: CATH database



Schematic representation of the (C)lass, (A)rchitecture and (T)opology/fold levels in the CATH database

Experimental techniques

- ✦ As in the case of genomic analysis, also many proteomic analyses are limited by the currently available experimental techniques
- ✦ Unfortunately, from the perspective of proteomics, the very nature of proteins makes the laboratory analyses particularly difficult and much less precise than those available for the genome
- ✦ In fact, the generation of gene copies can be easily achieved with PCR, in an efficient, controlled and cell-free environment; instead, obtaining a huge quantity of proteins, usable for analysis, requires that they be chemically isolated, in an inefficient and laborious way, from a large number of living cells
 - Two-dimensional electrophoresis
 - Mass spectrometry
 - Protein microarray

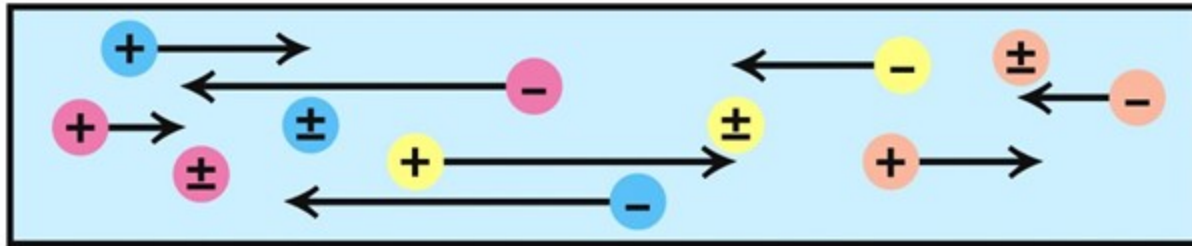
Two-dimensional electrophoresis – 1

- ✦ The **two-dimensional electrophoresis** is a technique that allows the separation of proteins according to their molecular weight and charge
- ✦ This process starts from the extraction of proteins from a tissue
- ✦ Proteins, placed on a polymeric support strip and in presence of an immobilized gradient of acidity, migrate according to their intrinsic electrical charge, reaching their isoelectric point (the isoelectric point is the pH at which a particular molecule carries no net electrical charge) and forming some “bands”
- ➡ This “first dimension” is called **isoelectric focusing** (IEF)

Two-dimensional electrophoresis – 2

(A)

Low pH
(+)



High pH
(-)

(B)

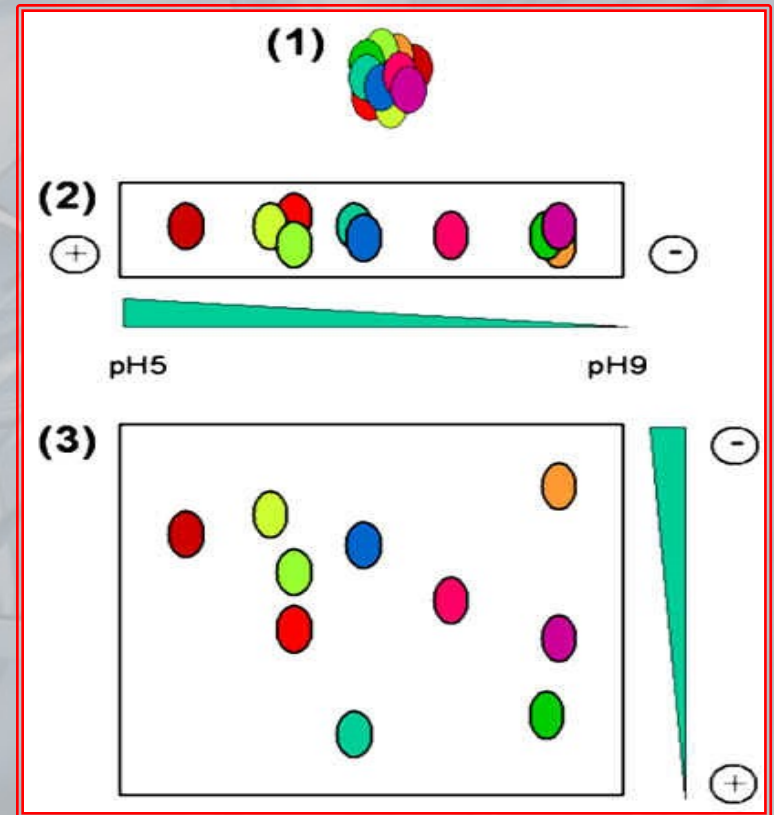
Low pH
(+)



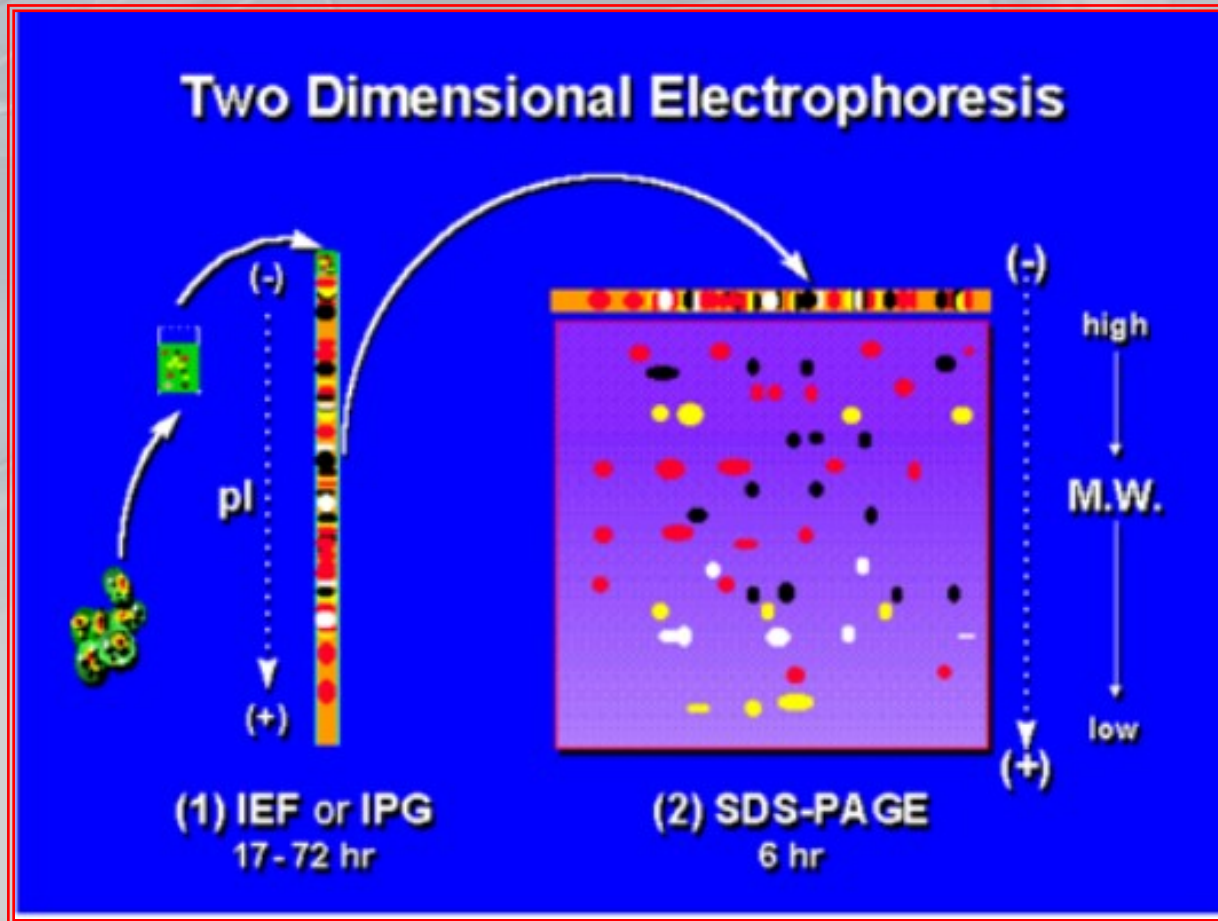
High pH
(-)

Two-dimensional electrophoresis – 3

- ✦ At this point, the support is placed on the margin of an electrophoresis gel that allows the separation of proteins according to their molecular weight, based on the application of an electric field
- ✦ In fact the gel is composed of polyacrylamide and sodium dodecyl sulfate (SDS), a detergent that binds uniformly to all proteins and gives them a negative charge



Two-dimensional electrophoresis – 4

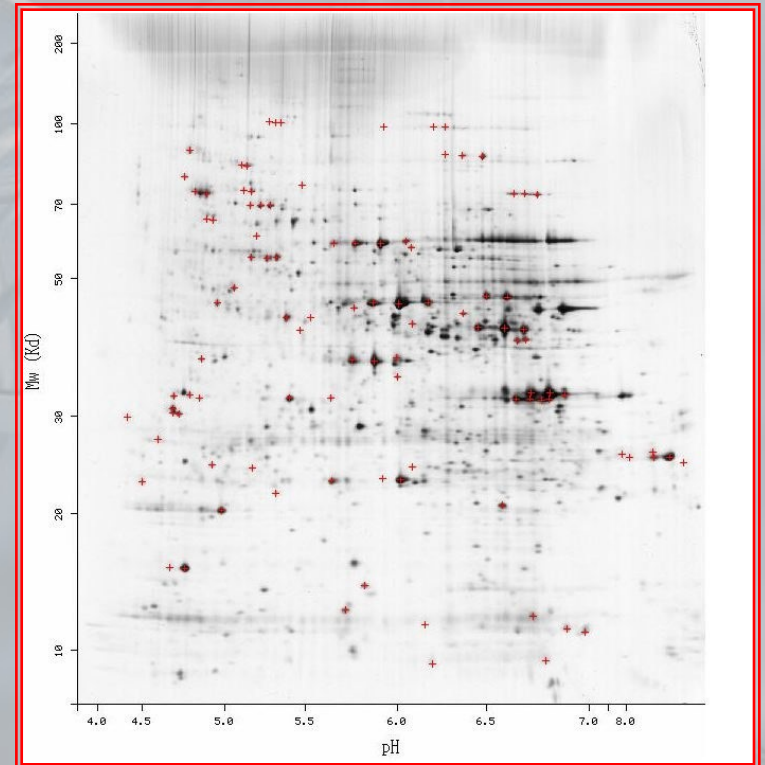


Two-dimensional electrophoresis – 5

- ✦ The final result is a gel in which each protein virtually occupies a point in the two-dimensional space; therefore, it is easily detectable by appropriate intensities or colors
- ✦ The last step consists in isolating each protein from the gel, to carry out the analysis that allows its identification
- ✦ Such analysis can be achieved either manually, by cutting a small piece of gel containing a single protein, and then proceeding with mass spectrometry, or through automatic techniques able to “read” directly from the gel

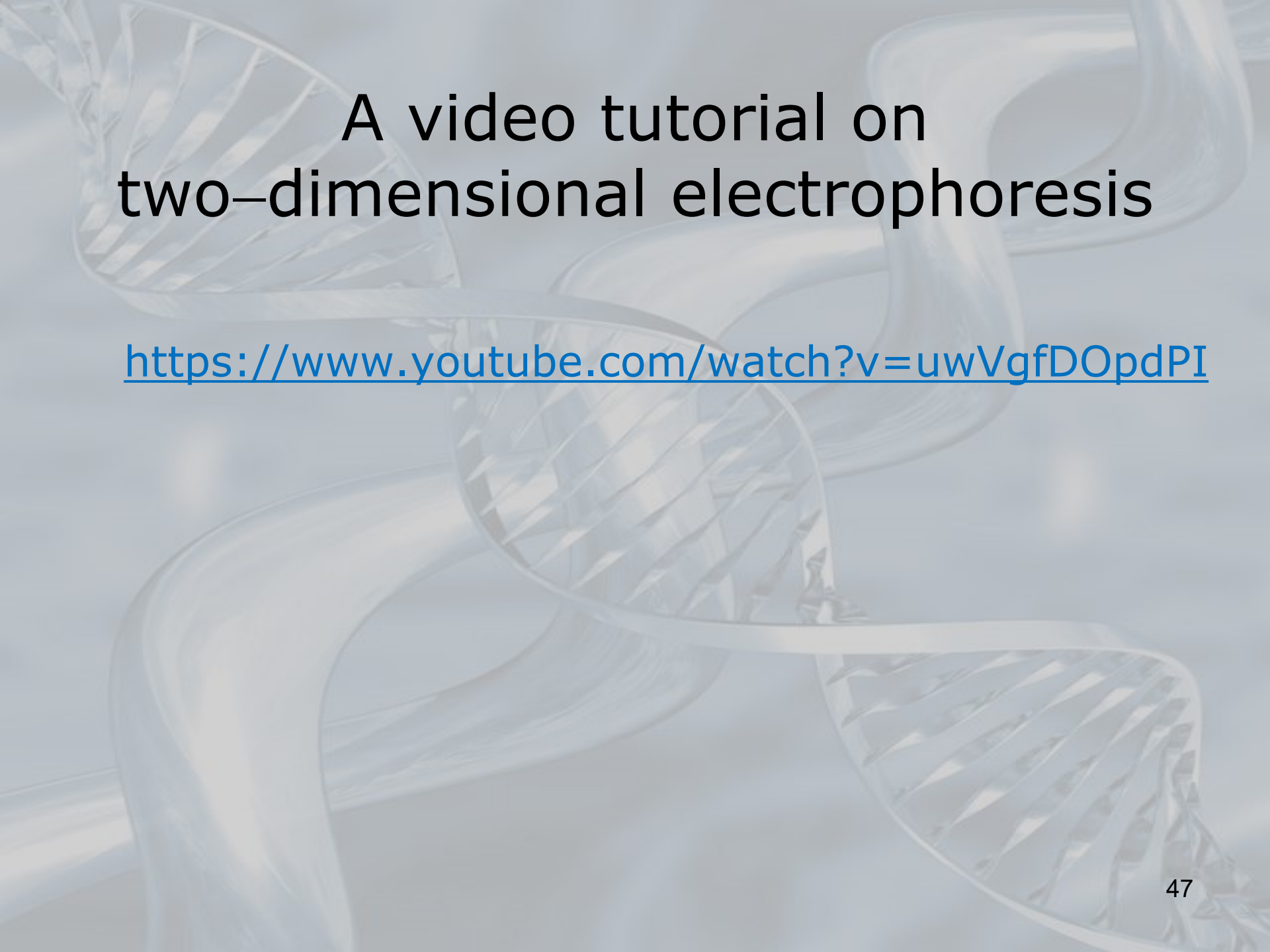
Two-dimensional electrophoresis – 6

- ✦ In practice, a cell or a tissue are considered and, using the described techniques, information on all the proteins that make up the sample are obtained, with just a single analysis
- ✦ With the 2D electrophoresis, images are obtained in which each dot represents a protein
- ✦ Comparing photos obtained from different samples, proteins that differ both for their presence or for their quantities can be evidenced, w.r.t. various experimental conditions – different tissues, normal or pathological conditions, development stages



Two-dimensional electrophoresis – Some limitations

- ✦ The 2D electrophoresis actually has several serious limitations that prevent its extensive use
 - The human genome encodes many tens of thousands of proteins
 - The 2D electrophoresis is inadequate for the analysis of proteins that are very small or endowed of a little electrical charge, such as intramembrane plasma proteins (which play an important role in many diseases), due to their poor solubility in preparations and gels
 - Relatively low sensitivity of detection methods
 - Difficulty in precisely determining which protein is represented by each spot
 - Anyway, up to 20000 proteins can be found and analyzed in a single tissue electrophoresis



A video tutorial on two-dimensional electrophoresis

<https://www.youtube.com/watch?v=uwVgfDOpdPI>

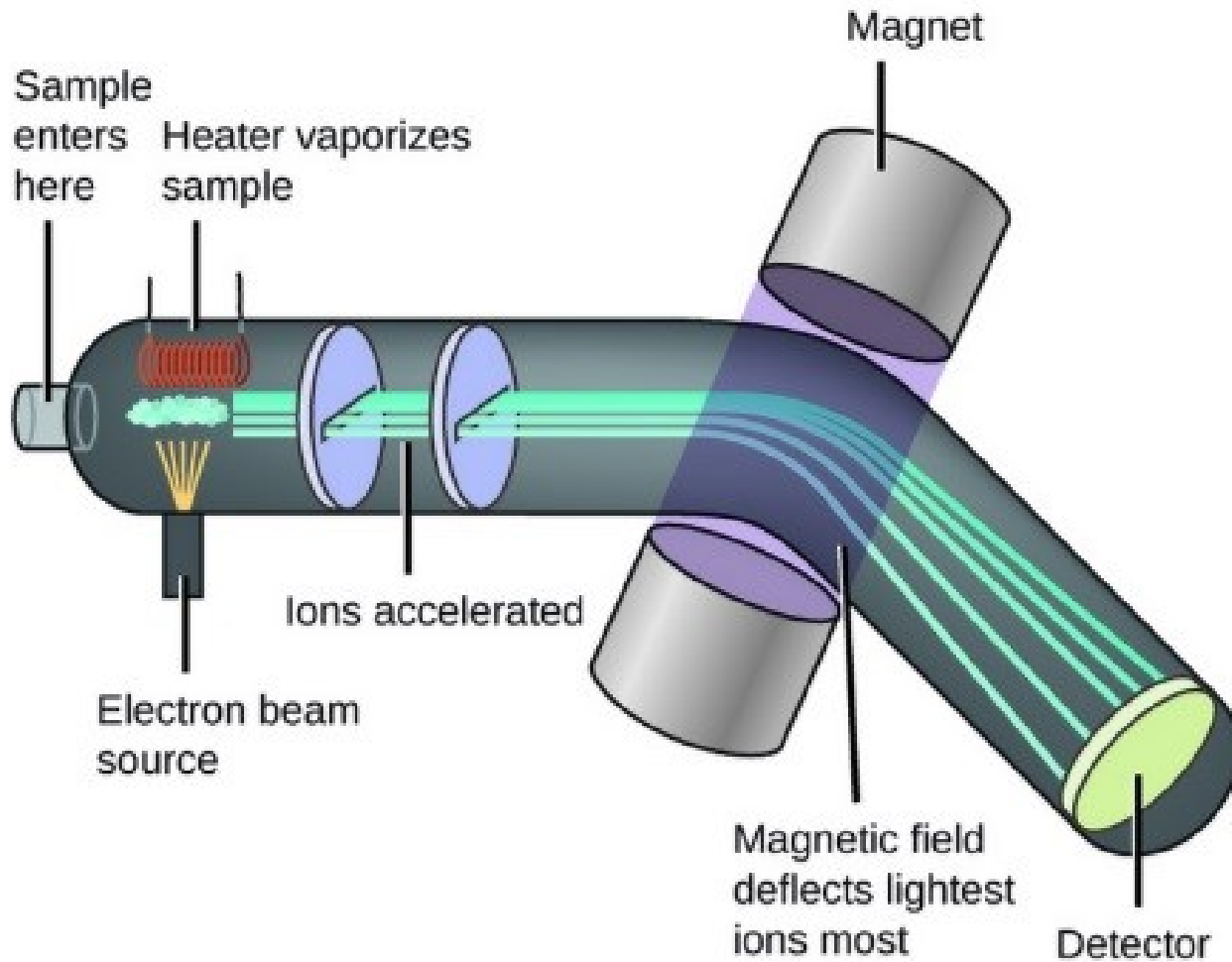
Mass spectrometry – 1

- ✦ **Spectrometers** are instruments able to split the spectrum of a source and measure its components
- ✦ There are spectrometers which measure the spectrum of an electromagnetic radiation and spectrometers which measure the mass spectrum of a substance, that is the masses of its constituents (atoms, molecules, compounds)
- ✦ In a mass spectrometer, samples may be introduced in the solid, liquid or gaseous state
- ✦ The solid or liquid substances must be made volatile before starting the ionization phase, during which the molecules of the compound are ionized, in the most common case by the interaction with an electron beam

Mass spectrometry – 2

- ✦ This cause some of the molecules to break into charged fragments or simply become charged without fragmenting
 - Only the ions are revealed by the spectrometer, and are separated according to their mass/charge ratio
 - In fact, the paths of the protein fragments (within an analyzer) are deviated by a magnetic field
 - The collision of the ions with a collector, placed at the end of the analyzer, generates an electrical current, that can be amplified and detected as a series of peaks, corresponding to a **fingerprint of the peptide mass** (mass fingerprint)

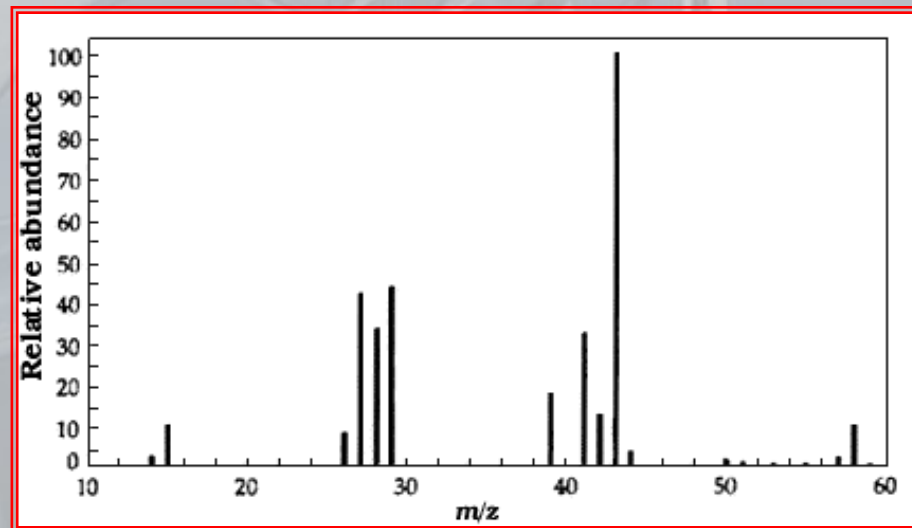
Mass spectrometry – 3

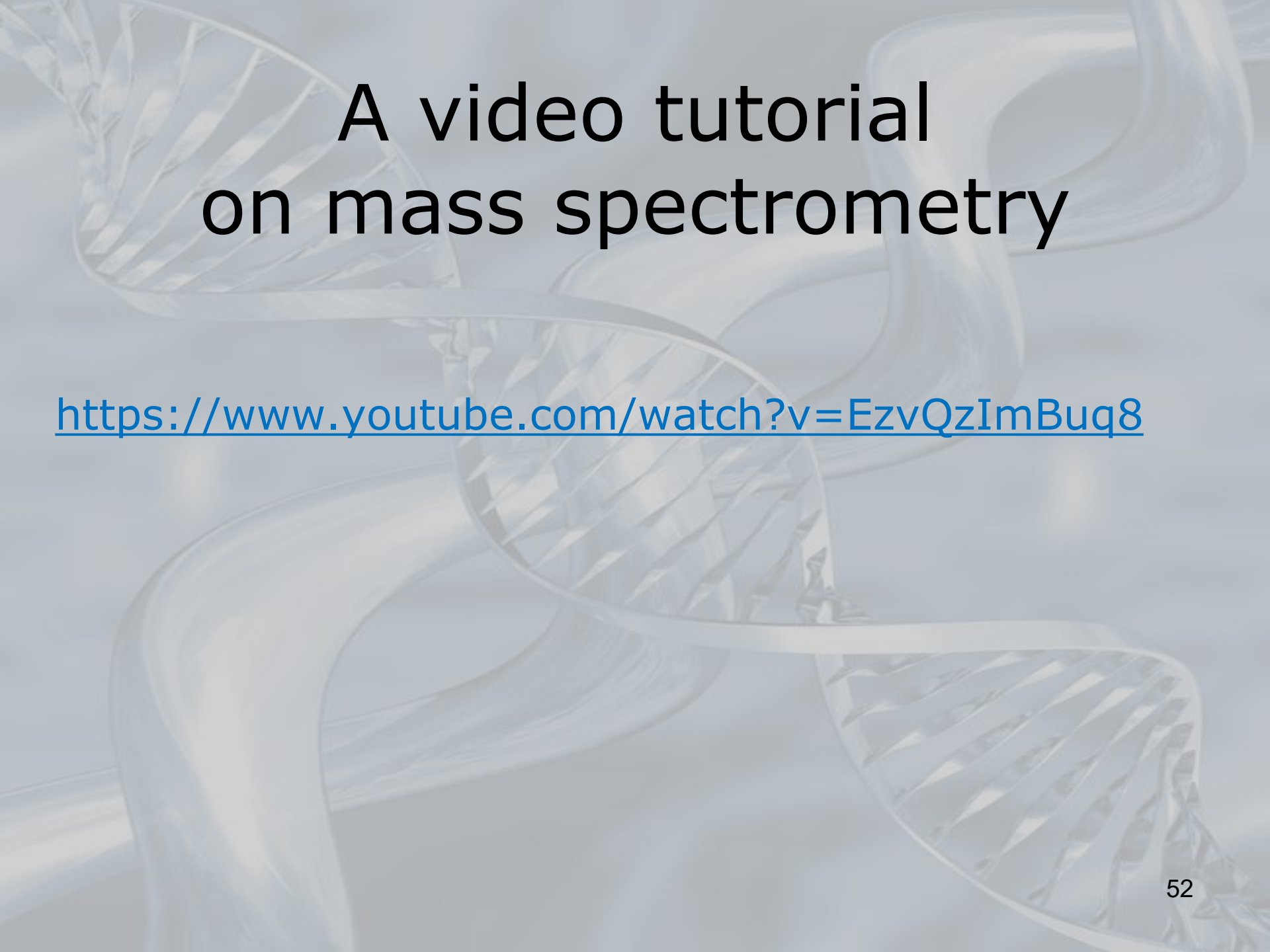


Mass spectrometer

Mass spectrometry – 4

- ✦ In a mass spectrum graph, the x-axis shows the values of the mass/charge ratio, whereas on the y-axis, the relative abundance of the analyzed ions is reported
- ✦ If the instrument resolution is high enough, the exact mass of each individual ion can be determined, from which the composition of the ions can be deduced





A video tutorial on mass spectrometry

<https://www.youtube.com/watch?v=EzvQzImBuq8>

Protein microarrays – 1

- ✦ Protein microarrays or protein chips are high throughput methods, where many proteins in parallel can be tracked and analysed
- ✦ Protein microarrays are used in five main fields:
 - Proteomics: detecting protein expression profiles
 - Protein functional analysis and protein–protein interaction
 - Detection of antigens and antibodies in blood samples
 - Antibody characterisation
 - Antigen–specific therapies and treatments

Protein microarrays – 2

- ✦ **Protein microarrays** are widely used because of their ability to perform protein analysis on a large scale, in the same way in which genetic chips have revolutionized the transcriptome analysis
- ✦ The basic concept of the protein chips is very similar to that of the gene chips: small amounts of individual probes are covalently linked to the surface of the silicon chip in a *high-density* array
- ✦ Proteins extracted from the cells are labeled with fluorophores and flushed on the chip
 - Just as with gene chips, the amount of material (in this case, protein) bound to the probes is determined by the excitation of the fluorophore

Protein microarrays – 3

- ✦ Apart from microarrays able to detect protein–protein interactions and protein–compound interactions, also arrays of capture probes (for example, antibodies) can be used – that bind to proteins of a sample so as to evaluate the relative expression levels
- ✦ Protein microarrays do not have the same impact of gene chips
 - Unlike DNA sequences, with their unique bonds dictated by base coupling, it is reasonable to expect that a single protein can interact with multiple different probes
 - The binding kinetics of each probe may vary, and differences in the intensity of the signal could be due to differences in the intensity of the binding
 - Proteins are known to be sensitive to the chemistry of their environment and to the surface they encounter, and both target proteins and probes may result in an unexpected behavior when subjected to control procedures

Protein microarrays – 4

- For the time being (and waiting for reliable automatic analysis techniques), it is easier to use gene chips as a base, to point to the study of the proteins of interest
 - After this preliminary phase, the proteomic analysis of small subsets of proteins will be performed on *ad hoc* protein chips

<https://www.youtube.com/watch?v=9TYgNWyENRE>

Inhibitors and drug design – 1

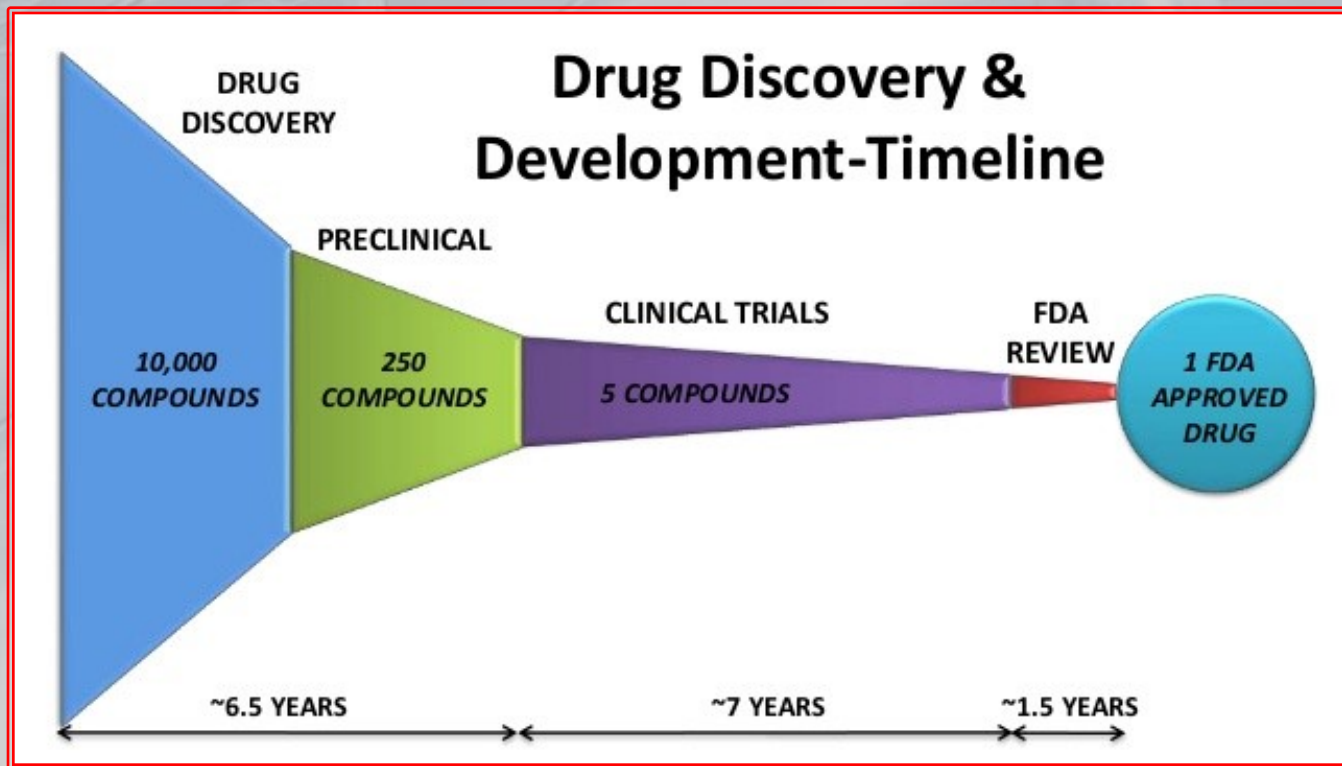
- One of the most important applications of bio-informatics is represented by the search for effective pharmaceutical agents to prevent and treat human diseases
- The development and testing of a new drug is expensive both in terms of time – often takes up to 15 years –, and of money – with a cost of hundreds of millions of euros

Inhibitors and drug design – 2

- ✦ Functional genomics, bioinformatics and proteomics promise to reduce the work involved in this process, accelerating time and lowering costs for developing new drugs



Drug discovery and development pipeline



Discovery and development

Some insights

- ✦ Typically, researchers discover new drugs through:
 - New insights into a disease process that allow them to design a product to stop or reverse the effects of the disease
 - Many tests of molecular compounds to find possible beneficial effects against any of a large number of diseases
 - Existing treatments that have unanticipated effects
 - New technologies, such as those that provide new ways to target medical products to specific sites within the body or to manipulate genetic material

Drug design – 1

- ✦ While the exact phases of the development of a drug are variable, the overall process is divided into two basic steps: discovery and testing
- ✦ The discovery process, which is rather laborious and expensive and provides a breeding ground for bio-informatics, can again be divided into three main steps
 - Target identification
 - Discovery and optimization of a “lead” compound
 - Toxicology and pharmacokinetics (which quantitatively studies absorption, distribution, metabolism and elimination of drugs)
- ✦ The testing process, which involves pre-clinical and clinical tests and trials, is generally not subject to significant improvements with the use of automated methods

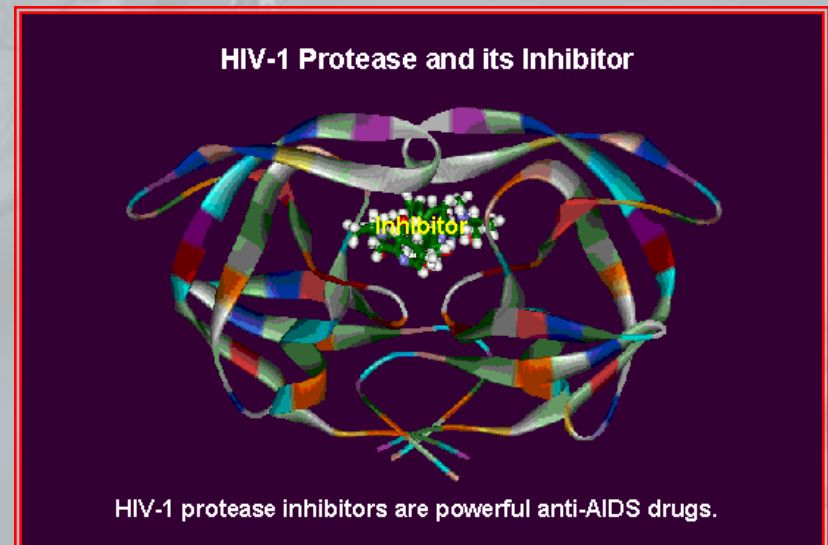
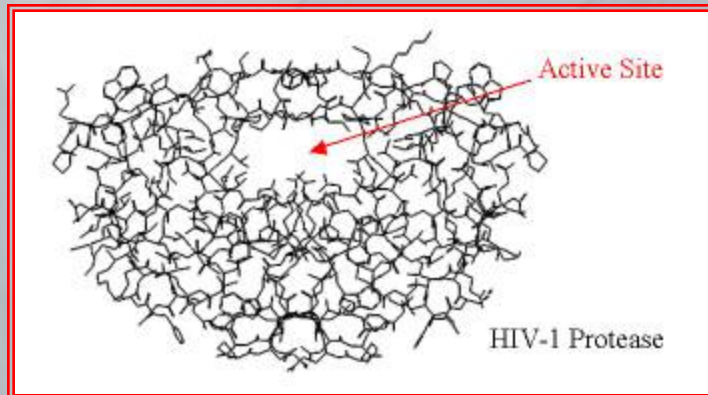
Drug design – 2

- ✦ **Target identification** consists of isolating a biological molecule that is essential for the survival or the proliferation of a particular agent, responsible for a disease and called **pathogen**
- ✦ After the target identification, the objective of drug design is the development of a molecule that binds to the target and inhibits its activity
- ✦ Given that the function of the target is essential for the vital process of the pathogen, its inhibition stops the proliferation of the pathogen or even destroys it
- ➡ Understanding the structure and the function of proteins is a key component in the development of new drugs, since proteins are common targets for drugs

Drug design: An Example – 1

✦ HIV Virus

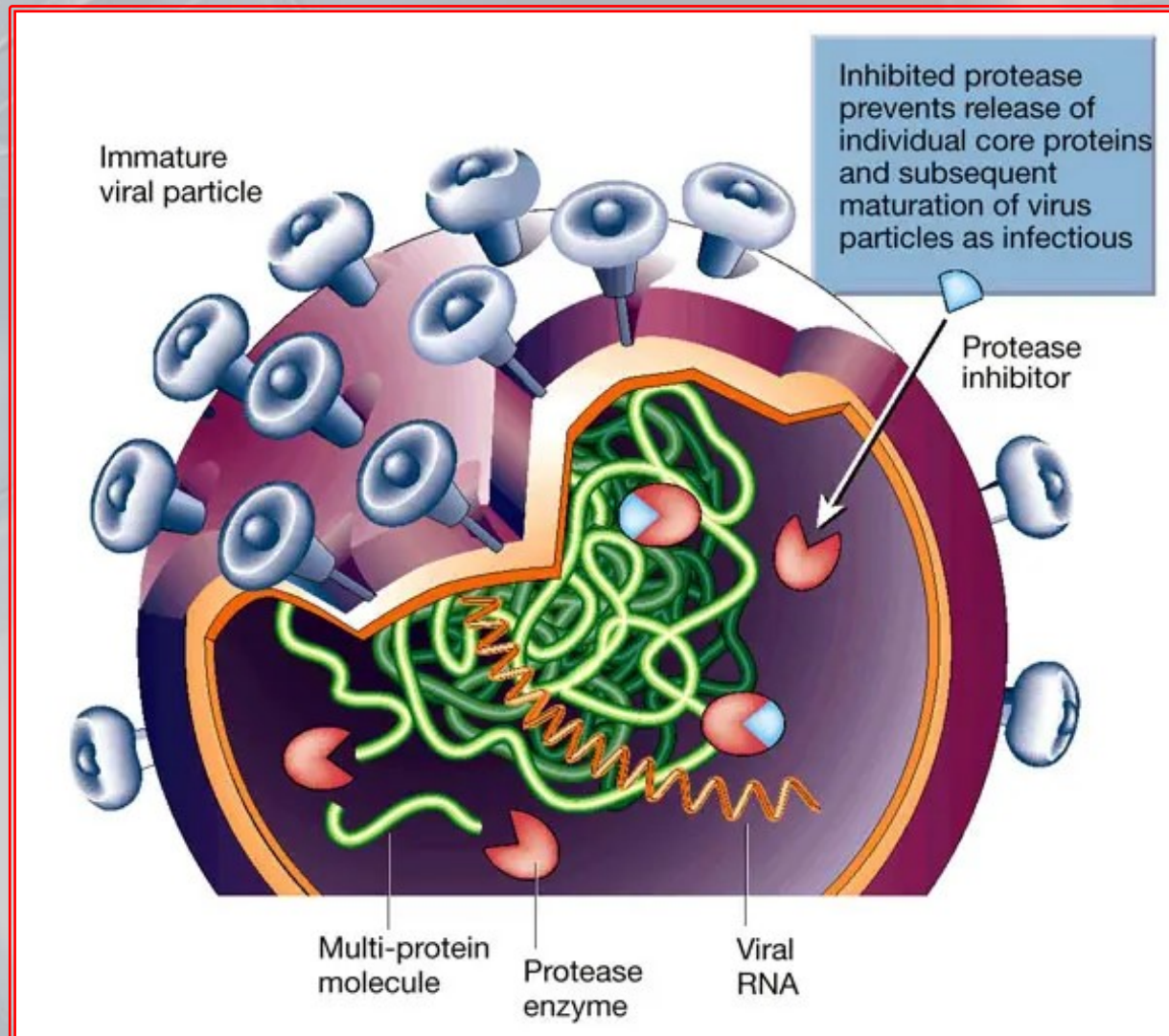
- The **HIV protease** is a protein produced by the human immunodeficiency virus (HIV), the pathogen that causes AIDS, in the context of a human cell host
- It is a small enzyme made up of two identical protein chains, each only 99 amino acids long; the two chains are joined to form a long tunnel that crosses the molecule; the tunnel is covered by two flexible flaps
- The HIV protease is essential for the virus proliferation: the inhibition of such protein destroys the effectiveness of the virus and its transmission capacity



Drug design: An Example – 2

- ✦ How can a molecule inhibit the action of an enzyme, such as the HIV protease?
 - Proteases are proteins that digest other proteins, such as restriction enzymes used to cut the DNA molecule in a specific way
 - Many of the proteins that HIV needs to survive and proliferate in a human host are codified as a single long polypeptide chain
 - This polypeptide must then be cut into the functional protein components by the HIV protease
 - Like many other enzymes, the HIV protease has an active site, to which other molecules can bind and “operate”
 - ➡ Design a molecule that binds to the active site of the HIV protease, so as to prevent its normal operation

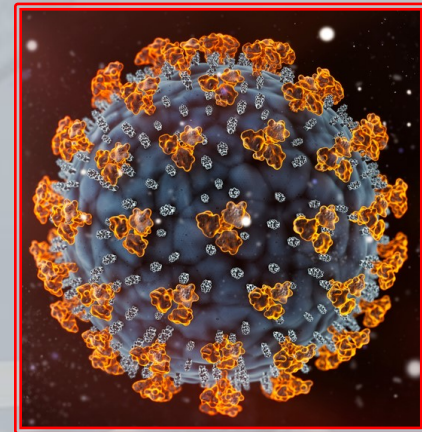
Drug design: An Example – 3



A possible strategy to fight COVID-19 – 1

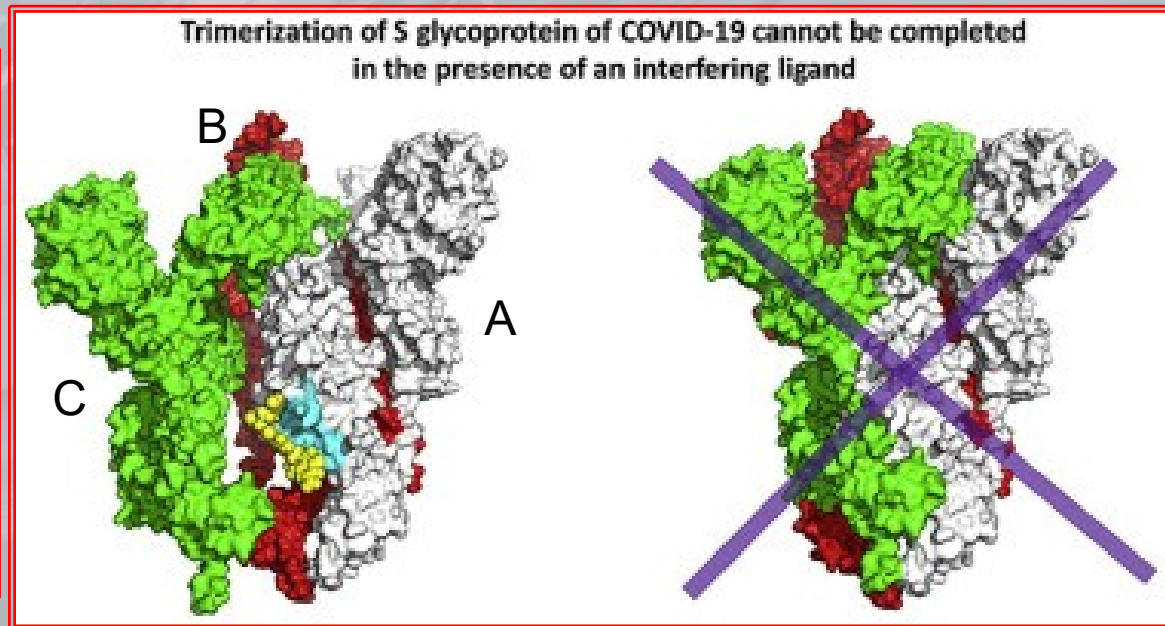
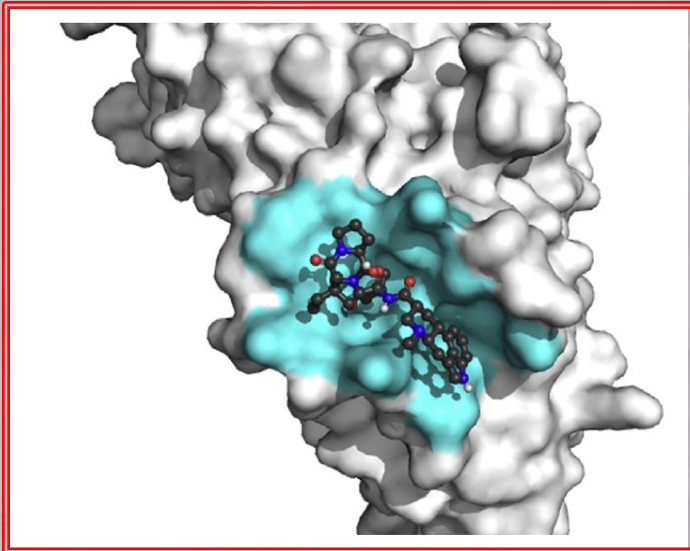
✦ COVID-19

- The S glycoprotein is central for COVID-19 infection as it mediates attachment of virions to the host cell receptor
 - Assembly of protomers into the bioactive form of S glycoprotein trimeric structure has been experimentally proved to be the rate-limiting step for transmissible gastro-enteritis coronavirus (TGEV) infecting cycle
 - By comparing the S glycoprotein sequences of COVID-19 and TGEV coronaviruses, their similarity was evidenced
- ⇒ Search for superficial pockets at the interface between two forming protomers, in which a small molecule can be inserted to prevent the quaternary structure formation (trimerization)



A possible strategy to fight COVID-19 – 2

- Eight small molecules were obtained (two of which are nutraceutical) able to fit in a pocket of protomer A to avoid its interaction with protomer C



A video on drug design

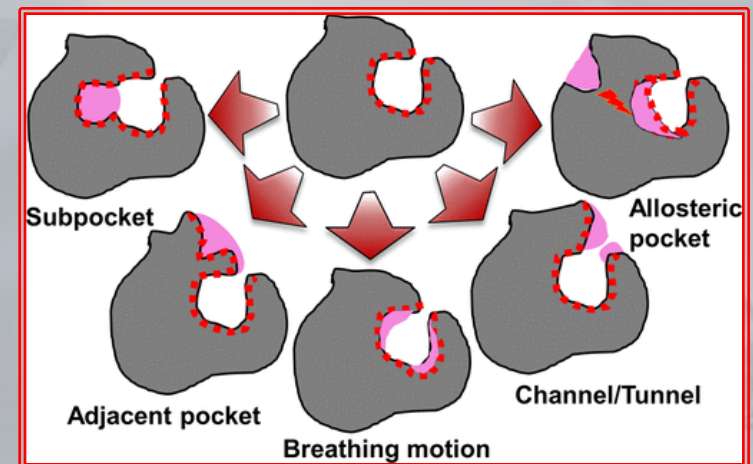
- ✦ Some more insights on inhibitors and drug design
<https://www.youtube.com/watch?v=sENZ6KCYDjU>

Ligand screening – 1

- ✦ The first step toward the discovery of an **inhibitor** for a particular protein is usually the identification of one or more **lead compounds**, which bind to the active site of the target protein
- ✦ Traditionally, the search for lead compounds has always been a *trial-and-error* process, during which several molecules were tested, until a sufficient number of compounds with inhibitory effects were found
- ✦ Recently, methods for high throughput **screening** (HTS) have made this procedure much more efficient, even if the underlying process is still an exhaustive search of the greatest possible number of lead compounds

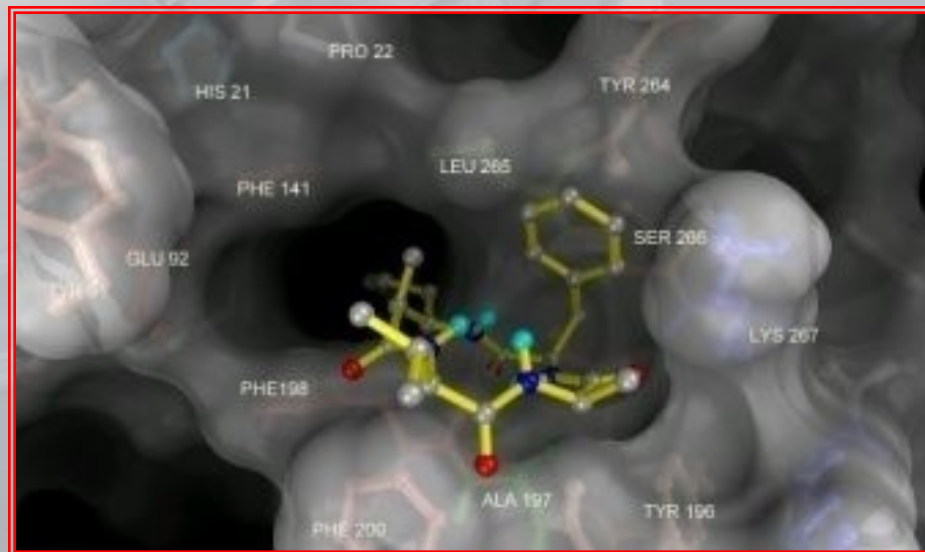
Ligand screening – 2

- ✦ The active sites of enzymes are housed in pockets (cavities) formed on the protein surface, with specific physico-chemical characteristics
- ✦ The protein–ligand interaction is dictated mainly by the complementary nature of the two compounds: hydrophobic ligands will bind hydrophobic regions, charged ligands will be recalled from charged regions of opposite sign, etc.
- ✦ **Docking** and **screening** algorithms for ligands try to produce a more efficient discovery process, moving from the world of *in vitro* testing to that of defining abstract models that can be automatically evaluated



Ligand docking – 1

- ✦ **Docking** is just the silicon simulation of the binding of a ligand with a protein, or, in other words, the docking aim is to determine how two molecules of known structure can interact
 - Surface geometry
 - Interactions between related atoms
 - Electrostatic force fields



Ligand docking – 2

- ✦ In many cases, the three-dimensional structure of a protein and of its ligands are known, but the structure of the formed compound is unknown
 - In drug design, molecular docking is used to determine how a particular drug binds to a target or to understand how two proteins can interact with each other to form a compound
- ✦ Molecular docking approaches have much in common with protein folding algorithms
 - Both problems involve calculating the energy of a particular molecular conformation and searching for the conformation that minimizes the free energy of the system
 - ⇒ Many degrees of freedom: heuristic searches and suboptimal solutions

Ligand docking – 3

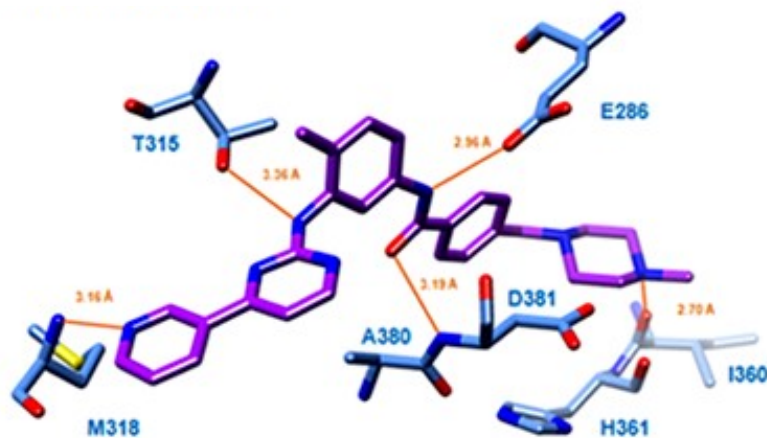
- ✦ As in protein folding, there are two main considerations to take into account when designing a docking algorithm
 - Define an energy function for evaluating the quality of a particular conformation and, subsequently, use an algorithm for exploring the space of all possible binding conformations to search for a structure with minimum energy
 - Managing the flexibility of both the protein and the putative ligand
 - ✗ The **key-lock approach** assumes a rigid protein structure which binds to a ligand with a flexible structure (a computationally advantageous approach)
 - ✗ The **induced fit docking** allows flexibility of both the protein and the ligand
 - ✗ Compromise: assuming a rigid protein backbone, while allowing the flexibility of the side chains near the binding site

Ligand docking – 4

Experimental conformation



Best predicted conformation

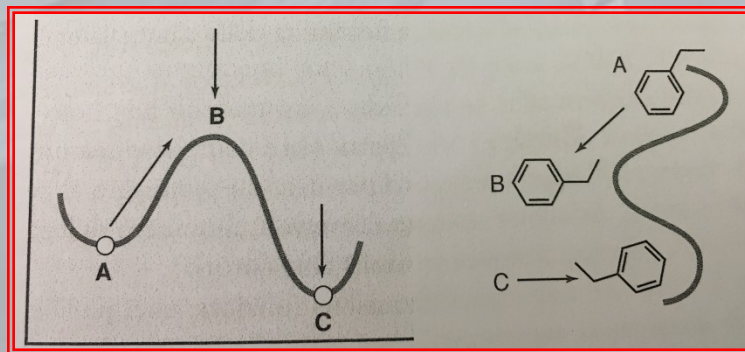


Ligand docking – 5

- ✦ **AutoDock** (<https://autodock.scripps.edu/>) is a well-known method for docking of both flexible and rigid macromolecules
 - It uses a force field based on a grid in order to evaluate a particular conformation
 - The force field is used to give a score to the resulting conformation, according to the formation of favorable electrostatic interactions, the number of established hydrogen bonds, van der Waals interactions, etc.

Ligand docking – 6

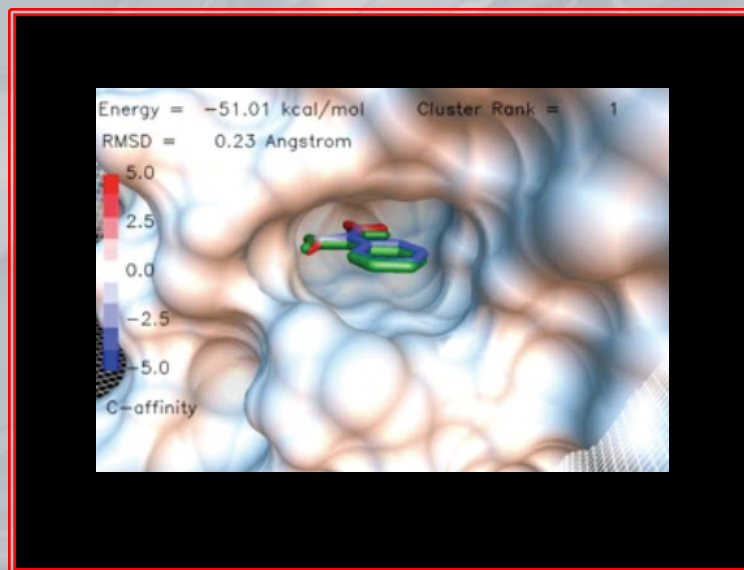
- ✦ AutoDock originally used a Monte Carlo/Simulated annealing approach
 - Random changes are induced in the current position and conformation of the ligand, keeping those that give rise to lower energy arrangements (w.r.t. the current one); when a change leads to an increase in energy, it is discarded
 - However, in order to allow the algorithm to find low-energy states, overcoming energy barriers, such changes that lead to higher energies are sometimes accepted (with a high frequency at the beginning of the optimization process, which slowly decreases for subsequent iterations)



Ligand docking – 7

- Recent AutoDock releases are equipped also with genetic algorithms, optimization programs that emulate the dynamics of natural selection in a population of competing solutions

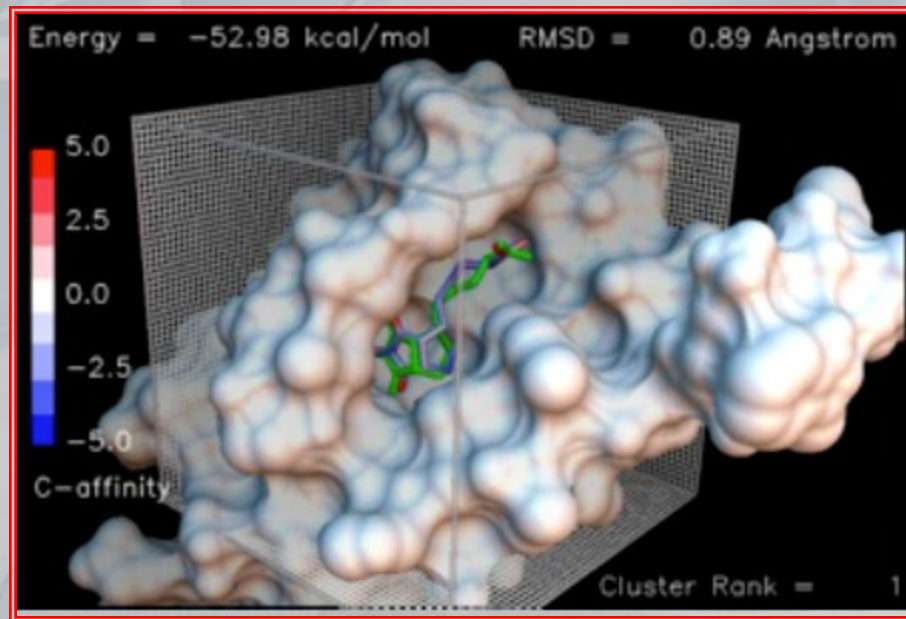
Benzamidine binding to β -Trypsin



The blue color of the phenyl ring indicates that it has a favorable interaction energy with the macromolecule; note that NH_2 groups are colored red

Ligand docking – 8

Biotin binding to Streptavidin



The ligand biotin and its final docked conformations, after 100 AutoDock dockings, to streptavidin; the crystal structure conformation of biotin is shown in green; to give an idea of the scale, the sides of the grid shown in the background are 22.875 Å long

Database screening – 1

- ✦ The main compromise in designing docking algorithms is the need of balancing between a complete and accurate search of all possible binding conformations, while implementing an algorithm with a “reasonable” computational complexity
 - For screening databases of possible drugs, searching algorithms must in fact perform the docking of thousands of ligands to the active site of a protein and, therefore, they need a high efficiency

Database screening – 2

- ✦ Methods specifically designed for database screening, such as the **SLIDE** (Screening for Ligands by Induced-fit Docking, Efficiently) algorithm, often reduces the number of considered compounds, using techniques of database indexing, to a priori discard those lead compounds which are unlikely to bind the active site of the target
- ✦ **SLIDE** characterizes the active site of the target in accordance with the position of potential donor and acceptor of hydrogen bonds, and considering hydrophobic points of interaction with the ligand, thus forming a model
- ✦ Any potential ligand in the database is characterized in the same way, in order to construct an index for the database

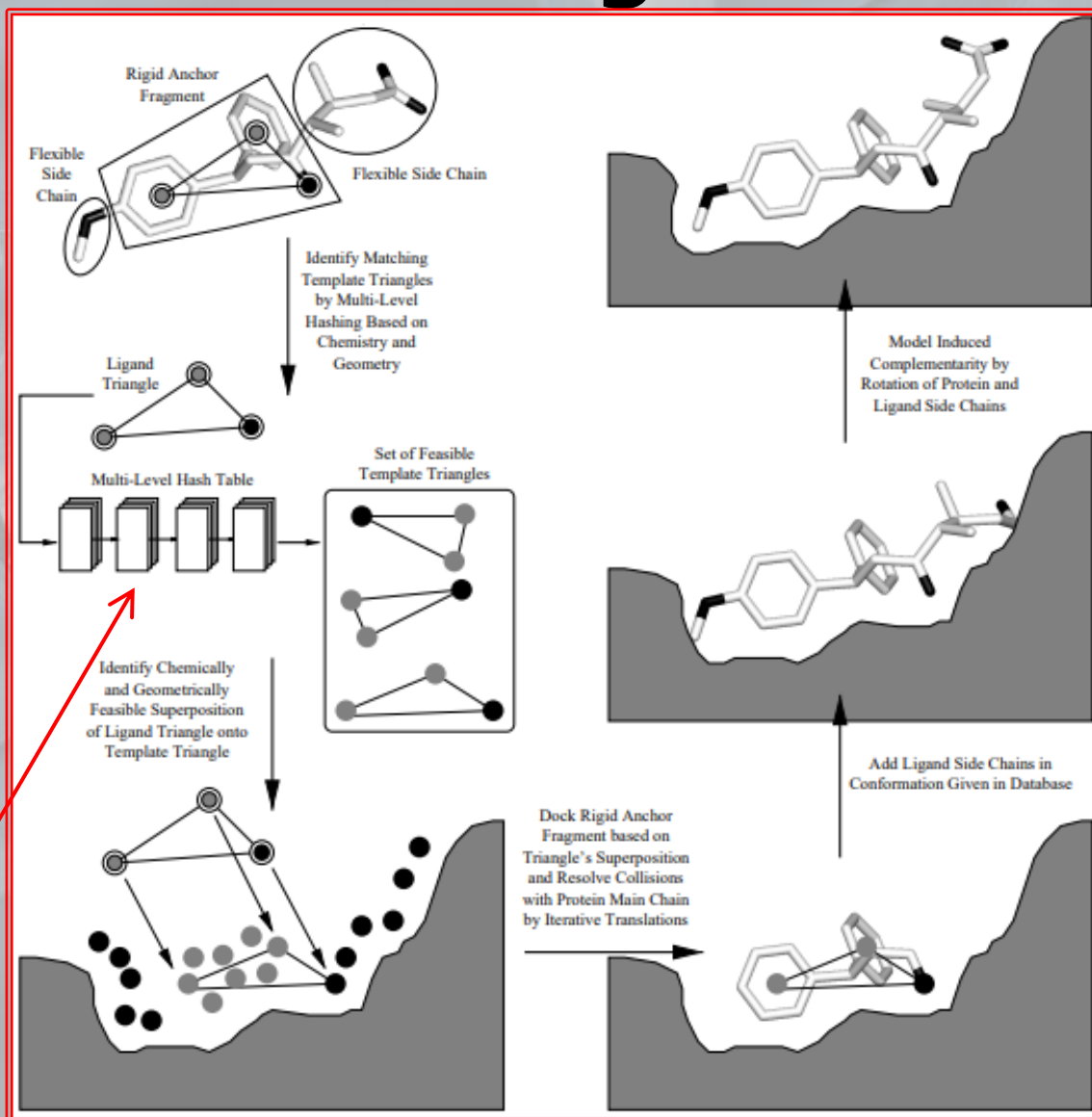
Database screening – 3

- ✦ The indexing operation allows SLIDE to rapidly eliminate those ligands that are, for example, too large or too small to fit the model
- ✦ By the reduction of the number of ligands that are subjected to the computationally expensive docking procedure, SLIDE (together with similar algorithms) can probe large database of potential ligands in days, or hours, compared to months
- ✦ Actually, SLIDE can screen 100,000 compounds within a few days and returns a ranked list of sterically feasible ligand candidates, ranked by complementarity to the protein binding site

Database screening – 4

SLIDE docking of potential ligands into the binding site is based on mapping triplets of ligand interaction centers (H-bond donors, acceptors, or hydrophobic ring centers) onto triangles of template points located above the protein surface; feasible template triangles for each possible triplet in a screened molecule are directly accessed via a multi-level hash table, and the corresponding mapping is used to dock the rigid anchor fragment of the ligand; single bonds in the flexible parts of both molecules are rotated to generate a shape-complementary interface, before the complex is scored by the number of intermolecular hydrogen bonds and hydrophobic complementarity of the contact surfaces

The geometric features considered are the length of the longest side of the triangle, the perimeter, etc.



An overview of protein–ligand interaction and affinity databases

- ✦ Binding DB: contains 2.9M data, for 1.3M compounds and 9.4K targets

www.bindingdb.org

- ✦ <https://www.youtube.com/watch?v=u49k72rUdyc&t=6s>

Crystal structures solved by X-ray – 1

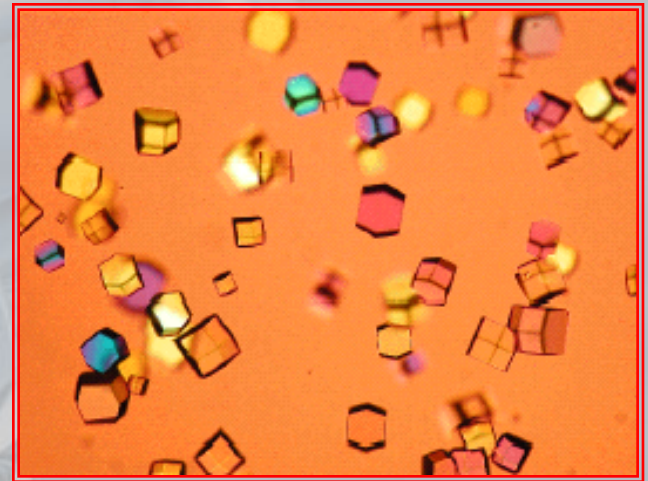
- ✦ Even the most powerful microscopic technique is insufficient to determine the molecular coordinates of each atom of a protein
- ✦ Instead, the discovery of X-rays by W. C. Roentgen (1895) has allowed the development of a powerful tool for the protein structure analysis: [the X-ray crystallography](#)
- ✦ In 1912, M. von Laue discovered that **crystals**, solid structures composed by an ordered lattice of atoms or molecules, diffract X-rays forming regular and predictable patterns
- ✦ In the early '50s, pioneering scientists such as D. Hodgkin were able to crystallize some complex organic molecules and to determine their structure by observing how they diffracted an X-ray beam
- ✦ Today, the X-ray crystallography was used to determine the structure of more than 189000 proteins with high resolution

Crystal structures solved by X-ray – 2

- ✦ The first step in the crystallographic determination of a protein structure is the growth of its crystal
- ✦ Crystallization is a very delicate and challenging process, but the basic idea is simple
 - Just as the sugar crystals can be produced through the slow evaporation of a solution of sugar and water, the protein crystals are grown by the evaporation of a solution of pure protein
 - Protein crystals, however, are generally very small (from about 0.3mm to 1.5mm in each dimension) and are composed by about 70% water, with a consistency more similar to gel than to the sugar crystals

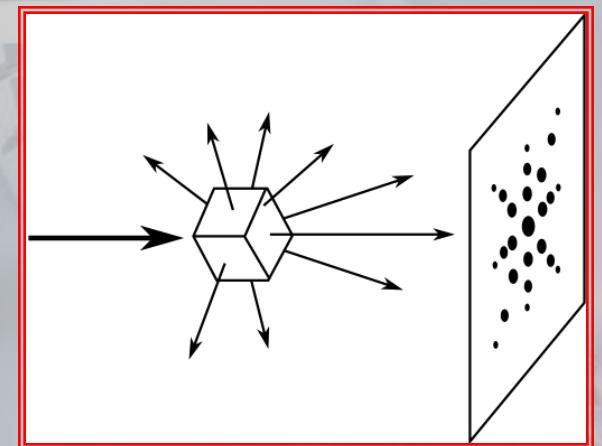
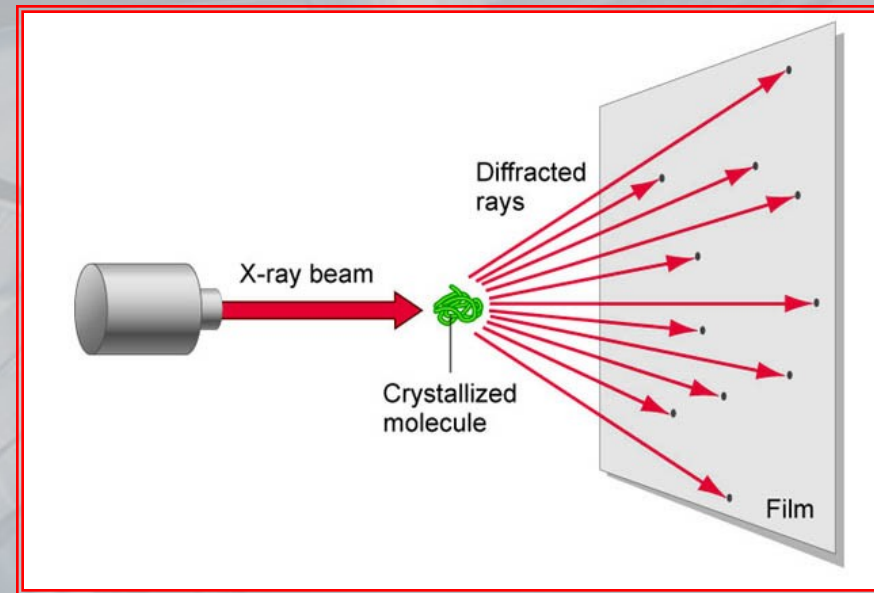
Crystal structures solved by X-ray – 3

- ✦ The growth of protein crystals requires carefully controlled conditions and a large amount of time: reaching an appropriate crystallization state for a single protein can take months of experiments
- ✦ Once obtained, protein crystals are loaded inside a capillary tube and exposed to an X-ray beam, which is diffracted by the crystals



Crystal structures solved by X-ray – 4

- ✦ Originally, the diffraction pattern was captured on a radiographic film
- ✦ Modern tools for X-ray crystallography are based on detectors that transfer the diffraction patterns directly on computers, for their successive analysis
- ✦ Given the gathered diffraction data, numerical methods, *ad hoc* software and known protein models can be used to determine the 3D structure of the protein



Crystal structures solved by X-ray – 5

✦ In detail...

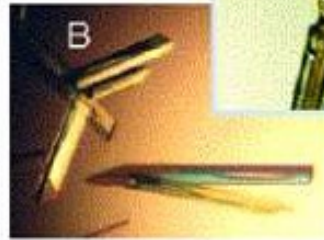
- From the X-ray diffraction spectrum of the crystal, crystallographers are able to calculate the electron density map, which, in practice, is an image of the molecules that form the crystal, magnified about a hundred million times
- The electron density map can then be examined using computer graphics techniques to verify its agreement (fitting) with a molecular model
- In this way, and eventually after some adjustments, a molecular model can be obtained for the protein, with an average error on the coordinates of 0.3–0.5 Å, which allows a very detailed examination of its 3D structure

Crystal structures solved by X-ray – 6

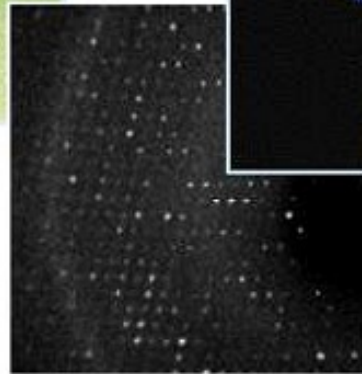
- ✦ Finally, note that the obtained crystallographic structure is essentially averaged over multiple copies of a single protein crystal and with respect to the time during which the crystal is exposed to X-rays
- ✦ The crystallized proteins are not completely rigid and the mobility of a particular atom within a protein can “confuse” the crystallographic signal
- ✦ Moreover, the problem of the presence of water molecules within the crystal (which are often included in the entry of the protein databases) is difficult to solve and causes “noise”
- ✦ However, crystallography is currently the main method for visualizing 3D protein structures at an atomic resolution

Crystal structures solved by X-ray – 7

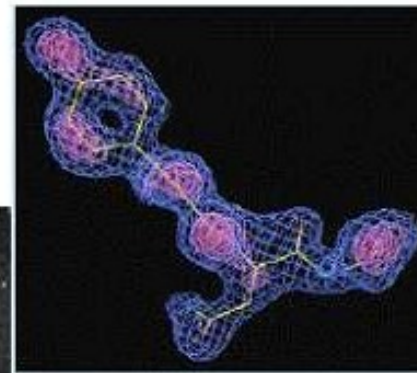
- ✦ The **Protein Data Bank (PDB)**, (<http://www.pdb.org>) is the leading database that collects protein structures derived from X-ray crystallography



Crystallization and crystal characterization

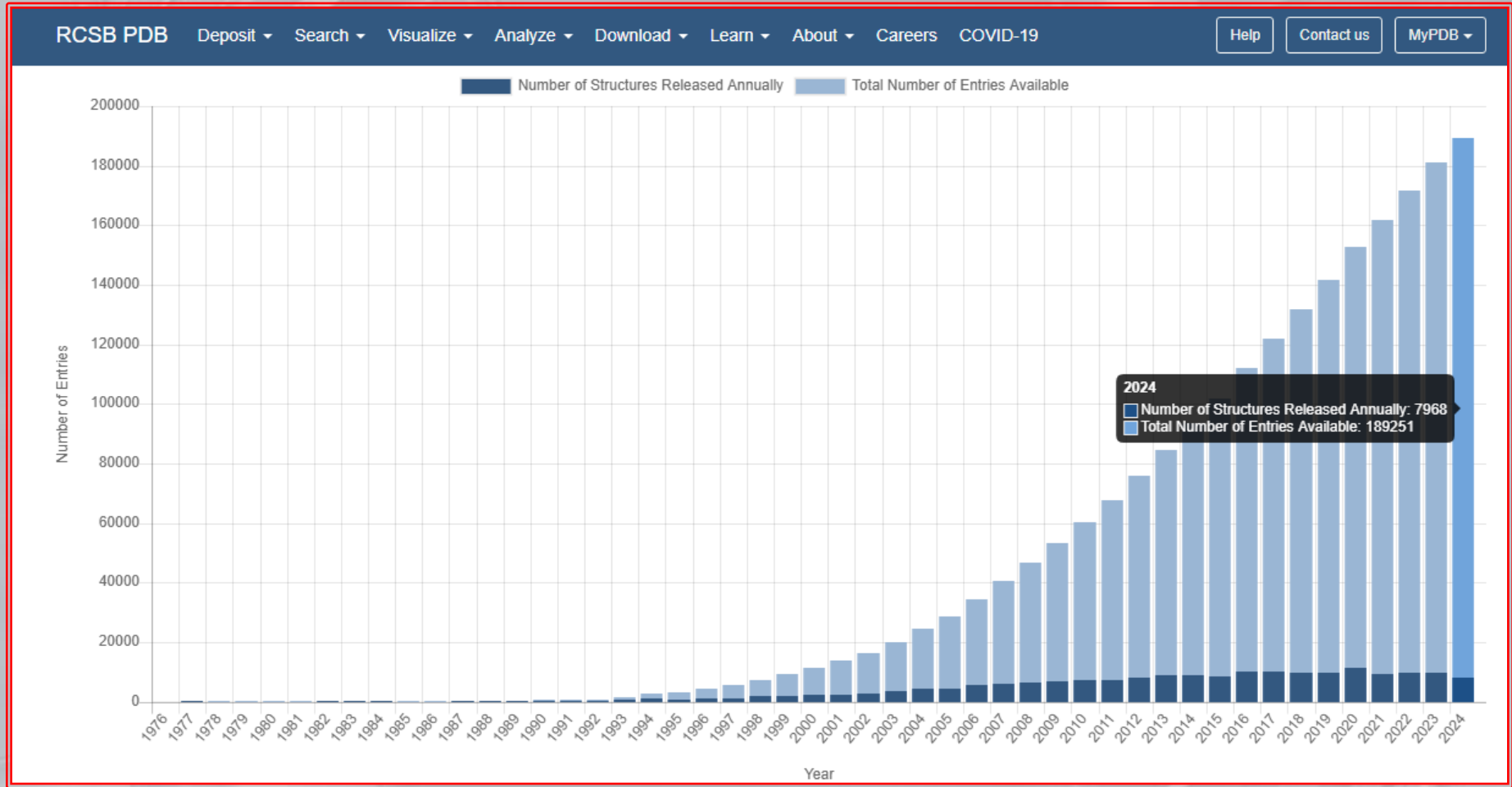


Data collection: diffraction spectrum of the crystal



Evaluation and interpretation of electron density maps

Yearly growth of structures solved by X-ray



Source: PDB statistics

NMR structures – 1

- ✦ The spectroscopic technique called **Nuclear Magnetic Resonance (NMR)** provides an alternative method to determine macromolecule structures
- ✦ At the basis of NMR, there is the observation that the atoms of some elements – such as hydrogen and radioactive isotopes of carbon and nitrogen – vibrate or resonate, when the molecules to which they belong are immersed in a static magnetic field and exposed to a second oscillating magnetic field
- ✦ Atomic nuclei try to align themselves with the static magnetic field, in a parallel or antiparallel configuration, and when the oscillating magnetic field provides them with an energy equal to the energy difference between the two states, the phenomenon of resonance occurs, which can be detected by external sensors, such as NMR spectrometers

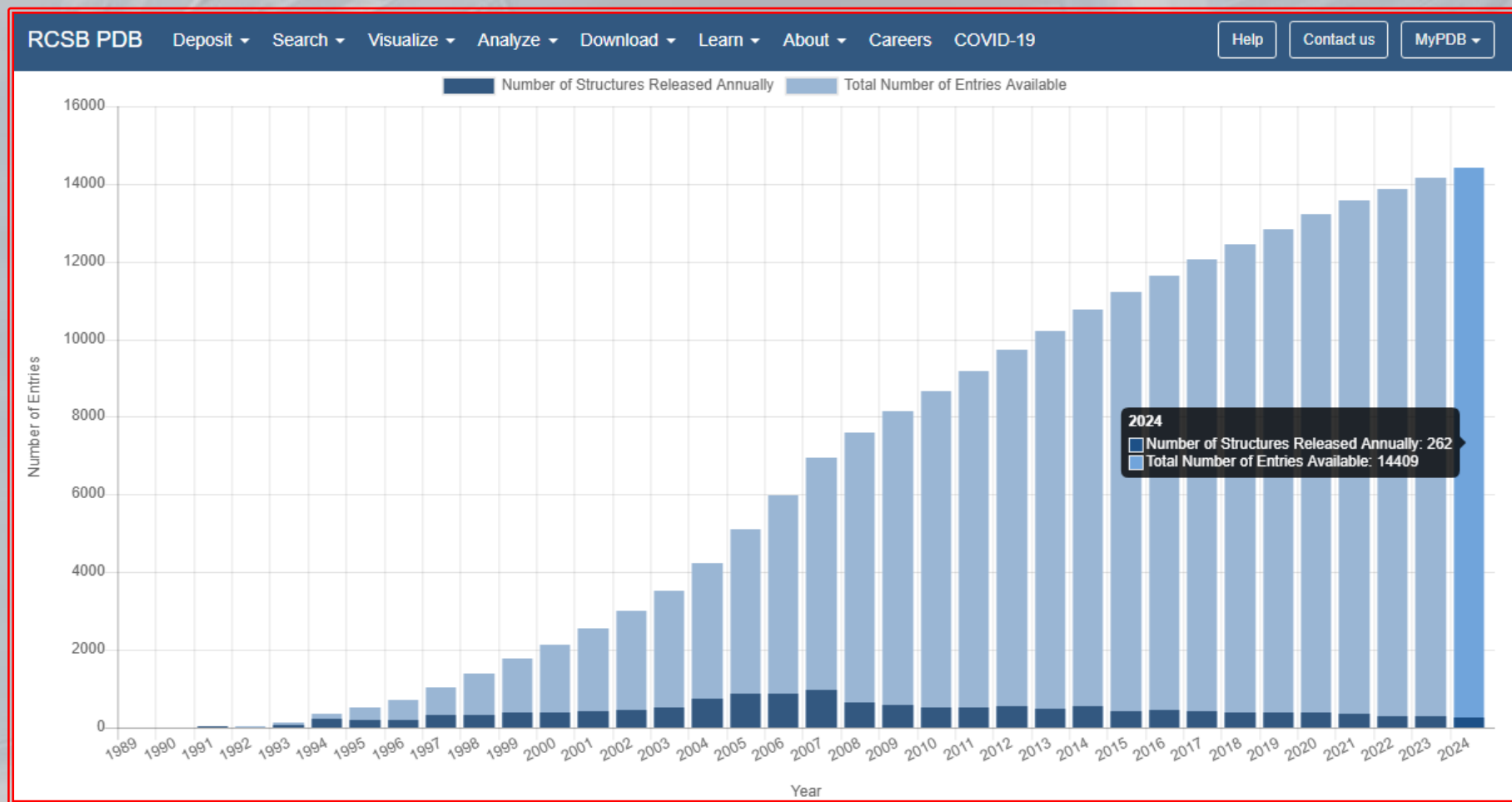
NMR structures – 2

- ✦ The behavior of each atom is mainly influenced by the neighboring atoms, i.e. those placed within adjacent residues
- ✦ Data analysis and interpretation requires complex numerical techniques and limits the utility of the approach for each protein or protein domain (a domain is composed of globular or fibrous polypeptide chains folded into compact regions, that represent “pieces” of the tertiary structure) no longer than few hundreds amino acids
- ✦ NMR methods do not use crystallization: they are very advantageous in the case of proteins that cannot be crystallized (especially integral membrane proteins)

NMR structures – 3

- ✦ The result of an NMR experiment is a set of constraints on the interatomic distances within a macromolecular structure
- ✦ These constraints can then be used, together with the protein sequence, to describe a model of the 3D protein structure
- ✦ However, in general, many protein models can actually satisfy constraints obtained from the NMR technique; therefore, NMR structures usually contain different models of a protein, that is, different sets of coordinates, while the crystallographic structures, normally, contain only one model
- ✦ PDB collects more than 14400 protein structures derived from NMR

Yearly growth of structures solved by NMR



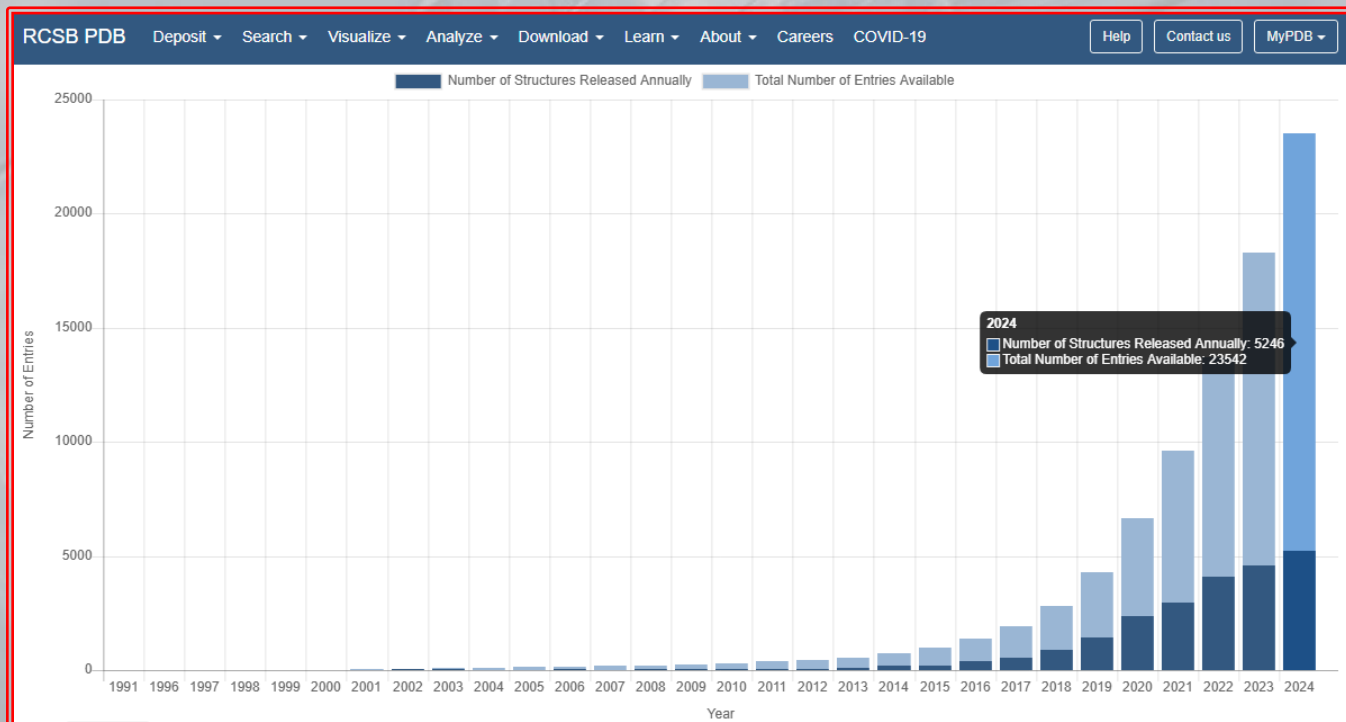
Source: PDB statistics

X-ray crystallography vs NMR

- ✦ Finally, we can notice that X-ray crystallography and NMR contain complementary information for the structural characterization of biological macromolecules (location of atoms and distances among them)
- ✦ X-ray diffraction is primarily sensitive to the overall shape of the molecule, whereas NMR is mostly sensitive to the local atomic arrangements
- ✦ Their combination can therefore provide a stronger justification for the resulting structure

Electron Microscopy structures

- ✦ Electron Microscopy (EM) is a versatile tool in the structural analysis of proteins and biological macromolecular assemblies
- ✦ It provides opportunities for the 3D structural determination of macromolecular complexes that are either too large or too heterogeneous to be investigated by conventional X-ray crystallography or nuclear magnetic resonance



PDB again – 1

- ✦ We have seen some methods to obtain macromolecular structures, but where can we find them stored?
- ✦ The **Protein Data Bank (PDB)** format provides a standard representation for macromolecular structure data derived from X-ray diffraction and NMR studies
- ✦ This representation was created in the 1970's and a large amount of software using it has been written
- ✦ Some documentation describing the PDB file format is available from wwPDB at

<http://www.wwPDB.org/documentation/file-format.php>

PDB again – 2

- ✦ Protein structures contained in **PDB** are stored in text format (or in **mmCIF**, for macromolecular Crystallographic Information File)
 - Each line of a PDB file contains the coordinates (x,y,z) , in angstroms (10^{-10} m) of each atom of a protein, with other useful information
- ✦ Also, a navigable 3D representation of a protein can be obtained
- ✦ For each structure in the PDB database, a four character code is assigned
 - **Example:** 2APR identifies the data file of rizopuspepsine, which is an aspartic protease
 - Files in PDB format are generally called XXXX.pdb or pdbXXXX.ent, where XXXX is the four-digit code related to the particular structure

PDB again – 3

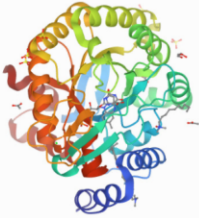
<https://www.rcsb.org/>

RCSB PDB - 6GK0: HUMAN DIHYDR...

← → ↺ rcsb.org/structure/6GK0

RCSB PDB Deposit Search Visualize Analyze Download Learn More MyPDB

Biological Assembly 1



6GK0
HUMAN DIHYDROOROTATE DEHYDROGENASE IN COMPLEX WITH CLASS III HISTONE DEACETYLASE INHIBITOR
DOI: [10.2210/pdb6GK0/pdb](https://doi.org/10.2210/pdb6GK0/pdb)
Classification: **ANTITUMOR PROTEIN**
Organism(s): *Homo sapiens*
Expression System: *Escherichia coli*
Mutation(s): No

Deposited: 2018-05-17 Release Date: 2018-06-13
Deposition Author(s): [Hakansson et al.](#)

Experimental Data Snapshot
Method: X-RAY DIFFRACTION
Resolution: 1.85 Å
R-Value Free: 0.193
R-Value Work: 0.168

3D View: [Structure](#) | [Electron Density](#) | [Ligand Interaction](#)

Global Symmetry: Asymmetric - C1
Global Stoichiometry: Monomer - A1

Display Files Download Files

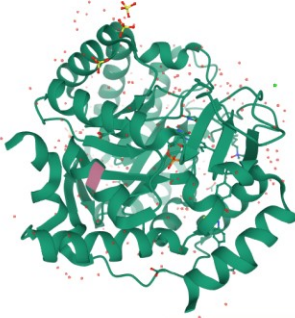
Contact Us

RCSB PDB - 6GK0: HUMAN DIHYDR...

← → ↺ rcsb.org/3d-view/6GK0/1

Sequence of 6GK0 | HUMA... 1: Dihydroorot... A

MATGDERFYAEHLMTLQGLLDPESAHRLAVRFTSLGLLPRARFQDSIMLEVRVLGHKFRNFVGIAAGFDKHGEAVDGLYKMGFGFVEIGSVTFKPKQEGNFRPRVFR
140 150 160 170 180 190 200 210 220 230 240
LPEDQAVINRYGFNSHGLSVVEHRLRARQQKQAKLTEDGLFLGNLGNKNTSDVDAEDYAEGRVVLGLADYLVVNVSSPNTAGLSLQGGKAELELLTKVLQERDG
250 260 270 280 290 300 310 320 330 340 350
LRRVHRPAVLVKIAPDLTSQCKEDIAVVKELIGDLIVTNTTVSRPAGLQGLRSETGGLSGKPLRDLSTQTIREMYALTQGRVPIIGVGVSQGDALRIRAGA



Dihydroorotate dehydrogenase (quinone), mitochondrial
6GK0 | Model 1 | Instance ASM_1 | A | GLU 315 [auth 344]

Structure

6GK0 | HUMAN DIHYDROOROTATE...

Type Assembly

Asm Id 1: Author And Softwar...

Nothing Focused

Measurements

Components 6GK0

Preset + Add

Polymer Cartoon

Ligand Ball & Stick

Water Ball & Stick

Ion Ball & Stick

Unit Cell P 32 2 1

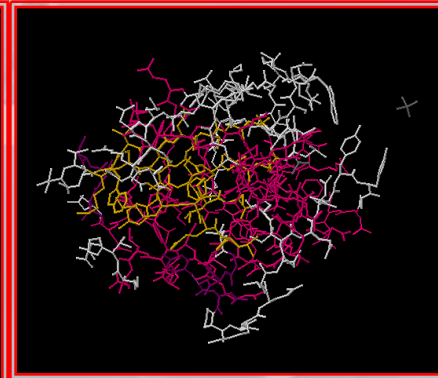
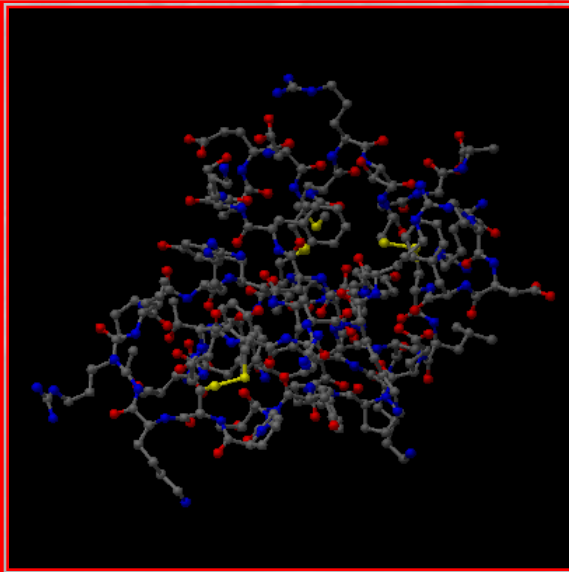
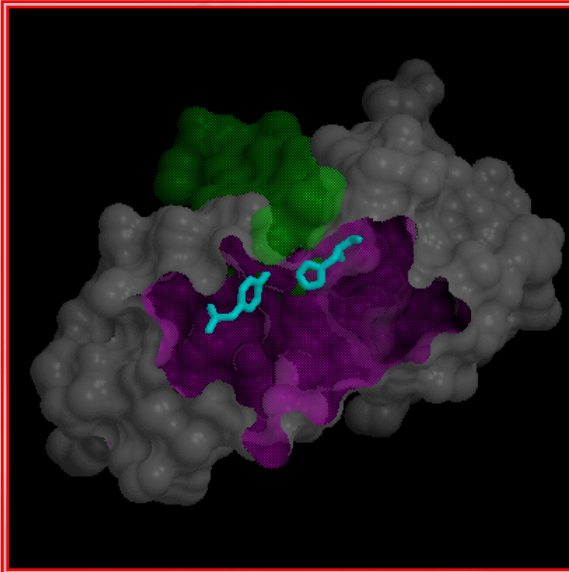
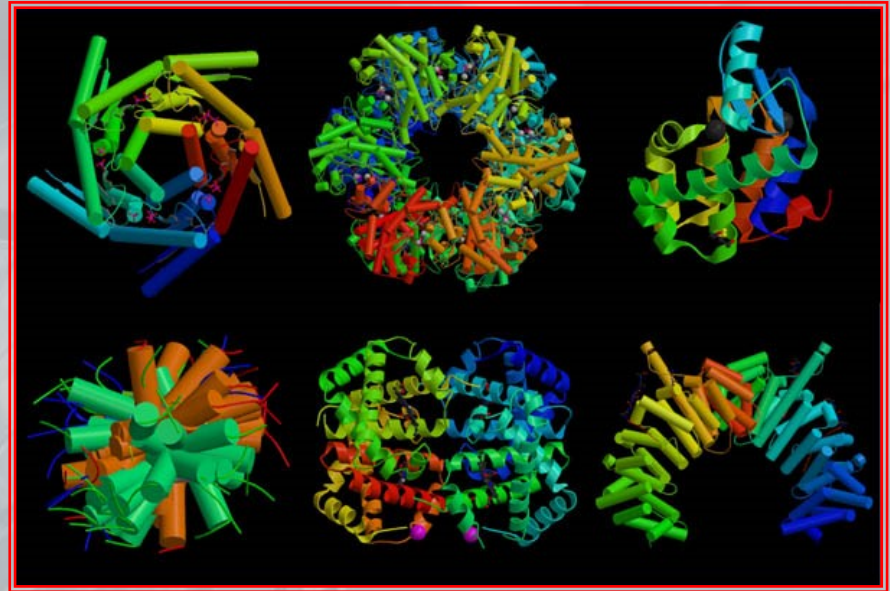
Density

Assembly Symmetry

Tertiary structure representations

The **cartoon** method evidences regions of secondary structure

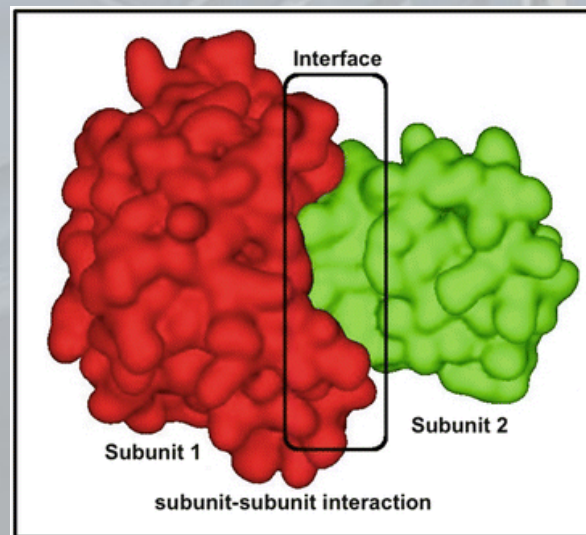
The representation of the **molecular surface** reveals the overall shape of the protein



The 3D **thread-like representation** (with balls & sticks or wireframe) illustrates molecular interactions

What are protein tertiary structures useful for?

- ✦ **Example problem:** Definition of an algorithm which, given the 3D structure of a protein, is able to predict which residues are most likely involved in protein–protein interactions
 - A very important question since many proteins are active only when they are associated with other proteins in a multienzymatic complex



Protein–protein interactions

- ✦ Protein–protein interactions are characterized as stable or transient, and they could be either weak or strong
- ✦ Hemoglobin and core RNA polymerase are examples of multi–subunit interactions that form stable complexes
- ✦ Transient interactions are temporary and typically require a set of conditions that promote the interaction, such as phosphorylation
- ✦ Proteins bind to each other through a combination of hydrophobic contacts, van der Waals forces, and salt bridges in specific binding domains on each protein
- ✦ These domains can be small binding clefts or large surfaces, and can be just a few peptides long or span hundreds of amino acids; the strength of the binding is influenced by the size of the binding domain

Empirical methods and predictive techniques – 1

- **Solution:** from the PDB database, select a set of sample structures, which are constituted by two or more proteins that form a complex
 - There will be interfacial residues, involved in the contact surface, and non interfacial residues
 - **Input:** For each residue, a set of features, to be measured and used to solve the prediction problem, must be selected
 - **Output:** Prediction of residues belonging to protein–protein interfaces (for unknown proteins)

Empirical methods and predictive techniques – 2

✦ Possible features:

- Number of residues within a given radius with respect to the test residue
- Net charge of the residue and of the neighboring residues
- Hydrophobicity level
- Potential of the hydrogen bonds
- ➡ Construction of a feature vector describing the given residue

✦ In conjunction with the feature vector, a target is given, attesting the membership (or not) of the particular residue to the protein–protein interface

➡ **Application of machine learning methods**

Protein–protein interactions and drug design

- ✦ The importance of unveiling the human protein interaction network is undeniable, particularly in the biological, biomedical and pharmacological research areas
- ✦ Even though protein interaction networks evolve over time and can suffer spontaneous alterations, occasional shifts are often associated with disease conditions
- ✦ These disorders may be caused by external pathogens, such as bacteria and viruses, or by intrinsic factors, such as auto-immune disorders and neurological impairment
- ✦ Therefore, having the knowledge of how proteins interact with each other will provide a great opportunity to understand pathogenic mechanisms, and subsequently to support the development of drugs focused on very specific disease pathways, and for re-targeting already commercialized drugs to new tasks

Post-translation modification prediction

- ✦ The wide variety of protein structures and functions is partially due to the fact that proteins are subjected to many modifications also after being translated
 - Removal of protein segments
 - Formation of covalent bonds between residues and sugars, or phosphate and sulphate groups
 - Formation of cross-links involving (possibly far) residues within a protein (disulfide bonds)
- ✦ Many of these modifications are carried out by other proteins, which must recognize specific surface residues, appropriate to trigger such reactions
- ➡ Neural network based prediction techniques

Protein sorting – 1

- ✦ Some processes take place during and immediately after the translation process
- ✦ **Protein sorting** is the biological mechanism by which proteins are transported to their appropriate destinations in the cell or outside it
- ✦ Eukaryotic cells are different from the prokaryotic ones, given the presence of membrane-enclosed organelles within their cytoplasm
- ✦ The organelles provide separated compartments in which specific cellular activities take place and the resulting subdivision of the cytoplasm allows eukaryotic cells to function efficiently in spite of their large size (which are about a thousand times the volume of a bacterium)

Protein sorting – 2

- ✦ Therefore...
 - The presence of internal cellular compartments surrounded by membranes is a eukaryotic peculiarity
 - Both the chemical environment and the protein population may differ greatly in different compartments
- ✦ It is imperative, for energetic and functional reasons, that eukaryotic organisms provide to transport all their proteins into their appropriate compartments
- ✦ For example, histones – proteins that bind to DNA and are associated with the chromatin – are functionally useful only inside the nucleus, where chromosomes reside
- ✦ Other proteins – such as proteases, which are located inside peroxisomes (specialized metabolic compartments) – would even be dangerous for the cell if they were found in any other place inside the cell

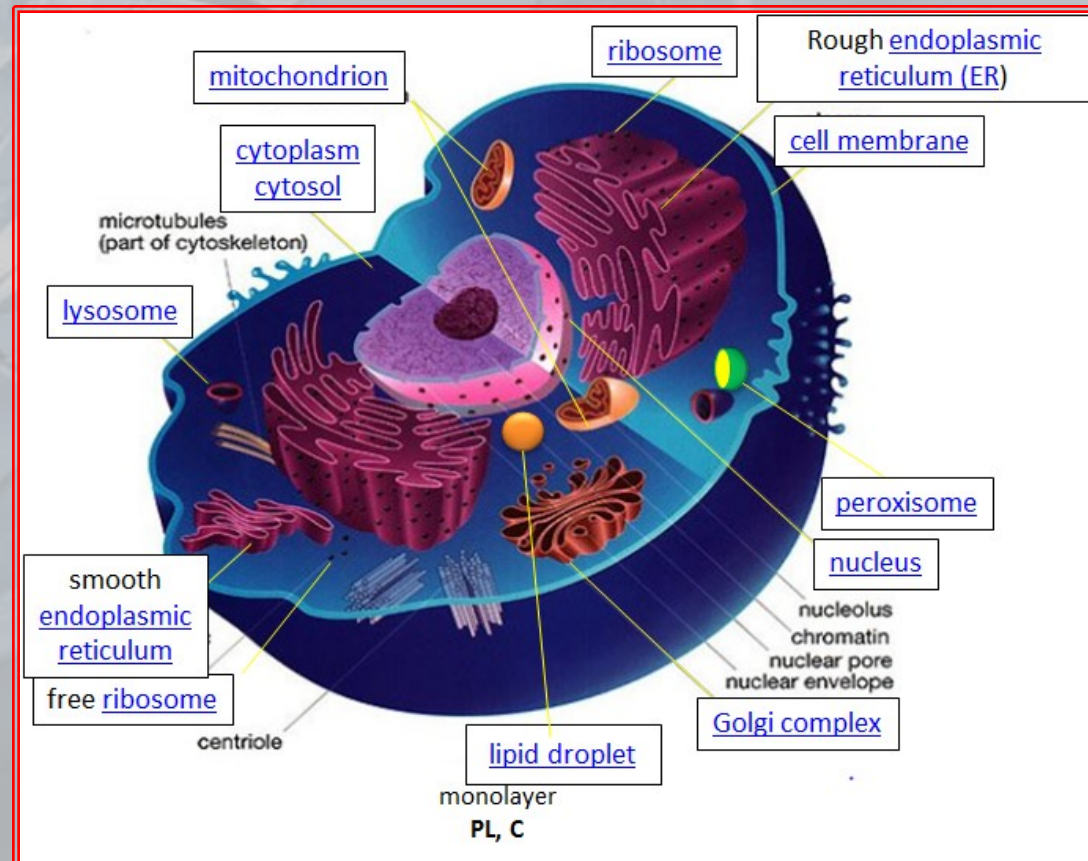
Protein sorting – 3

- ✦ Because of the complex internal organization of eukaryotic cells, sorting and targeting proteins to their appropriate destinations are considerable tasks
- ✦ It seems that eukaryotic cells regard proteins as belonging to two distinct classes, according to their location: **proteins not associated with or attached to the membranes**
- ✦ The first set of proteins is exclusively translated by ribosomes which “float” inside the cytoplasm

Protein sorting – 4

✦ Subsequently, proteins, translated by the floating ribosomes, may remain in the cytoplasm or may be transported...

- within the nucleus
- in the mitochondria
- in the chloroplasts
- in the peroxisomes



Protein sorting – 5

- ✦ Cytoplasm appears to be the default environment for proteins; on the contrary, the protein transport in different compartments, separated by membranes, requires the presence and the recognition of specific localization signals

Organelles	Signal localization	Type	Signal length
Mitochondria	N-terminal	Amphipathic helix	12–30
Chloroplasts	N-terminal	Charged	~25
Nucleus	Internal	Basicity	7–41
Peroxisomes	C-terminal	SKL serine-lysine-leucine	3

Protein sorting – 6

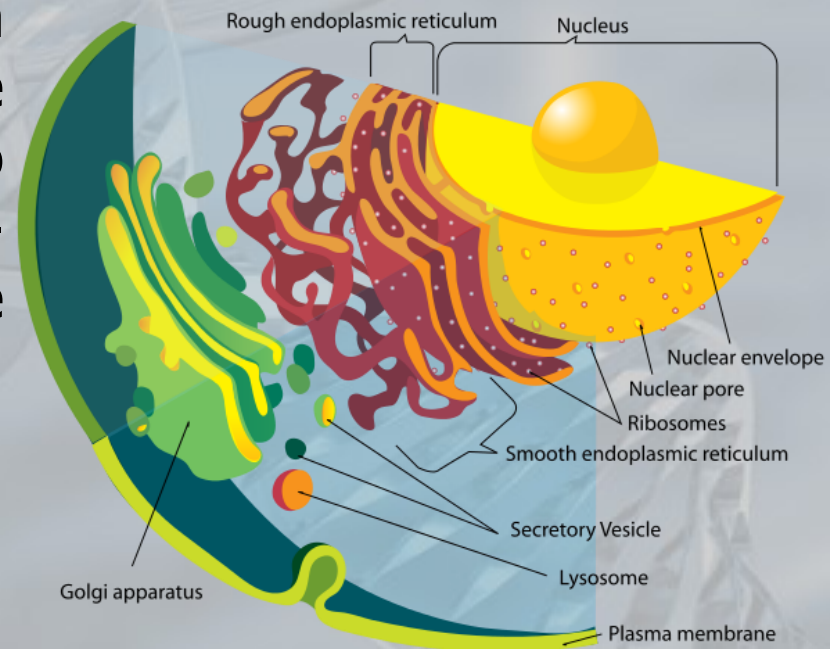
- ✦ Nuclear proteins possess a **nuclear localization sequence**: an internal region, composed by 7 to 41 amino acids, rich in lysines and/or arginines (positive charged)
- ✦ Mitochondrial proteins possess an amphipathic helix (with both hydrophobic and hydrophilic amino acid residues arranged in such a way as to create two faces on opposite sides of the helix), composed by 12 to 30 amino acids, located at their N-terminal
 - This **mitochondrial signal sequence** is recognized by a receptor on the mitochondrion surface, and it is often removed to activate the protein as soon as it is transported inside the organelle

Protein sorting – 7

- ✦ The chloroplast proteins have a **chloroplast transit sequence** (about 25 charged amino acids located at the N-terminal), which is recognized by the protein receptors on the chloroplast surface
- ✦ Finally, proteins destined to peroxisomes possess the **peroxisomal target signal** (serine–lysine–leucine) which is recognized by the *ad hoc* receptors, that ensure their transport to the correct destination

Protein sorting – 8

- ✦ The second set of proteins is translated by ribosomes, bound to the membrane, which are associated with the **endoplasmic reticulum** (ER)
- ✦ The endoplasmic reticulum is a network of membranes intimately associated with the Golgi apparatus, where additional modifications to proteins (such as glycosylation and acetylation) take place



Protein sorting – 9

- ✦ All the proteins translated by the ER ribosomes, in fact, begin to be translated by floating ribosomes within the cytoplasm
 - When the first 15–30 amino acids to be translated correspond to a particular **signal sequence**, a molecule, which recognizes the protein, binds to it and stops its translation, until the ribosomes and its mRNA are transported into the ER

Protein sorting – 10

- ✦ Although no particular consensus sequence exists for the signal, almost always there is a hydrophobic sequence, 10–15 residues long, that ends with one or more positive charged amino acids at the N-terminal
- ✦ When the translation resumes, the new polypeptide is extruded through a pore in the membrane of the ER, inside the lumen (the interior space) of the same ER
 - A **peptidase signal** protein cuts the target N-terminal sequence from the protein (unless the protein should be retained permanently as a membrane-bound protein)
- ✦ Neural nets: [SignalP - 5.0 - Services - DTU Health Tech](#)

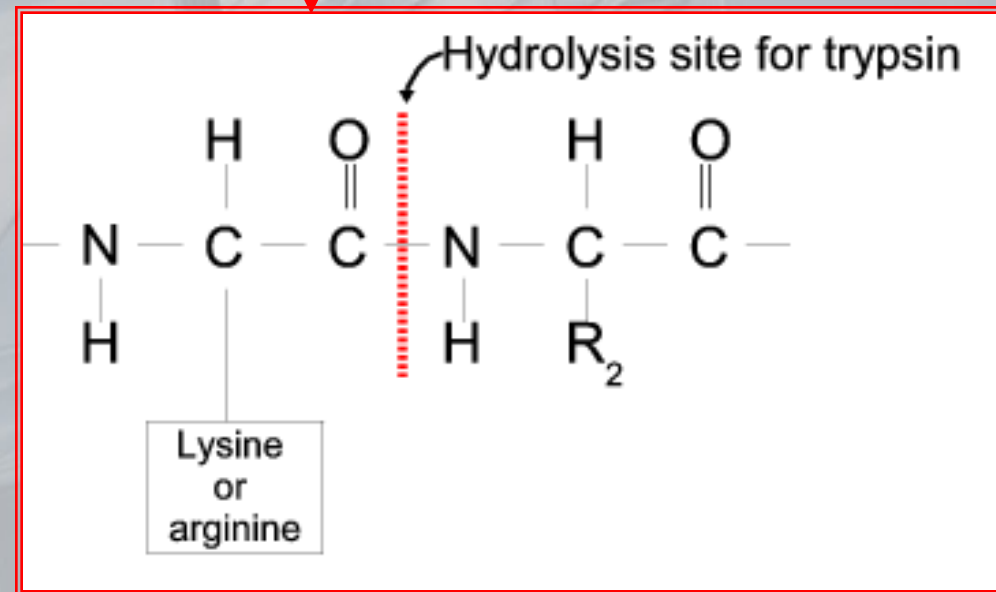
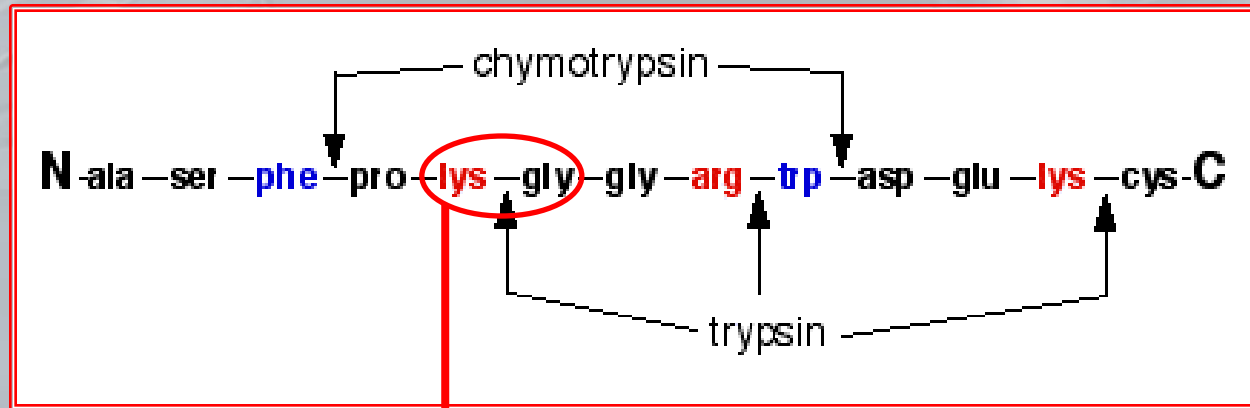
Proteolytic cleavage – 1

- ✦ Both prokaryotes and eukaryotes possess several enzymes responsible for the cutting and the degradation of proteins and peptides
- ✦ There are different types of proteolytic cleavage:
 - Removal of the methionine residue present at the beginning of each polypeptide (since the start codon also codes for methionine)
 - Removal of the signal peptides
 - ...

Proteolytic cleavage – 2

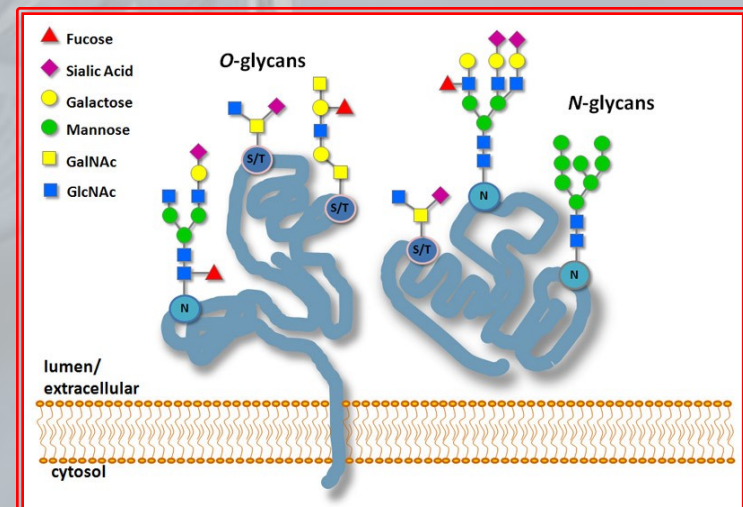
- ✦ Sometimes, the cleavage signal is constituted by a single residue
- ✦ However, in many cases, the sequence motif is longer and ambiguous
 - Chymotrypsin cuts polypeptides at the C-terminal of bulky aromatic residues (containing a ring), as the phenylalanine
 - Trypsin cuts the peptide bond on the carboxyl side of lysine and arginine residues
 - Elastase cuts the peptide bond on the C-terminal of small residues, such as glycine and alanine
- ✦ Neural networks: prediction accuracy > 98% (see <http://www.paproc.de>, a prediction algorithm for proteasomal cleavages)

Proteolytic cleavage – 3



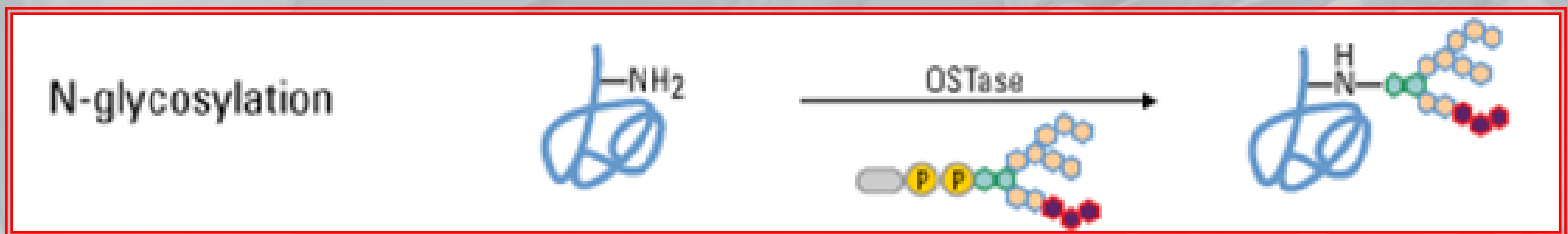
Glycosylation – 1

- ✦ **Glycosylation** is the process that permanently binds an oligosaccharide (a short chain of sugars) to the side chain of a residue on the protein surface
- ✦ The presence of glycosylated residues can have a significant effect on protein folding, location, biological activity and interaction with other proteins
- ✦ In eukaryotes:
 - N-glycosylation
 - O-glycosylation



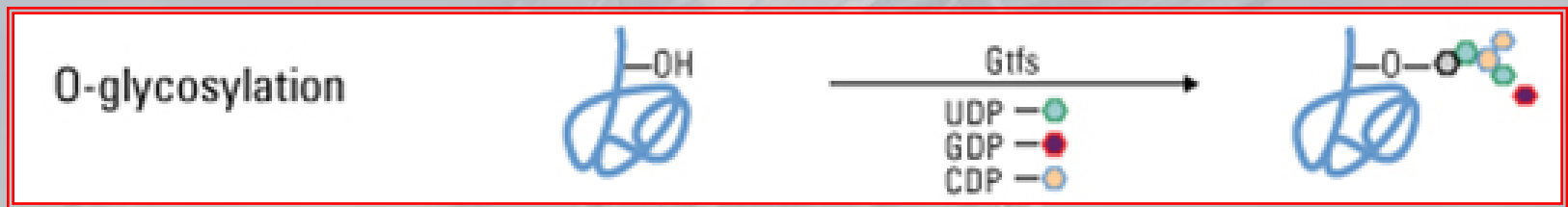
Glycosylation – 2

- ✦ The N-glycosylation is the addition of an oligosaccharide to an **asparagine** residue during protein translation
 - The main signal which indicates that an asparagine residue (Asn) has to be glycosylated is the local amino acid sequence Asn-X-Ser or Asn-X-Thr, where X corresponds to any residue except proline
 - However, this sequence alone is not sufficient to determine glycosylation (as we can observe in Nature)



Glycosylation – 3

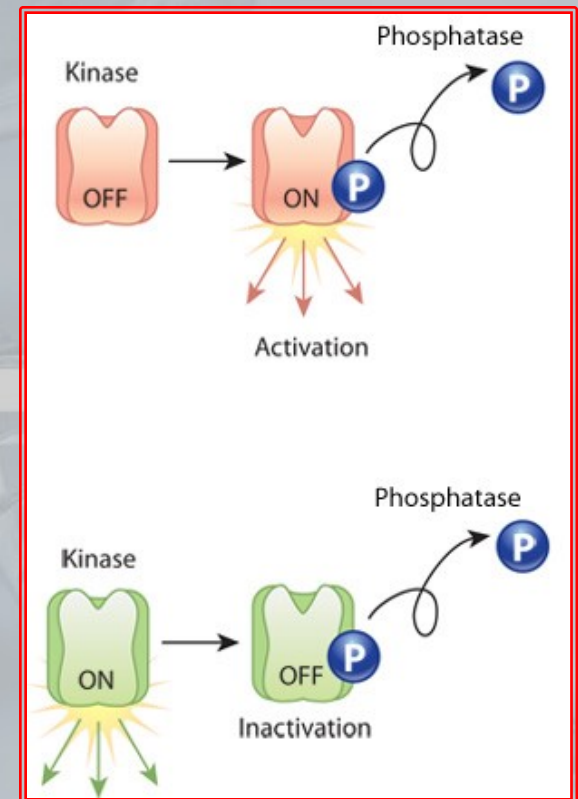
- ✦ The O-glycosylation is a post-translational process in which the N-acetyl-glucosaminyl transferase binds an oligosaccharide to an oxygen atom of a **serine** or a **threonine** residue



- ✦ Unlike the N-glycosylation, known sequence motifs that mark a site for O-glycosylation do not exist, except for the presence of proline and valine residues near the Ser or Thr which must be glycosylated
- ✦ Neural Networks: accuracy of 75% for N-glycosylation and higher than 85% for O-glycosylation

Phosphorylation – 1

- ✦ **Phosphorylation** (binding of a phosphate group) of surface residues is probably the most common post-translational modification in animal proteins
 - **Kinases**, which are responsible for phosphorylation, are also involved in a wide variety of regulatory pathways and signal transmissions
 - Since phosphorylation frequently serves as a signal for the enzyme activation, it is often a temporary condition
 - **Phosphatases** are the enzymes responsible for removing phosphate groups from phosphorylated residues



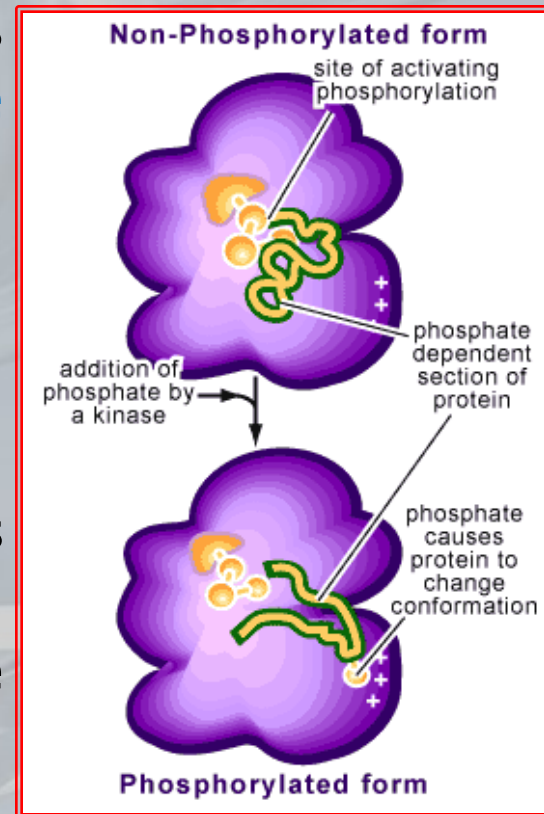
Phosphorylation – 2

- ✦ Since phosphorylation of key residues of **tyrosine**, **serine** and **threonine** serves as a regulatory mechanism in a wide variety of molecular processes, the diverse kinases involved in each process must show a high specificity in the recognition of particular enzymes

- No single consensus sequence identifies a residue as a phosphorylation target
- Some different not easily identifiable patterns exist

- ✦ Neural Networks: accuracy ~70%

([NetPhos 3.1 - DTU Health Tech - Bioinformatic Services](#))



Concluding... – 1

- ✦ While genomics is rapidly becoming a highly developed research area, proteomic techniques are only beginning to identify proteins, encoded in the genome, together with their various interactions
- ✦ The proteome characterization promises to bridge the gap between our knowledge of the genome and morphological and physiological effects due to genetic information
- ✦ Various taxonomies have been developed to classify and organize proteins, according to their enzymatic function, and to sequence and 3D structure similarities

Concluding... – 2

- ✦ Equipped with databases of families, superfamilies and protein folds, together with advanced experimental techniques, such as 2D electrophoresis and mass spectrometry, analysts are able to separate, purify and identify the diverse proteins expressed by a cell at a given time
- ✦ Important proteomics applications are in drug design
 - Recent advances in protein structure understanding and in X-ray crystallography have allowed the development of some automated methods for screening and docking of ligands and significantly contribute to the process of drug discovery

Concluding... – 3

- ✦ Although the 3D structure of proteins is fundamental to understand their function and their interaction with other proteins, some useful information can also be obtained from the primary structure
- ✦ Actually, the location signal and the various post-translational modifications are described by sequence motifs well conserved within the primary structure of proteins
 - Post-translational modifications (also) account for the fact that the same gene may encode for many proteins

A great Ted Talk on Genomics

- <https://www.youtube.com/watch?v=s6rJLXq1Re0>

A great TED Talk about proteins

- Leading network expert — Albert Laszlo Barabasi — talks about how we can model proteins and understand disease occurrence using complex systems and network theory
- <https://www.youtube.com/watch?v=10oQMHadGos>