Sequence Alignment:

Homology can be defined as existence if a shared ancestry of 2 biological units. Two homologous comes from a common ancestor. A particular kind is paralogy which is homology derived from a gene duplication event. The two genes can derive in this way can accumulate mutations independently. When it happens the fitness of the organism does not change until at least a copy remain functional. Usually what happens in evolution is that one of the two copies undergoes genetic drift and become a pseudogene losing its ability to codify. Very rarely one of the copy is able to acquire a new function and codify a new protein with similar sequence and structure.

Ortology is homology which derive from gene separation. This kind of separation does not happen in the same organism through duplication for example but is related to the process of speciation:
If there is a population which share the same genome, when a part of a population separates from the whole group either ecologically or in space until those population cannot breed and have fertile offspring anymore, the two species accumulate mutation for one gene independently from each other and the two-protein codified will be different even conserving same function and structure.

We can understand if two proteins are homologous comparing their structures and sequences. If we superpose the two and they are similar off if they share a sequence similarity of 20 percent, it means that they are homologous. before 20 we are in the twilight zone Occam's razor.

To compare the protein sequences, we must align them. The amount of sequence information of today is much higher compared to the information of protein structures. Align two sequences means to superpose the letter in an optimal way without changing the order of the letter and introducing gaps in order to have a specific functional and evolutionary meaning in a sense that the model obtained for the alignment is more sensible than a random one.
LA CASA E` BELLA
LA CASSA E` BUIA example slide (include photo)
If we were to align those sentences, we would superimpose the identical letters trying to maximize the number of the identical superpose letters by introducing the gaps (indels) we can have some symbols '|' for equal letters '.' or ':' low and high similarity respectively and conservation of physic chemical property in the alignment. Align two biological sequence is absolutely different form align two strings. When we align two biological sequences is to try to understand if two proteins are homologous and once, we are sure of that our aim is to obtain functional structural and evolutionary information from the alignment. Optimally alignment or superposition is the one which maximize the chance to obtain an evolutionary model compared to a random one.

PAIRWISE ALIGNMENT

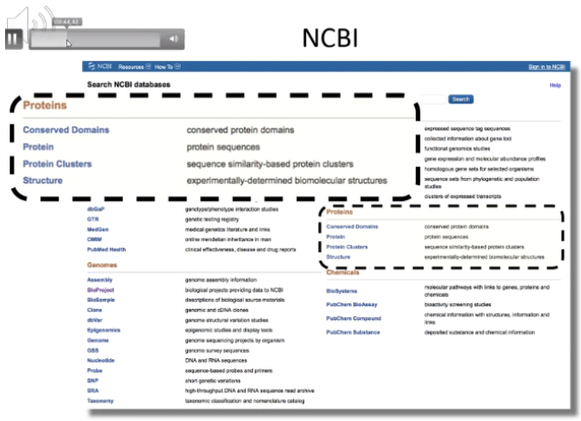
Is easy to align by hand two short sequences. Is harder for longer sequences as it happens for the protein sequence in human. In this case we should use an algorithm to perform it in an optimized way. The kind of algorithm is called dynamic programming. We cannot use a deterministic approach and explore every possible solution and combination for every sequence because the total number of possible pairwise alignment is the number and depend by the length of the sequence. If we have two sequences of length n and m the possible number of pairwise is: (m+n)!/m!*n!
The total number of possible pairwise alignment increases factorially with the lengths of the two sequences.

The recipe for the sequence alignment through dynamic programming:
Dynamic programming is a way to automatize the sequence alignment using an algorithm.
1) take two sequences at least
2) build a sequence matrix
That is, putting letters of the two sequences on the rows and columns of a matrix respectively.
3) choose a scoring matrix
That is a matrix of the scoring that we obtain when an amino acid od one sequence is aligned with the amino acid on the other sequence.
4) use an algorithm to identify the optimal path
Use one of the algorithms at disposal.
5) choose a scoring scheme for indels
Each algorithm can associate a certain penalty when we insert indels in our alignment.

$$(m+n)!/m!*n!$$

e.g., m=n=**10**,
2432902008176640000/3628800*3628800 = **184756**

DATABASES



We have special protein databases like NCBI databases for proteins where we can find sequences, domains, structures or clusters. Using keyword, we can search the protein that we are looking for.

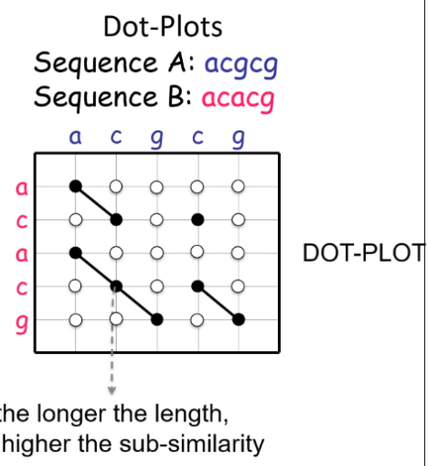
NCBIs databases might present redundant sequences for the same protein or genes because they have been sequenced for different labs from all over the word also using different technologies.

In order for NCBI to get rid of these NCBI utilize a subset of preferences sequences called refseq which are manually inserted. Each entry of each protein is represented once and only once.

UNIPROT

Is another important database which is really of pivotal important for bioinformaticians and is a free accessible database of protein sequences and information's and the entries in manually curated as in refseq and use a large amount of information about proteins.

SEQUENCE MATRIX



Dot-plot have a very strict connection with scoring matrix and sequence matrix. A dot-plot is a simplified version of the sequence matrix.

A dot-plot is a simple matrix and then we use a very simple scoring function that is a system to assign a score to the alignment. O in case of mismatching amino acid and 1 in case of Match, this is called a binary system scoring. Is called scoring matrix because using a scoring scheme we can build a matrix in which we have alle the amino acid and having the scoring for each match or mismatch with all the possible combinations.

In the case of a binary scoring matrix is too simple to assign 0 or 1 because in this way we are excluding the possibility of giving to each match a specific score based on the similarity of the two amino acid either physiological or evolutional.

What we would like to reach is a sort of scoring scheme in which we have positive or negative value corresponding to the similarity of residues

Ex: glutamate for aspartate positive score and aspartate for tryptophan would have a negative score.

Is not easy to quantify the score to give to each amino acid based on their properties because there are many factors to consider. Indeed, the scoring matrix are made by using a completely different method which is related to evolutionary relationship between the amino acids.

we assume that nature tends to conserve evolutionary the substitution that are favorable for the fitness of the organism, so we want to give to the evolutionary probability a >0 score and a value <0 to the random one.

The score is the quantity that we are going to insert in our scoring matrix and is called ODD SCORE.

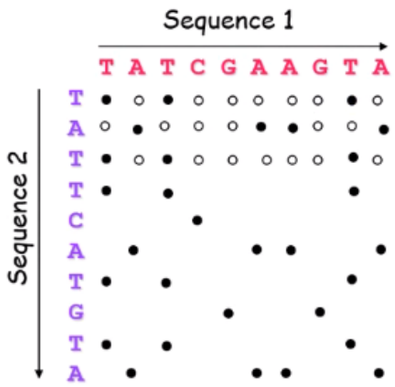
When the evolutionary prob. is stronger than the one by chance the score will be positive and vice versa.

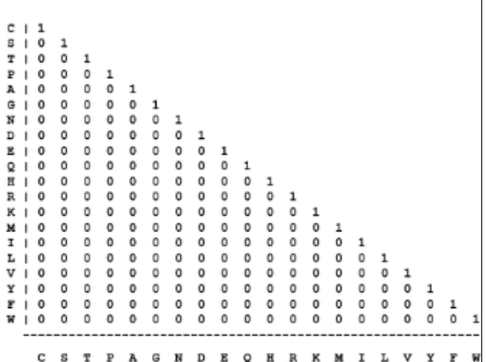
If we have a score of 0 it means that the numerator and denominator are the same.

To get these values we need to start with biological sequences and perform statistical analysis on them

F(x,y|E) x,y observed in nature/total substitution in nature

F(x,y|C) f(x)*f(y) because these probability are not related





● = 1

○ = 0

To derive F(x,y|E) we must observe the evolution of homologous sequences to access the number of substitutions but since we can't observe the evolution directly we must rely on already made models like PAM or BLOSUM.

The scoring matrix of PAM is made by having some score that represent the probability that a certain change is associated to a random mutation of evolutionary.
Notice that a change of leu for asp is associated to a negative score. We can derive a score also by considering the probability that some amino acid will be conserved during evolution.
Margaret Dayhoff 1978 developed Point Accepted mutation model (PAM)
M.Dayhoff is the first author of a protein database called atlas of protein sequences and structures.

She started with 71 homologous proteins with 85 percentage of sequence identity. The percentage of identity was so high because in this way we can assert that two protein are homologous for sure and easily aligned and second because it was easier to detect the single point mutation without the presence of hidden mutation which are:
If 2 sequences are evolutionarily closed in that case when there is a single substitution perhaps is the case of a single events while in more distant sequences perhaps between the two mutations, there are other mutation between the two mutations that we can notice, and those intermediates are hidden because they cannot be traced.
This led to the production of the first PAM matrix called PAMI which was aimed to describe the probability of just ONE amino acid replacement from one residue to another between two sequences without intermediate passages.
If we focus on three sequences A, B and C very similar but with some substitutions, using a phylogenetic algorithm called maximum parsimony we can build an evolutionary tree, choosing among the many possible trees, which minimizes the number of changes from one sequence to the other. That is why is called maximum parsimony.
We can use the evolutionary trees to derive the sequence of common ancestor and to access the type and manner of the substitution which were present during evolution present during the evolution of this family.
This simple point mutations detected during evolution are called PAM point accepted mutations and the scoring of PAM matrices is based on the row count of PAMS.
PAM I:1 pam PAM II:2 pams
Concept of PAM can be used as unit of evolution distance: if we have two sequences at an evolutionary distance of 1 pam, it is sufficient to have 1 mutation on 100 residues to reach the identity.
Pam 001 scoring matrix is a matrix that describes the probability that x mutates into y if the two sequences are at an evolutionary distance of 1 pam.

Scoring-Matrices

0 = the probability of an AA substitution is purely random

>0 = the probability of an AA substitution is **higher** compared to random (substitution **favoured** by Evolution)

<0 = the probability of an AA substitution is **lower** compared to random (substitution **disfavoured** by Evolution)

F(x,y|**E**) = Probability of Sub. (x, y) during EVOLUTION

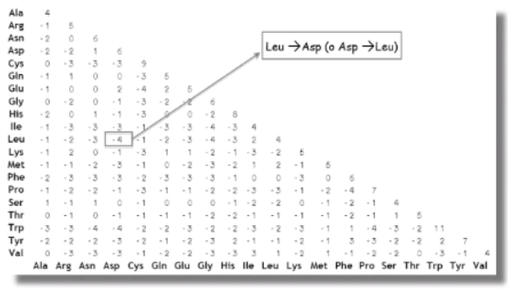
F(x,y|**C**) = Probability of Sub. (x, y) by **Chance**

$$\text{Score (S)} = \text{int}\left[\log_{10}\frac{F(x,y|E)}{F(x,y|C)}\right]$$

Scoring-Matrices

$$F(x,y|E) = \frac{\text{x,y substitutions observed in Nature}}{\text{Total substitutions observed in Nature}}$$

We must observe the Evolution of homologous sequences (2 main models: PAM e BLOSUM)

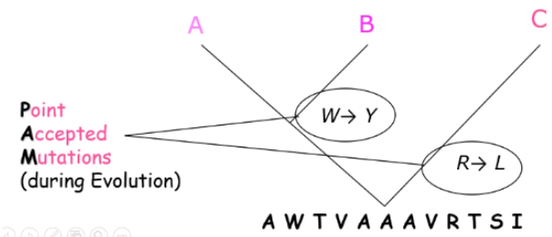


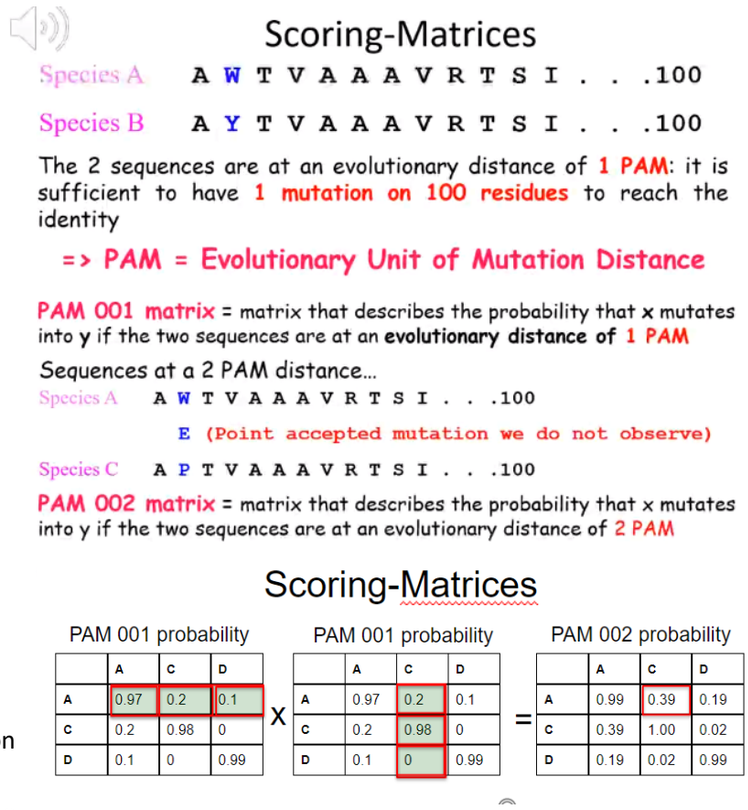
Scoring-Matrices

71 sequence families with at least 85% identity

Species A	A	W	T	V	A	A	A	V	R	T	S	I							
Species B	A	Y	T	V	A	A	A	V	R	T	S	I							
Species C	A	W	T	V	A	A	A	V	L	T	S	I							

Maximum parsimony (MP): choose tree that minimizes number of changes required to explain data





Starting from this matrix of probability, in order to have a full range of matrices to use at different evolutionary distances between two sequencing, Dayhoff was able to develop the other matrices by simply multiplying the first one for itself generating pam 002 probability and so on:

DEMOSTRATION:
if we consider the substitution from cystein to alanine for example in two sequneces at a distance of 2 pam we can made some oservation:
since there are 2 evolutionary steps from the two sequneces we must consider the case in which the mutation happens in the same place that we are analyzing (hidden mutation).
We must consider 4 cases:
the direct mutation fro ala to cys
the conservation of alanine in the first step and mutation in the second
mutaions in the fist step plus conservation of the mutation in the second
mutation from alanine to another residue in the first step plus mutation of this one to cysteine in the second step

A → C A → A → C A → C → C A → D → C

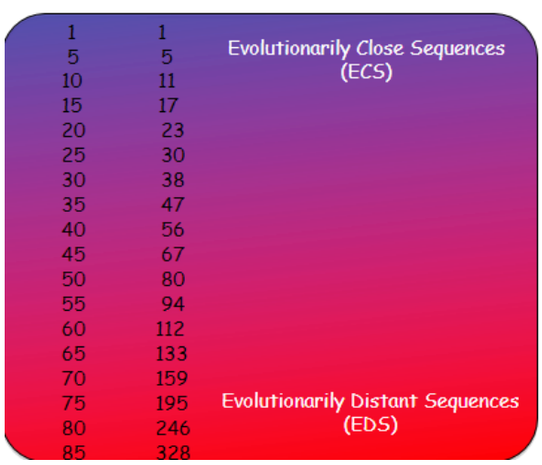
if we consder all these probability and sum them

P(A->C) = P(A-A)xP(A-C) + P(A-C)xP(C-C) + P(A-D)xP(D-C)

P(A->C) = 0.97 x 0.2 + 0.2 x 0.98 + 0.1 x 0 = 0.39

What we obtain is exactly a multiplication of matrix.
Now we can get the scores.
Dayhoff defined the score S of two aligned residues i,j as 10 times the (base 10) logarithm of how likely it is to observe these two residues aligned in nature divided by the background probability of finding these amino acids by chance.
The more two sequences are far from an evolutionary point of view and the higher is the matrix that we should use.
Low PAM matrices describe the probabilities of mutations in evolutionarily close sequences.

High PAM matrices describe the probabilities of mutations in evolutionarily distant sequences.
It may seems weird that two sequneces can have so high distances in PAM. This is due to the hidden mutations. For the same reason the number of pams increases faster then the percentage of difference.



Scoring matrices are used to find the optimal path to be used by alignment algorithm. There is a smart way to identify the best path in alignment and this way is such that we can maximize the score and to use less indels as possible.

DYNAMIC PROGRAMMING

Dynamic programming:
Given two sequences and a series of scores, the purpose of the "game" is to find the "path" (alignment) that maximizes the score, in order to Maximize the probability of obtaining an evolutionary model, compared to the random one (conserved evolutionary -> higher and positive score score)
Now that we have a setted scoring matrix and scoring sequnece we can start to think about the alignment pathway.

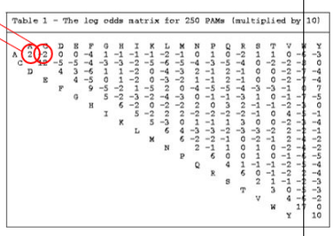
The first box of our matrix in the example is the score of alanine conservation, the second box is the replacement of adenine for cysteine and we assign a score for that and so on.

There are several rules to follow to find the best path:
I rule: I can only move one cell at a time, starting from the first one in the upper left
II rule: can only move diagonally, horizontally or vertically, and gather the score only if I move diagonally
III rule: If I move diagonally,
I pair the corresponding residues
IV rule: if I move horizontally, I introduce a gap in the vertical sequence, and vice versa
V rule: ALWAYS move from left to right and / or from top to bottom (NEVER GO BACK!)
Notice that following the rule the best pathway for our example is like that.

Dynamic Programming for sequence alignment: recipe

	A	C	D	E	F	G	H	I
A	2	-2	0	0	-4	1	-1	-1
C	-2	12	-5	-5	-4	-3	-3	-2
D	0	-5	4	3	-6	1	1	-2
H	-1	-3	1	1	-2	-2	6	-2
I	-1	-2	-2	-2	1	-3	-2	5

Dynamic programming: Given two sequences and a series of scores, the purpose of the "game" is to find the "path" (alignment) that maximizes the score



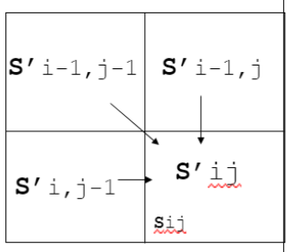
From a computational point of view, we need to consider a particular trick.
Analyzing 4 cell in each point of the cell S'i,j is the initial score of the initial position. S'i-1, j-1 of the previous row and column to with respect to S'i,j and so on. The question is that we need to assign the maximum score which is S'ij. We have in this case three pathway and the score assigned will be the maximum possible according to the three possibilities at disposal.
The basic rules are valid for any kind of string not just biological. But in order to make a sense talking about biological sequences and to consider the evolutional point of view we need to consider the potential presence of indels. In order to understand what indels are we need to understand the role of their presence in the sequences of proteins.

Dynamic Programming for sequence alignment: recipe

	A	C	D	E	F	G	H
A	10	10	10	10	10	10	10
D	10	10	20	20	20	20	20
F	10	10	20	20	30	30	30
H	10	10	20	20	30	30	**40**

A C D E F G H
A - D - F - H
10 10 10 10

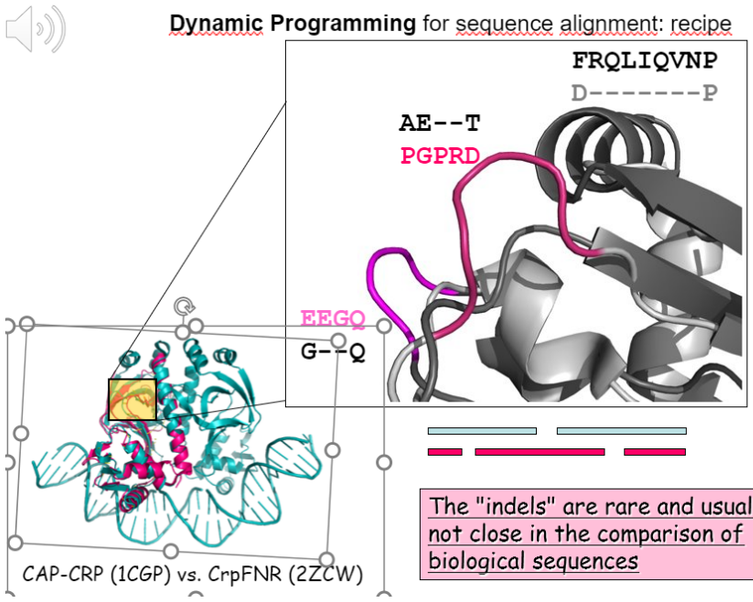
$$S'_{ij} = \mathbf{Max}\begin{cases} S_{ij} + S'_{i-1,j-1} \\ S'_{i-1,j} \\ S'_{i,j-1} \end{cases}$$



Taking into account the superposition of biological proteins structure.by magnified a particular region we can notice that in some position of the two proteins there are some insertion and deletion of polypeptide chain. That is why they are called indels (insertion and deletion). In one case there's and indels made of asp and gly in the other case the indel is made by insertion or deletion of pro and arg and in the last case there is an indels of a whole alpha helix. By observing many of these events into real proteins we can derive a general rule to implement in our algorithm. The rule is the one in pink.

Long story short insertion and deletions are rare in proteins from a biological point of view, and they are not fixed in the evolution process.
Just to consider our observation in the algorithm we need to insert in our algorithm a penalty each time we introduce an indels. This penalty is called GOP (gap open penalty).
GOP value is usually arbitrary. This kind of 'biological' approach is called Needleman-Wunsch algorithm. We can notice that the alignment obtained in this way would be different from the one obtained without taking into account the biological assumption.

Dynamic Programming for sequence alignment: recipe



The "indels" are rare and usually not close in the comparison of biological sequences

One last observation, to be precise and realistic, we need to take into account also the fact that once indels are present (rare event) they can be quite extended. That is because usually indels are usually present in the surfaces of proteins excluding problem within the core of the protein which is related to folding. The algorithm that come from the integration of this observation is called Needleman-Wunsch-Gotoh algorithm. This strategy consists in decrease the GOP each time we are introducing more than one indels one next to the other.

Dynamic Programming for sequence alignment: recipe

	A	C	D	E	F	G	H	I
A	10	5	0	-5	-10	-15	-20	-25
C	5	20	15	10	5	0	-5	-10
D	0	15	30	25	20	15	10	5
H	-5	10	25	30	25	20	25	20
I	-10	5	20	25	30	25	20	35

Match = 10
Mismatch = 0
GOP = -5

Needleman-Wunsch algorithm

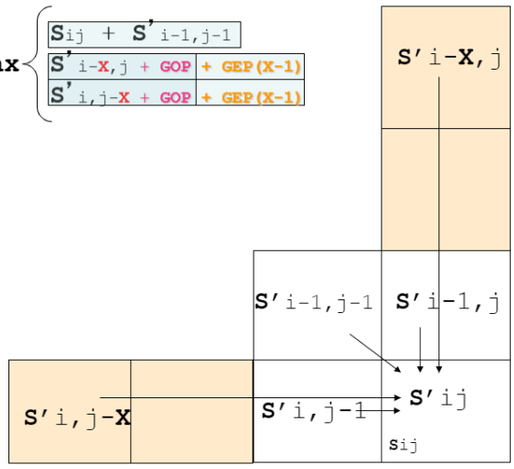
$$S'_{ij} = \mathbf{Max} \begin{cases} S_{ij} + S'_{i-1,j-1} \\ S'_{i-1,j} + GOP \\ S'_{i,j-1} + GOP \end{cases}$$

ACDEFGHI
ACD---HI

Dynamic Programming for sequence alignment: recipe

Needleman-Wunsch-Gotoh algorithm

$$S'_{ij} = \mathbf{Max} \begin{cases} S_{ij} + S'_{i-1,j-1} \\ S'_{i-X,j} + GOP + GEP(X-1) \\ S'_{i,j-X} + GOP + GEP(X-1) \end{cases}$$



Quaternary Structure

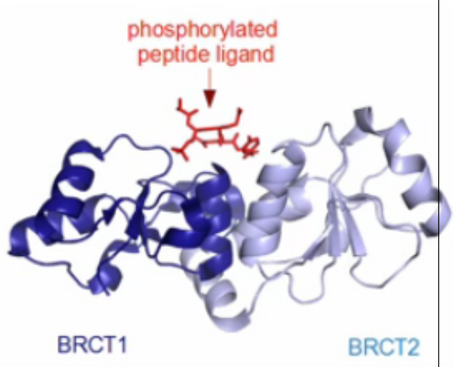
Nomenclature:

A protein domain is an evolutionary conserved part of a protein sequence which can fold and evolve independently from the resto of the protein chain.

A motif is a functional definition, is not necessarily linked to the ability of this region to fold independently from the rest of the protein. proteins with multiple functions usually have more than one domain and are called multidomain protein. They can also have multiple motifs and in this case are called multimotifs protein or both multidomain/motifs proteins.

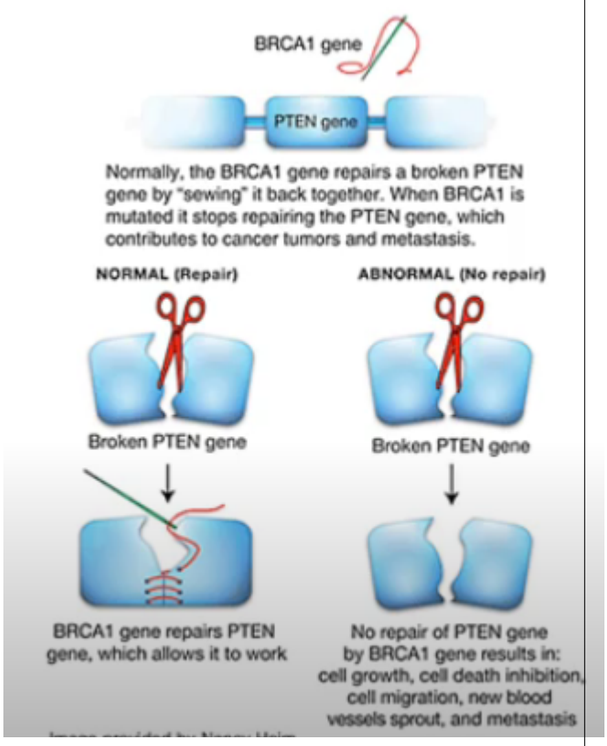
One example is the 'Rossmann Fold' or 'Helix turn Helix'. The former is able to fold independently from the rest the latter is not able. Another example is the MarA which is very simple and has got 2 main function, for this reason this protein is made of a domain and a motif. Each one is associated with a different function: motif is for the interaction with RNA polymerase and the domain is a DNA binding domain.

Modules refers usually to domains or motif which are repeatedly found in diverse proteins, modules can perform tasks which individual constituent are incapable of. One example is the BRCT (breast cancer protein) which is part of a family of evolutionary related proteins. BRCT mutations in proteins are related with breast cancer. This kind of module is found in proteins which are able to bind DNA, it is interesting that tandem repeats of this module allow the new protein to gain novel functions like the ability to phosphorylated linear motif of other proteins (in the case of BRCT the novel binding site is formed between the junction of the two domains) individual units are incapable of this function while two modules repeated can.



PTEN is a tumour suppressor and a phosphatase able to dephosphorylate different tyrosine, threonine and serine residues in protein. PTEN is also a lipid phosphatase which is able to remove the P at 2 and 3 position of inositol ring from

phosphatidyl inositol 3,4,5 triphosphate. BRCA1 is able to repair the P10 gene which is fundamental for many activities, in particular BRCA1 is able to seal back together broken PTEN gene. In case of mutation in BRCA we have an uncontrolled growth of cells which lead to metastasis and tumours.

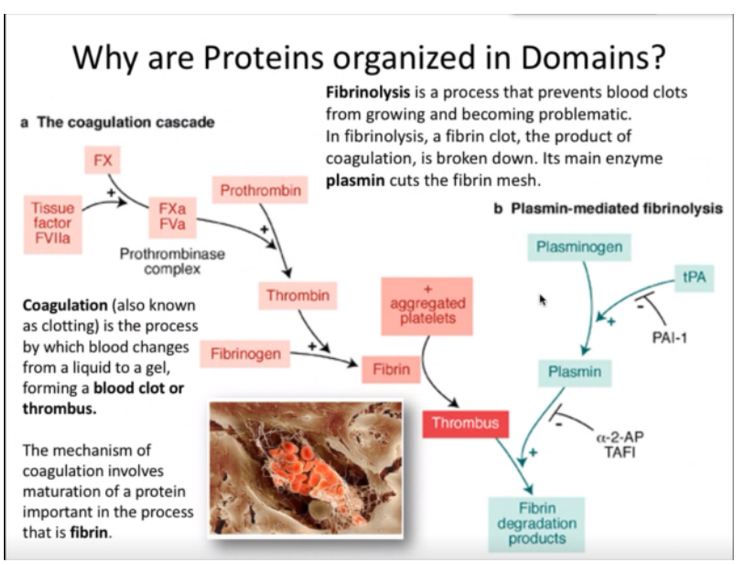


Protein repeats are therefore modules repeated in tandem, there are many kinds of them with different features. Usually, we have two kind of arranged:
Circular folded repeats and liner folded repeats. In the case of the first one, the singular circular repeats form the single blades of the propeller. This is very common in nature and is used to interact with organic molecules and peptides. The linear have the tendency to form semicircles and basket-like structures and usually interacts with large proteins.
Nucleophosmin is an example of circular repeats. Nucleophosmin is an example of multifunctional nuclear protein implicated in cell cycle control, DNA repair, apoptosis due to stress, mitotic spindle and many other functions. The majority of those functions are played through a variety of interactions with other protein partners thanks to each blade which is specialized in a precise interaction with the other proteins participating in a process. This protein is also overexpressed in tumours and is the most frequently mutated protein in the acute myeloid leukaemia.
C.A.T.H. Hierarchical organization of protein. Also name of a server in charge of classifying the domain of a protein

Another important server is the Pfam database which is a collection of protein families and domains, each of which is represented by a domain description and multiple sequence alignment. This server generates also high-level groupings of related entries known as clumps. A clump is a collection of Pfam entries which are related by similarity of sequencies or structures.
For example, PLP dependent aminotransferase superfamily contains a variety of PLP depended enzymes like SHMT, Decarboxylases etc.

WHY ARE PROTEIN ORGANIZED IN DOMAINS?

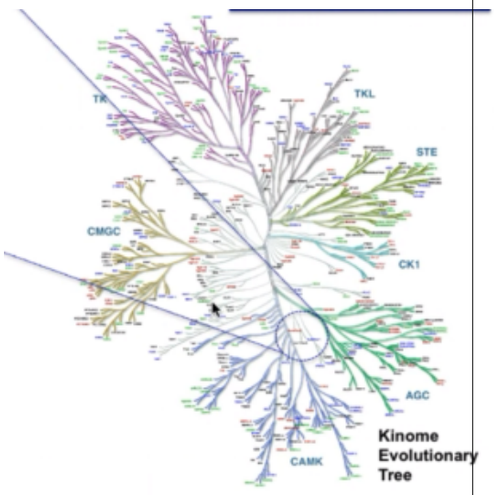
The evolution of a new protein function is too slow by taking into account point mutations compared to domain shuffling (exon shuffling) It happens in nature that various domains are recycled are reused in the formation of other proteins with different functions or gained functions. Proteins can be seen as a puzzle made by different pieces which arose during evolution form different sources. Taking as an example Blood clotting protein we can study the evolution behind this protein. Coagulation is the process by which blood is transformed into a jelly state form a liquid one forming a blood clot and it potentially results from the so called hemostasis.



Mechanism of coagulation involve a series of processes which are activation, adhesion and aggregation of platelets together with the activity of a very important protein called fibrin. During the process fibrin must undergo a maturation carried by proteases. Fibrinolysis is the revers process in a sense that prevents the formation of blood clots too grow too much developing thrombosis. These processes require activation of cascade of different evolutionarily related proteases, created during evolution through genetic shuffling.
The origin of three of such proteases involved in blood clotting is due to genetic shuffling events which is the mixing of domains and modules to compose new proteins. Such shuffling occurs by gross rearrangements such as deletion, inversion, duplication or translocations, homologous recombination during meiosis and slippage of DNA polymerase during replication (usually responsible for duplication).

THE DOMAIN ORGANIZATION IS IMPORTANT TO CREATE NEW FUNCTION AND SPECIFICITY.
Another important aspect of having domain is the allosteric regulation of proteins.
AURORA PROTEINS is a mitotic kinase involved in different phases of mitosis. Such complex proteins are regulated by relatively few numbers of kinases which are clustered in 5 different groups:

- CYCLINE DEPENDENT
- POLO-LIKE
- AURORA
- NIMA-LIKE
- OTHERS

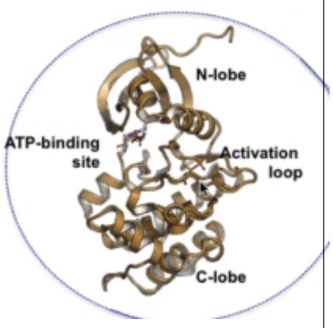


The aurora family is composed by 3 homologous members: Aurora a, b and c.
Kinase evolution can be seen used a KINOME EVOLUUTIONARY TREE because kinases is a very large family of evolutionarily related proteins, homologous proteins. Regarding the Aurora kinases family, we can say that those are serine-threonine kinase which means that they can phosphorylate with ATP serine and threonine of other proteins or of themselves in a process called autophosphorylation.

Aurora family proteins are composed by different regulatory N -terminal domain which are intrinsically disordered low complexity region in which there are important motifs used to regulate the half-life of those proteins.

n-terminal 

c- terminal



The A box of auroraA is recognized by other proteins which can prevent or activate Aurora's degradation in a way that such mitotic kinases are regulated in time and space during mitosis.
The C-terminal part is the catalytic/kinase domain constituted by N lobe subdomain and C lobe subdomain. The active site for ATP is present at the interface of these subdomains. A common feature between many kinases is the so called activation loops which is very long and can undergo large conformational changes during the activation of the kinases.

During G2 of interphase aurora is localized at the level of the centrosome where is important for centrosomes activity and maturation.
During Prophase aurora phosphorylate itself to activate a cascade of polo-like kinases involved in centrosome maturation and separation.
During prometaphase and metaphase, the role of Aurora changes and indeed aurora bind its main coactivator TPEX2 and phosphorylating it in order to migrate to the mitotic spindle participating to its maturation and elongation (spindle assembly).
.
pymol session on aurora […] from 12:10 to 27:17
.
TPEX can open the two lobes of AURORA.
In summary the 2 main function of domains organization are:
First to create complex functionality protein made up of multiple domain each of which is responsible of a simple task. These complex proteins can derive from the process of exon shuffling (Blood clot protein)
Second is to achieve an allostery regulation by changing the reciprocal orientation of the domain within the protein. (AURORA)

Quaternary structure:

Usually, proteins are assembled into oligomers and the single units are called subunits. The reason of this organization is that it presents the same advantage of the domain organization. Evolutionarily speaking is simple to obtain a stable oligomer from monomers. The process of monomer aggregation must be finely tined during evolution cause not always aggregation is stable and can happens that monomers need to disaggregate. There is a fine tuning of the kind of forces at interface between monomers that can be split in two cases:

- The first is the one of the interface of more stable oligomers where we can find residues such as valine alanine glycine isoleucine phenylamine leucine. The common feature is the hydrophobicity, the hydrophobic interactions are sticky and therefore stable. The Hydrophobic interactions once formed exclude water molecules and so there's no need for those aggregates to have polar residues.
- The second case, there are subunits which can assemble and disassemble more easily. In the case they are assemble there re force which assure the interaction, when they are disassembled there must be forces that are able to react with water. Indeed, the residues involved are polar.

By carrying some residues composition analysis at the interface of the monomers we can predict if the interface is stable in the cell and so the subunits are always aggregate together or if the interaction of the protein is dynamic and the subunits can be disassembled.

The reason of oligomerization is the same of the domain organization. In the case of enzymes for example, if we have a monomer usually the active site is exposed to the solvent and this can be a problem for reaction that require hydrophobic environment. In the case of a dimer for example the enzyme can control the active site in order to be open or closed depending by the reaction.

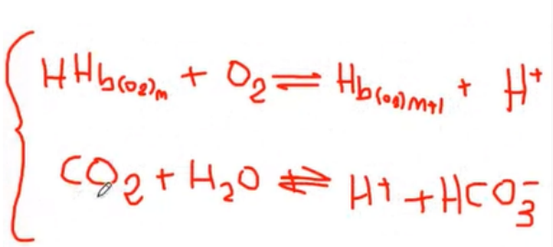
1) The active site, where reaction takes place, can be buried from solvent. This is thermodynamically favourable and avoid side reaction with solvent

2) The active site entrance can be finely regulated

1HVR protease is a great dimer example of the open and closed active site controlled by the protein itself.

HEMOGLOBIN

Cooperative binding can be achieved only in the oligomer form. Myoglobin saturation curve is hyperbolic because is made only by 1 monomer, this means that it can't have cooperativity. The saturation curve of hemoglobin is sigmoidal.
At low oxygen concentration hemoglobin is able to release O2 while at high O2 tension hemoglobin can be charged. This is the base of the Bohr effect which is the fact that when hemoglobin is bound to a certain number of oxygen and another O2 is added, hemoglobin is able to release a proton shifting the equilibrium of CO2 to the right.
In periferical tissue the euilbrium is shifted to the right, viceversa in the oxygenated tissue



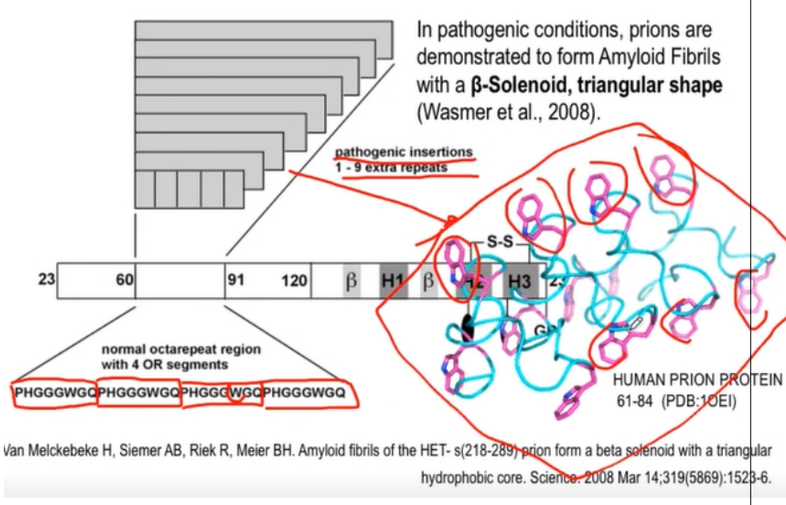
ABBERRANT OLIGOMERIZATION

Is the process, which is related to aberrant aggregation of proteins, misfolding of proteins and very dangerous human diseases. One of these diseases is the sickle cell anemia in which some residues of the surfaces of hemoglobin are mutated to some other residues such as glutamate 6 to valine. The function of glutamate is to prevent aggregation of hemoglobin, when glutamate is mutated, fibrous structures are formed. Such fibres change the shape of red blood cells which are not elastic anymore and become very fragile. This results in anemia which is a low number of red blood cells.
The heterozygote for this disease is protected from malaria and the sickle cell anemia symptoms will be present only at certain altitudes where there is low concentration of oxygen. The heterozygous individual is present in population where malaria is highly present.

Aberrant oligomerization diseases are related usually to many neurodegenerative diseases like Parkinson or Alzheimer etc. the diseases are very different, but they share common features such as the presence of beta helico fibres. The identity of single protein responsible for the diseases are different but the process is always the same i.e., protein aggregation, misfolding and exposure of hydrophobic beta strand which tends to aggregate to form such kind of toxic fibres.

Such amyloid fibres are organized into filaments rich in beta sheets, the process is usual divided into two different phases: Lag phase can take years and is the process by which you have a physiological equilibrium between oligomers and monomers. Due to the toxic environment of or cell we can have the partial denaturation of monomers and the formation of hydrophobic sticky surfaces. Those surfaces can aggregate forming amyloidogenic aggregation in a way that they can polymerize and create amyloid seeds. At this point is too late and from the lag phase and enter in the Growth phase where seeds polymerize and form protofibrils which generates mature fibrils.
One of the best characterized amylogenic diseases is the Prion protein. A normal prion is present in our brain and the function is a mystery. The normal protein is made up alpha helices but due to genetic or environmental reason is that there is helix to beta helix shift in the prion protein.

Such beta helices can aggregate to form fibres. A single disease-causing prion that we can intake from food can catalyse the transformation of a normal protein into a disease-causing protein in an exponential way.
This is one of the few diseases transmitted by a single protein. The structural reason of the shift is the duplication of a particular region of the prior, composed by 4 repeated motifs (IN RED) and those are usually folded in alpha helix, when we have the pathogenic insertions of other repeats 1-9 we assist the change of the alpha helix to a sort of beta helix where the imidazole ring of the tryptophan residue exposed. Tryptophan residues are very sticky because are hydrophobic and can for stacking interaction and that's the reason of the aggregation of prion into amyloid structure.
aggregation of prior amyloid structure.

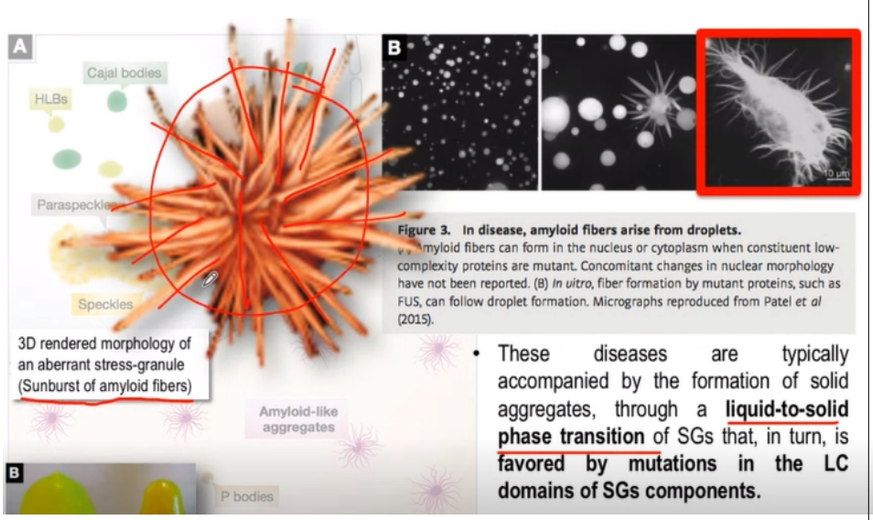


Van Melckebeke H, Siemer AB, Riek R, Meier BH. Amyloid fibrils of the HET- s(218-289) prion form a beta solenoid with a triangular hydrophobic core. Science. 2008 Mar 14;319(5869):1523-6.

The kind of disease depend also on the region of our brain in which the association happens. In any case he disease is very severe.
Kuru disease is a prion disease that manifest when a person eats the brain of the dead people (cannibalism associated)

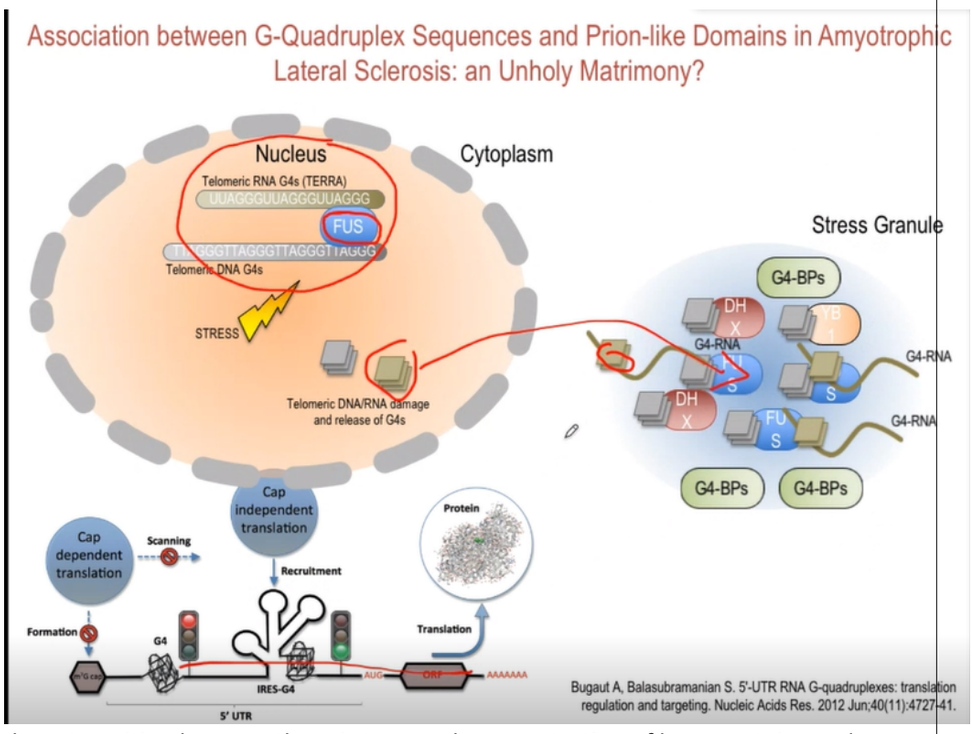
AMYOTROPHIC LATERAL SCLEROSIS
Amyloid lateral sclerosis is a disease in which we have the death of nerve cells which are associated to muscles. This disease is responsible for the progressively weakens of muscles until you die suffocated.
One of the symptoms is the presence in our cell of a form of particular stress granules called sunburst. Stress granules is an example of membrane less organelles, those can be regarded like liquid droplets Such organelles are used to carry out biochemical reaction which must be disjoined from the rest of the cell in time and space. They miss the lipid bilayer of membranes and therefor can form and dissolve in a very fast way. The trick adopted is that they undergo liquid-liquid phase separation (LLPS) like water with oil but in this case the nature of this reaction is the different nature of forces associated in protein with rna interaction because the membrane less organelles are always associated to rna processing and rna is a major component of such organelles, therefor the interaction that we are talking about are association with rna and particular proteins which are able to enter into the membrane less organelles (only those)
Most of these proteins are the ones with low complexity regions (sequence with low variation of amino acid) related with IDR. In particular the proteins able to enter are rich in arginine and tyrosine because those are the best residues for the interaction with RNA. Tyrosine is able to form stacking interactions and H-bond while arginine cation-pi interactions. Also, phenylalanine is present but in a less extent. The other proteins are excluded.
Stress granules are formed only in stress condition like UV radiations. The problem is that in certain condition when this process is repeated over time, the stress granules are not able to dissolve and undergo a process called liquid-solid phase separation which means that at the beginning they are liquid but if the stress is prolonged the granules become solid and tend to create sunburst of amyloid fibres which is pathogenic stress granules and is toxic for neurons.

This disease is associated with many different proteins but One common features of this proteins is that thy are in low complexity regions and that those are able to interact with g quadruplex structures (Hogsten interactions). Why arginine and tyrosine interaction interact with such complex region? Arginine can interact with nucleotide and P but some sequence motifs such as GYG or GWG (prion protein) which means that such proteins are similar to prion proteins.

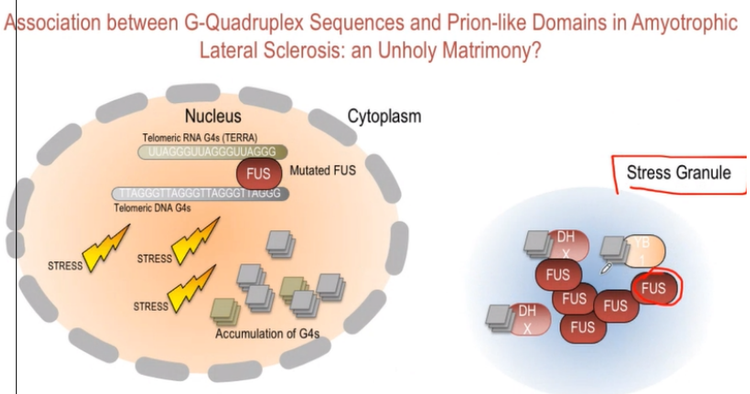


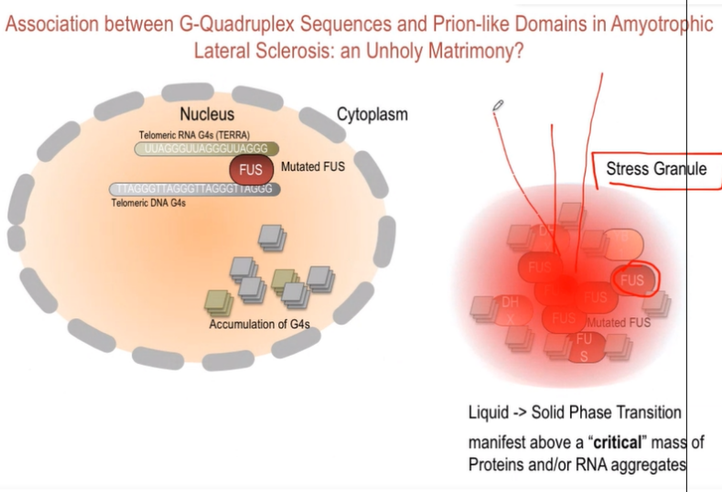
FUS protein is very important in this process and interact with telomeric DNA/rna quadruplex regions within nuclei and is composed of large number of low complexity regions. In normal condition you have stress which is associated to a damage to telomers of DNA or rna of chromosomes and some of such regions are able to be bound by such proteins which is a

sort of shuttle system between nucleus and cytoplasm and this protein is able to shuttle the quadruplex into cytoplasm signalling that there is something wrong in the nucleus, this signal aggregate other proteins able to interact quadruplex regions and is able to create a stress granules in which other rna involved in stress response with quadruplex regions are aggregated and activated. Stress is a signal to activate translation of RNA implied in stress response and so is a physiological process



In repeated stress condition or when FUS is mutated there are too much intrinsically disordered proteins in the granules and since those are able to form fibres they undergo liquid to solid phase transition because there is too much concentration of hose proteins and so there's formation of fibres and sunburst responsible for amyotrophic lateral sclerosis.





X-ray crystallography

Practical part

(Pymol⇒ 3jaf-3rd scene 00:00 to 4:00)

(referring to bacteriophage_emd_6068_volume and emd_6324_volume, which show the electron density) Structure is resolved by cryo-electron microscopy, now with the evolution of cryo-electron microscopy we can obtain the kind of resolution (that can be observed with this PDB) that atoms and the macromolecular complexes

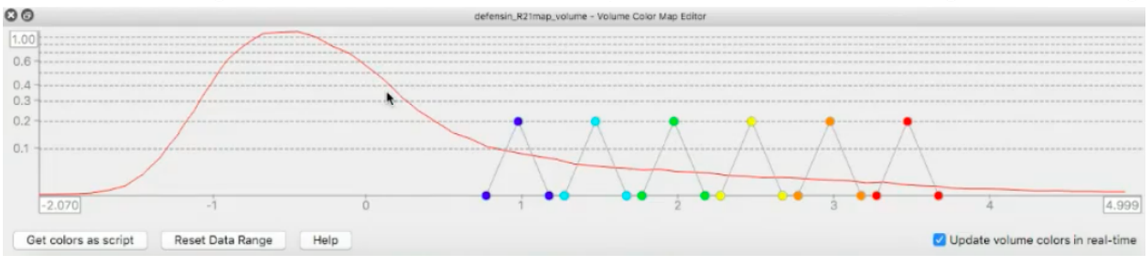
(referring to 3jaf by itself) this is the kind of the resolution obtained through x-ray crystallography and NMR, we are able to observe the structure and position of atoms in space at atomic resolution. What can be seen is only a model of experimental data because the source of experimental data in x-ray crystallography is the electron density. The aim of the crystallographer is to observe the distribution of electron density around atoms and the main work is to map the position of the side chains of the atoms into the electron density

(Pymol⇒ defensin_R21map and defensin_R21 from "files for practical", 5:30-30:00)
start from electron density map and try to model the positions of the side chains of the arginine21 within the defensin protein (like a crystallographer, obvs In the case of a crystallographer the task is repeated for each of the side chains and main chains of the protein)

We can download the electron density map of any protein from Protein data bank and also when we "get PDB" within Pymol we can select "2FoFc Map" before downloading. Once we open the document, the object will look like a white cube, called unit cell. In crystallography, for each point within the cell, we know the electron density and we can see the electron density by clicking on action⇒volume⇒… (there's multiple ones but the best one is rainbow2, it will create the object defensin_R21map_volume) . The electron density map is at very high resolution, we can easily map the position of the atoms within the map (not always the case but for the map in the tutorial it is). Once we open "defensin_R21" we can see how the map and the structure itself matches with each other. Select arginine22 within the protein, there's no side chain in the structure but it can be seen in the electron density map, so we will try to build the side chain from the map wizard⇒ mutagenesis⇒ protein. No mutation⇒ mutate to ARG⇒ click on C-alpha of the arginine22

It won't fit the electron density map right away, bc it will give us the suggested rotamer. We can look through all the available rotamers to check if any of them match the density map⇒ the second one is good, so we click on "apply" and "done". To make the atoms match the density map perfectly we go into mouse-editing mode. By pressing control+right button of the mouse and sliding the cursor around the C-beta (dihedral angle) we can change the position of the side chain (usually done in a hierarchical way, so we move C-beta to optimise the position of C-gamma and so on). We can do the same with the other dihedral angle at C-gamma and once we find an optimal position for the C-delta we go on with the chain until the last atom. Once satisfied we can save the file with the new position of the side chain
(this process is done to make any PDB file that we can find, from x-ray crystallography data)

Electron density: there are some units within the cell that are more or less dense of electrons, we can observe the difference in density by clicking on the color besides the volume object⇒ panel. It will give a graph of the normalized distribution of the electron density within the unit cell (a gumbel distribution, distribution of extreme values: a normal distribution with a tail on the right). The peak is the mean value of electron density within the cell and the numbers below represent the standard deviations away from the mean, each value corresponding to a different color within the map; we can change the intensity and value of each color by moving the dots of each color (if we want to highlight a specific value within the map). Obviously the closer to the atom, the higher the electron density. When looking at a map, we aren't interested in the regions with the mean electron density but we're interested In the regions with the highest density



X-ray crystallography (theory)

Importance of x-ray crystallography: Watson, Crick and Wilkins received the Nobel Prize in Physiology or Medicine for their 1953 determination of the structure of deoxyribonucleic acid (DNA), derived from a picture of x-ray crystallography data (diffraction map) that was stolen from Franklin (the initial model of Watson and Crick was completely wrong)
[video (36:30-41:00, 41:20-44:00) with explanation that can be watched from slides which explains the technique of fourier transform (which is able, starting from waves, to obtain an image and then from the image to get waves again) and how the technique can be used to tell if one image is random or not random. Fourier analysis states that any image can be decomposed into a very simple summary of sine waves with peaks and throws, the more waves we add the higher the resolution of the image]

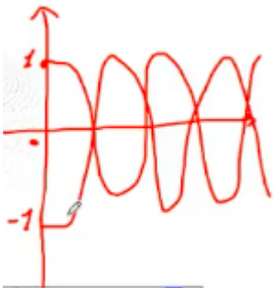
[video from youtube "Fourier Image decomposition and reconstruction" (44:34-47:36); when we convert an image from a heavier format to a lighter one we get rid of the sine waves that define the small details of the image]
[website "an interactive introduction to fourier transforms" (48:00-); demonstration shows that a unidimensional wave can be decomposed by Fourier transform into elementary sine waves; any monodimensional function that represents a unidimensional image (not just visual ones) can be converted in sine waves. If we look at the complex sinusoids waves we can notice they work in the exact same waves and instead of rotating in a circle they rotate in whatever other shape the image has but they can be simplified into sine waves.]

Fourier transform is important because electron density is an image and we're able to rebuild this image by using the raw data which are information regarding waves and we put them together in the Fourier transform to rebuild the image
images⇒ waves

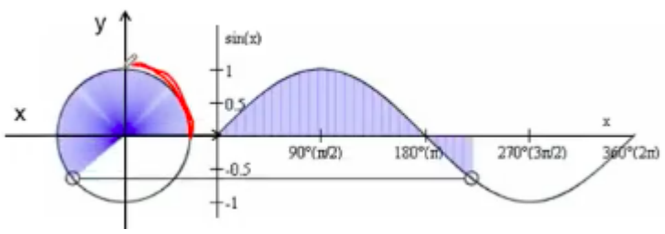
Even our brain behaves like a Fourier transform: we get information from the outer reality through light and light is made of waves, what we see is light scattered from objects and what our brain sees are not the objects themselves, they're the waves.

Fourier transform⇒ we transform 3D images into waves
Inverse fourier transform⇒ we transform 3D images into waves and then back into images
X-ray crystallography works in the same way⇒ we collect waves which are diffracted from crystal structures of proteins and we collect those waves, put the waves into an algorithm for computing the inverse Fourier trandof=m and rebuild the image

What is a wave? A wave is an undulatory phenomenon (perturbation of something through space, like water or electromagnetic) that propagates through space and time and is regularly repeated
The wave function (function that describes a bidimensional wave, like the one in example at 56:00)
$Y = A \sin[2\pi(hx - vt) + \varphi]$ A= amplitude, the size of the peak starting from 0. 2pi is a constant, h is the wave number (number of peaks per unit of space), x is the position, t is time, v is frequency (peaks in unit of time) and last variable is phase (the value of the wave function at 0, when it starts)



these two waves are counterphase as they start at opposite ends
In crystallography we work with waves that are time-independent: stationary waves $Y = A \sin[2\pi(hx) + \varphi]$
radiance: $\alpha = 2\pi(hx)$, the way to describe a wave going in circles, each radiance is a "slice of a circle"
If we have ¼ of a trajectory of a circle (a wave) it's pi/2 and if we have half of the circle it's pi
The relationship between the radiant representation and the linear representation



$Y = A \sin[\alpha + \varphi]$: the phase of the wave is 0 and the amplitude is 1 so we can simplify $Y = \sin\alpha$
In crystallography we don't represent the waves as sine waves because we have to sum many waves, it's better to sum vectors so we don't have to use trigonometry; so we represent waves as circulating vectors
To describe vectors we need polar coordinates while relying on a plane that is the complex plane and use this function, which represents a vector: $\underline{Y} = \bar{A}\, e^{-i[\alpha - \varphi]}$
Wave vectors are represented in a particular plane, in which you have real numbers and imaginary numbers, the vector is the sum of the cosine of the radiant and the sine the imaginary space:

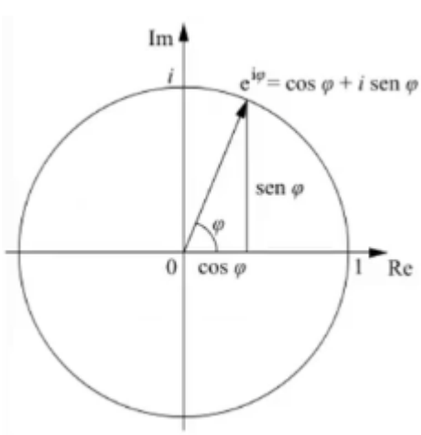
$$e^{iz} = \cos(z) + i\,\sin(z)$$

$$e^{z} = 1 + z + \frac{z^2}{2!} + \frac{z^3}{3!} + \cdots$$

$$\cos z = 1 - \frac{z^2}{2!} + \frac{z^4}{4!} - \frac{z^6}{6!} + \cdots$$

$$\operatorname{sen} z = z - \frac{z^3}{3!} + \frac{z^5}{5!} - \frac{z^7}{7!} + \cdots$$

$$e^{iz} = 1 + iz + \frac{(iz)^2}{2!} + \frac{(iz)^3}{3!} + \frac{(iz)^4}{4!} + \frac{(iz)^5}{5!} + \frac{(iz)^6}{6!} + \frac{(iz)^7}{7!} + \frac{(iz)^8}{8!} + \cdots$$



Any function can be represented as a polynomial function (Taylor expansions) (we don't need to know this for the exam)

If we put the imaginary numbers into the polynomial we get: $= 1 + iz - \frac{z^2}{2!} - \frac{iz^3}{3!} + \frac{z^4}{4!} + \frac{iz^5}{5!} - \frac{z^6}{6!} - \frac{iz^7}{7!} + \frac{z^8}{8!} + \cdots$

$= \left(1 - \frac{z^2}{2!} + \frac{z^4}{4!} - \frac{z^6}{6!} + \frac{z^8}{8!} + \cdots\right) + i\left(z - \frac{z^3}{3!} + \frac{z^5}{5!} - \frac{z^7}{7!} + \cdots\right) = \cos(z) + i\text{sen}\,(z)$

This is a Demonstration that the radiant waves and the linear waves functions are the same

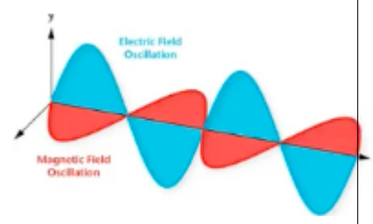
If we replace the radiant $\alpha = 2\pi(hx)$ with the linear representation we get: $Y = A\, e^{-2\pi i[hx - \varphi]}$, a stationary monodimentional wave

water wave has 3 dimensions: $Z = SIN(SQRT(X^2+Y^2) - Vt)$ time, x an y while a stationary one wouldnt'have time and would have 2 dimensions: Euler 2D wave $Z = A\, e^{-2\pi i[hx + ky - \varphi]}$)

The difference between the function monodiension and two dimensions (2 variables plus time) is just one additional variable

if we want to cope with a 3D wave (all the waves we perceive are 3D) we "just have to add another variable"

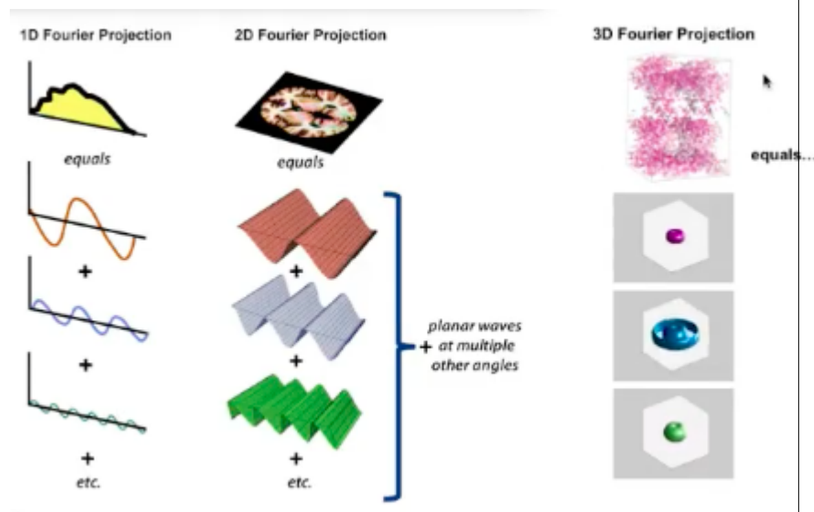
Euler 3D wave $\rho = A\, e^{-2\pi i[hx + ky + lz - \varphi]}$



When you hit the electron with x-rays you cause a scattering of 3D waves from electrons: electromagnetic waves (perturbation of electric and magnetic fields)

We have different kinds of electromagnetic waves according to the wavelength (distance between peaks of the waves, reciprocal of the frequency)

Most important formula in x-ray crystallography, function defining the electron density in a point of the unit cell given by the coordinates x, y and z: the electron density in a given point in space is the summary of waves in that point $\rho(xyz) = \sum_{-\infty}^{+\infty} A\; e^{-2\pi i[hx+ky+lz-\phi(hkl)]}$ ⇒ sort of fourier transport: electron density (a kind of image) can be rebuilt by knowing the 3D waves which are diffracted by that point.



2D projection is used in NMR in medicine, while an example for 3d projection is electron density, which is an image made of 3D waves. Another example is vision where we collect the 3D waves of light (electromagnetic) and our brain makes an inverse fourier transform to rebuild the image from the waves, crystallography works the same way: we start from the diffracted waves from the atoms, collect the waves and we put the info in an inverse fourier transport and we obtain the electron density.

Other way of writing the electron density formula, by normalizing it with 1/V, which is the volume of the unit cell (the cube we've seen in Pymol); in which case the electron density is the density of electrons normalized by the values into the volume $\rho(xyz) = \frac{1}{V}\sum_{-\infty}^{+\infty} A\; e^{-2\pi i[hx+ky+lz-\phi(hkl)]}$

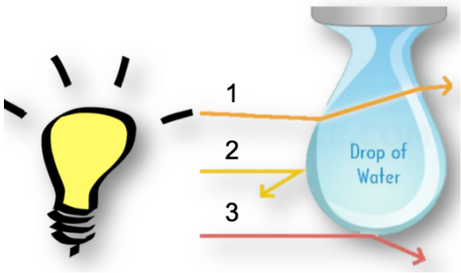
The amplitude in crystallography is called "structural factor" and it's represented this way: $\rho(xyz) = \frac{1}{V}\sum_{hkl}^{+\infty}_{-\infty} |F(hkl)|\,.\, e^{-2\pi i[hx+ky+lz-\phi(hkl)]}$ and it's the amplitude of the diffracted x-rays

At each point, the density is the sum of the interference of the different waves passing through that point

Xray Crystallography pt. 2

There are three ways an electromagnetic wave can interact with water:

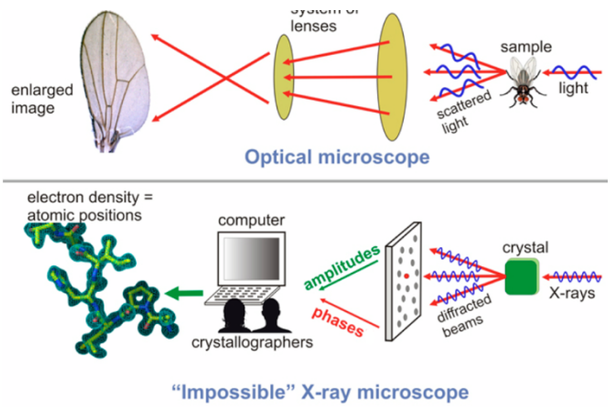
1. Refraction → The bending of light ray when entering a medium with a different constant of light speed.
2. Reflection → bouncing from the surface of the object back towards the source.
3. Diffraction → the bending of light when interacting with objects of size comparable with the light wavelength.



Diffraction and refraction are at the root of the function of the microscope. Light passing through tiny objects is diffracted and then refracted in a system of lenses, producing enlarged images showing the tiny details of the objects. The specimen must be very thin in order to avoid reflection of the light waves. In the first step, the scattered light from the object is refracted in the lens of the microscope, because the light is entering a medium of a different density. In this way, the image is magnified and the image we see in the eye piece is enlarged.

Why is it impossible to use a light microscope to see at the level of atoms? The problem is the maximum enlargement we can have for a light microscope is about 2000x. This limit is not a technical problem, but instead is based on a physical phenomenon called diffraction limit. This rule states: you cannot image things that are much smaller than the wavelength of the light you are using (diffraction limit). We have seen that light waves are diffracted by objects about the same size as the wavelength. If the object is too small, the light does not diffract, and the object can be said to be invisible for the light.

This is the reason why when we observe objects in a microscope, since visible light is 400-700nm, we can see objects such as E. Coli and bacteria, but no more than that. If instead we want to observe objects the size of Armstrongs, such as 0.1 nm we must use wavelengths of a similar length, in this case X-Rays.



With x-ray crystallography we observe atoms by experimentally determining their electron density. There is no system of lenses in this approach which can rebuild the enlarged image, such as in the case of light microscopes. For this reason, x ray requires computational power to rebuild the image from the electron density map. Based on the amplitudes and phases collected from the diffracted beams of x-rays passing through a crystal, a computer can generate atomic positions based on calculated electron densities. While in the case of light microscopes we have a sample on a thin film. Instead, in x-ray we use a crystal, the formation of which is the most difficult part. This crystal acts as an amplifier, a single specimen is too small to properly diffract x-rays in the right way, but in the crystal which has all the molecules positioned in the same way and in the same position, the crystal acts as an amplifier for the x-rays in a way that they can sum together into the same phases and be impressed in a diffractometer – the object used the measure the scattered x-rays from the diffracted crystal.

The diffractometer serves as a recorder for the amplitudes and phases of the diffracted waves, which a computer can take and use Fourier transforms to rebuild the atomic positions.

XRAYS

When a moving electron changes direction, it emits energy. When the electron is moving fast enough, the emitted energy is at the x-ray wavelength. In order to accelerate the electrons at the right velocity to emit x-rays, we need a large machine called a synchrotron. Synchrotrons are one of the most expensive tools every built by humans, usually used in physics to study subatomic particles. The large ring of the synchrotron accelerates the electrons extremely high energy and then makes them change direction periodically. The European synchrotron radiation facility (ESRF) produces x-rays 100 billion times brighter than the x-rays used in hospitals. A beam of x-ray is passed through a channel, and columnated after which it hits the crystal and defects its beams on a diffractometer. Other rooms exist in the same facility to constantly analyze the data collected to rebuild the electron density. During this process, the crystals are constantly rotated, resulting in a result which has a 3D image instead of the 2D image obtained from any single x-ray beam. This 3D result is called a diffractogram and allows the capture of all the molecular complexities within the crystal.

We do not have time and I will not enter into the details of the mathematical and computational details of how the diffractogram derives the amplitudes and phases of the waves, but it is possible from every spot to get these two values.

DEFFRACTION OF CRYSTALS

The biochemical procedure for obtaining crystals is very hard, this is the reason we do not have a three-dimensional structure of all the proteins we have sequenced. This gap is due to the inability to crystalize every protein we can purify. This process is trial and error starting from a seed of nucleation. From the seed we slowly increase the concentration of

our protein. Slowly in order to prevent aggregation and precipitation in an unordered way. The protein is mixed with a precipitation which is a Calotrophic agent that can start the process of aggregation. The concentration of protein and precipitant must be balanced to reach the nucleation point, which is where we have the first seeds of the protein. From this seed, the other proteins must aggregate in the same orientation and same order. To get to the nucleation point we have to play with the pH, ion concentration, temperature etc. in order to avoid the precipitation. The bounders between undersaturated state, crystal growth state, nucleation state and precipitation state are very thin, so we have to move extremely carefully with the physical chemical properties to avoid precipitation, this is why the process can even take years.

The most used technique in crystallography in order to increase protein concentration and precipitant at the same time is called vapor diffusion. A droplet containing purified protein, buffer, and precipitant can equilibrate with a large reservoir containing similar buffers and precipitants in high concentration. In this way, due to an osmotic process, water is balanced between concentration of protein and precipitant and slowly increases as the time proceeds, if we are lucky this results in a small crystal growth in the droplet.

Often a set of vials are used to try different conditions, which change conditions smoothly so containers next to each other will have similar but different conditions. This allows trying multiple conditions for the crystal growth at the same time, by adding buffer to the reservoir, and purified protein plus reagent to the top of the vial. Now, we have robots that set up such screens which test multiple conditions simultaneously to grow the crystals on the structure.

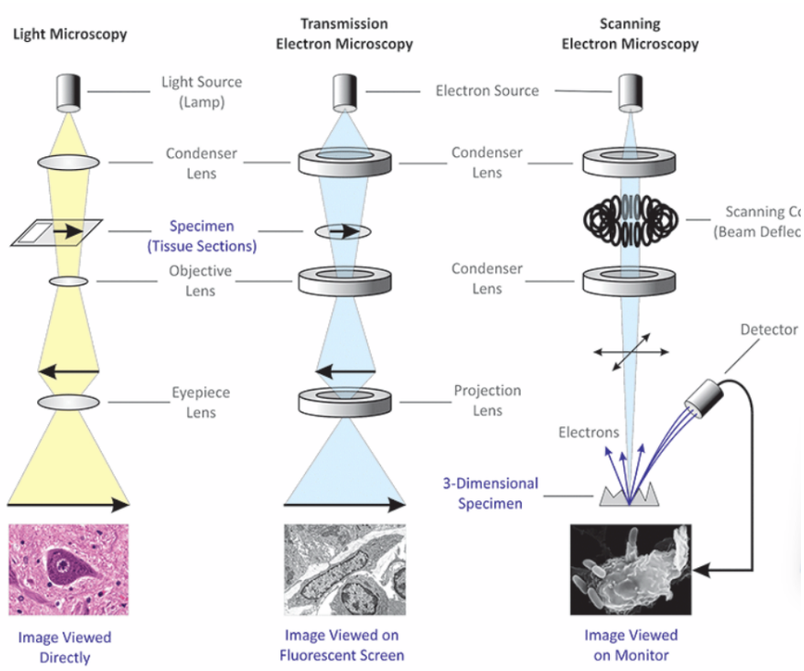
We will skip the physical analysis of the diffraction of crystals. (Slide 51)

ELECTRON MICROSCOPY

This technique used for solving 3D structures of proteins is quite new, and rapidly improving so we can solve many complexes that are impossible to solve via x-ray crystallography. Electron microscopy uses a beam of accelerated electrons instead of the electromagnetic wave of light. For each particle we can associate a wavelength, in the case of electrons this wavelength is tiny, 0.01 A. This means in theory we can use electrons to magnify a sample about 10 million times, compared to 2000 in a light microscope. Technical issues still make this not yet possible.

Transmission electron microscopes contain almost the same components as a light microscope, the light source (lamp) is instead an electron source. We also contain lenses to scatter the lights direction, in this case a magnet used to deviate the path of the electrons and focus them on our specimen, which must be very thin to avoid refraction of electrons as the same in light microscopes. The image we observe is similar to that observed in light microscopy, except in the case of TEM it is black and white.

A modified version of TEM is SEM, scanning electron microscopy. In this case instead of a 2D projection of our image, we can use a scanning coil, or a beam detractor, which is able to deviate the path of electrons in a way that we can obtain snap shots of our image from different orientations, and computer the 2D projections into 3D images using software to obtain stunning detailed images. Images from SEM are often colored manually post imaging, but the output is black and white.



Another recent modification of this technique is called cryo-electron microscopy. The principle is the same as TEM and SEM, the only difference is the preparation of the sample, we use a process called vitrification. In vitrification we build a very thin layer of our sample and freeze it, so the molecules are not moving. A grid is then imposed, and the sample is processed by SEM or TEM to retrieve the 2d projections. Next, a machine learning algorithm can cluster the images from the sample, the rebuild a 3D map, which can be used by a structural biologist to rebuild the molecule.

In pymol, we looked at the membrane channel 3J5P, used to sense heat(spice) and itch in humans. Membrane proteins are especially hard to be solved via x-ray crystallography, as they are ingrained in the bilayer, but such proteins are very important in pharmaceutical research. Fetch the PDB in pymol, then go. to PDB and search 3J5P, click download files in top right, and download EM Map. The resolution of the electron microscopy map is higher in electron microscopy than x-ray crystallography. Decompress the downloaded file and drag and drop the EMD map into pymol. In the case of cry-EM we do not use volume information as in the case of x-ray crystallography, but instead use iso-potential surfaces,

obtained by clicking action of the object, surface and choose a level higher resolution is the highest we can obtain, in the example we choose 3 and can now observe the 3D rebuilding of the channel. What is nice in cryo-EM, is the electron density is built from real images of the molecule, not a Fourier transform of the diffracted x-rays. The details are less in cryo-EM compared to x-ray crystallography, but this is improving, and the procedure is more reliable.

A small amount of purified protein is put into a grid, which is loaded into a machine called a vitrobot, which freezes rapidly the sample. The result is a very thin ice layer, such that the size of the layer is approximately the size of the molecule. Each molecule is shot with electrons, and the data we obtain is the shadows of such molecules, all in slightly different orientations. The software clusters molecules and tries to obtain a clear image of a side of the molecule. This process is repeated for each orientation, and the result is a 3D map built from such sides, first by building the mainchain then the side chains of our molecule. Since 2013 when the technique began, the resolution of cryo-EM has drastically improved, even to the same level of resolution as x-ray crystallography. Richard Anderson won the Nobel prize in 2017 for his improvement on this technique.

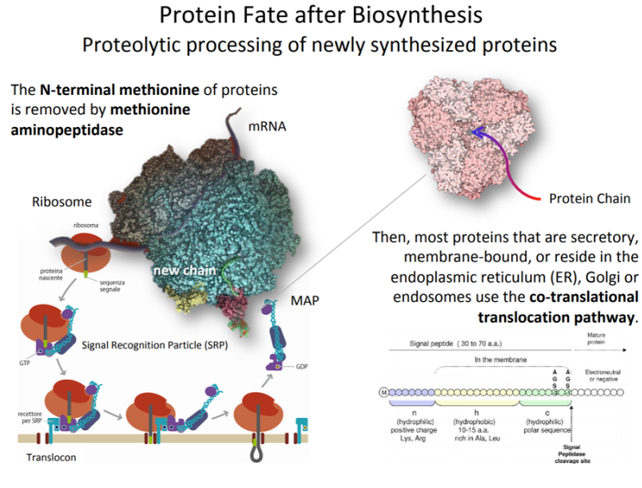
PROTEIN FATE AFTER BIOSYTNHESIS

Sometimes proteins are not ready as soon as they are translated. Proteins need to be modified in a covalent way to achieve their final function. THE DIVERSITY OF MODIFICATIONS is long and its ability of phosphorylation, ubiquitination and so on to regulate signaling and protein turnover.

We have discovered a plethora of other modifications such as sugars modification acetylation and so on. Thanks to those is possible to target proteins in a specific target protein in a specific cellular location and to coordinate extert dynamic control over protein function in a diverse biological context. On most cases protein function is obtained after those modification. Is a covalent addition of functional group but it also refers to proteolytic processing and maturation.
In this regard one of the most fundamental query about modification is the extent to which different modifications which can be reversible or not can increase the complexity of proteins. Thanks to human genome project we can now estimate the number of protein coding gene in a range of 20000. It means that each protein can be modified differently in a combinatorial way. We need also to consider the different combination of modification on each protein and to associate the modification to an output of functionality and each gene can be spliced in different ways. This of course increase proteome complexity to over 1 million.

MAP

N-terminal methionine is removed post-translational modification by a precise enzyme called methionine aminopeptidase map or by cleavage of n-terminal signal peptide used for localization of proteins. Signal peptide referred as signal sequence etc is a short sequence of peptide present at N-terminal of newly synthetized proteins that are destined to secretory pathway. These proteins also include proteins that are in the ER, Golgi and endosomes or proteins which are inserted in cellular membranes. The core of signal peptide is a long stretch of hydrophobic residues that usually forms an alpha helix. (that why is called H sequence) many signal peptides being with a short sequence of positively charged residues which may have to enforce proper topology in the polypeptide translocation by the positive inside role according to which the positively charged residues are located into membranes.



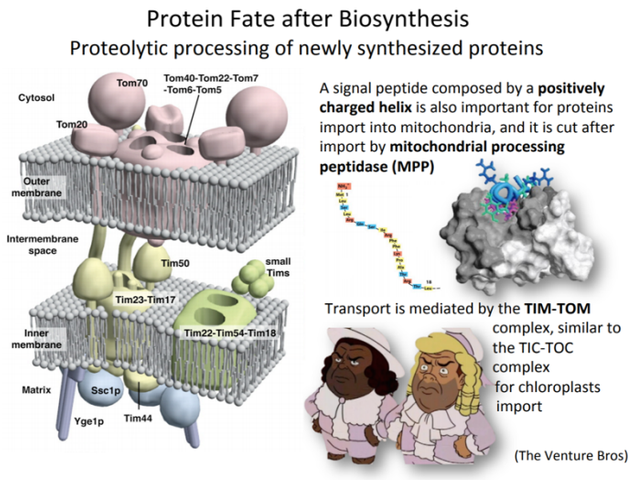
The cotranslational pathway starts as soon as the signal peptide emerges from the ribosome and is recognized by signal recognition particles or SRP. When SRP binds to signal peptide can halt the translation (of course only in eukaryotes). SRP direct the sequence ribosome mRNA complex to the SRP receptor located on the surface of the ER. Once this membrane targeting is completed the signal sequence is inserted into the translocon. In this way ribosome are physically docked into the cytoplasm of translocon.

Now protein synthesis can start again. SRP is a cytosolic ribonucleoprotein. The process of protein translocation happens during elongation and therefore we talk about elongation stop.

NET

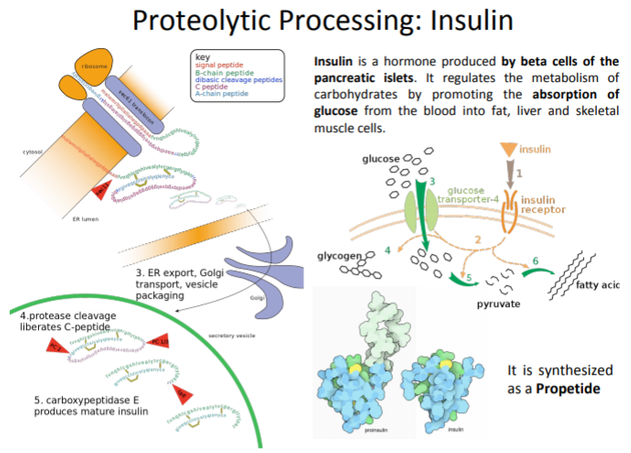
Another n-terminal process beside the trimming of n-terminal is the N-terminal acylation carried by n-terminal acyl transferase or NET protein which usually after the removal of the methionine by MAP the new terminal is stabilized by acetylation and is very useful in terms of protection from hydrolysis. A Signal peptide composed completely by positively charged residues and so positively charged helix is present in proteins that are needed to be imported into the

mitochondria. After translocation, this signal peptide is cleaved by mitochondrial process peptidase or MPP. The process of import is mediated by the TIM-TOM complex.



INSULIN

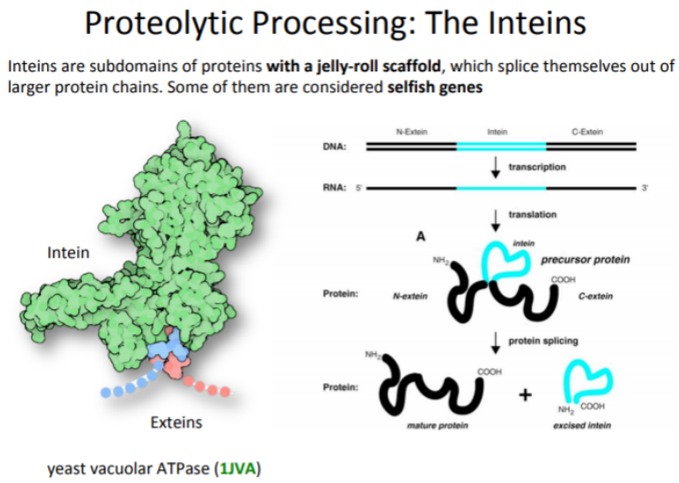
Another example of proteolytic process has insulin as protagonist. Insulin is produced by the beta cells of pancreatic islets. It regulates the metabolism of carbohydrates by promoting the absorption of glucose from blood into fat, liver and skeletal muscles. Glucose can then be accumulated as glycogen or converted into fatty acid. Insulin produced in the pancreas is released when stimuli are detected such as the rise in concentration of glucose in the plasma and other monomeric units like amino acids that results from the digestion of food.



Insulin consists of 2 polypeptide chain a and b chain put together by disulfate bridges. It is first synthesised by a polypeptide called pre proinsulin that contains signal peptide which direct the nascent polypeptide chain in the rough ER. The signal peptide is then cleaved into the lumen of the rough ER forming proinsulin and there the pre insulin fold in the correct conformation and the bridges are formed. Proinsulin is then transferred to the trans Golgi network where new granules are formed. Then it maturates thanks to endopeptidases known as PC1 and PC2 convertases as well as an exopeptidase that is a carboxypeptidase called carboxypeptidase E. The endopeptidase cleaves at two positions producing a fragment called C peptide leaving two chains A and B linked by 2 disulphide bonds. The cleavage side are each located after a period of basic residues, after cleavage of C peptide these 2 pair of basic residues are in turn removed by the carboxypeptidase. Resulting insulin is packed into granules and waits for the release signal.

INTEINS

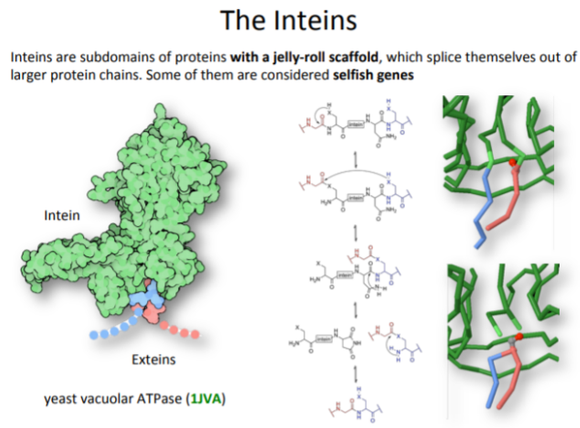
The exact function of intein is told that these proteins are selfish elements of genome. Those are similar of introns of mRNA but in peptides, indeed intein is surrounded by 2 polypeptides called N-Extein and C-Extein. After transcription of the intein mRNA splicing does not occur and the protein is codified as precursor. Then a particular phenomenon called protein splicing allow us to get the mature protein composed of jelly-roll scaffold subdomains.



PROTEN SPLICING MECHANISM

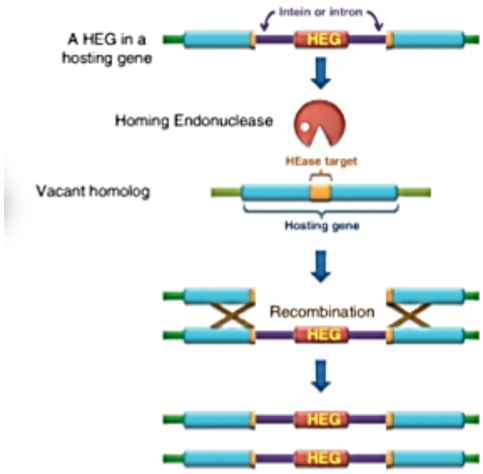
N terminal and c terminal Extein are represented in blue and red. Process begins with a chemical shift in the side chain of the 1 residue that can be ser, thr or cys of the precursor protein makes a nucleophilic attack to the peptide bond of the residue immediately upstream that is the final residue of the n terminal Extein to form an ester or thioester intermediate. Then a transesterification occurs when the side chain of the first residue of the C terminal Extein attack the newly formed thioester to free the N-terminal end of intein

The last residue of intein is usually an asparagine and an amide hydrogen atom of this side chain cleaves the peptide bond between intein and C-Extein resulting in a free intein able to fold into jellyroll with a particular terminal cyclic ring. Finally, the free amino group of the c Extein attacks the ester producing a peptide bond and new functional protein.



Gene codifying for intein are considered selfish genes. Indeed, intein usually include two main domains.
The first splices the intein out of the overall protein chain.
The second is called homing endonuclease HE because its function is to home in DNA that does not code for any intein.
HE Gene usually cleaves homologous gene without intein and in this way the cell normal repair mechanism will try to fix this DNA break by using homologous recombination using the intein including gene as a template. In this way when the damage is corrected the gene is left with another intein and so this selfish element can be propagated within the genome (that is why is called selfish).

INTEINS IN GENOME EDITING

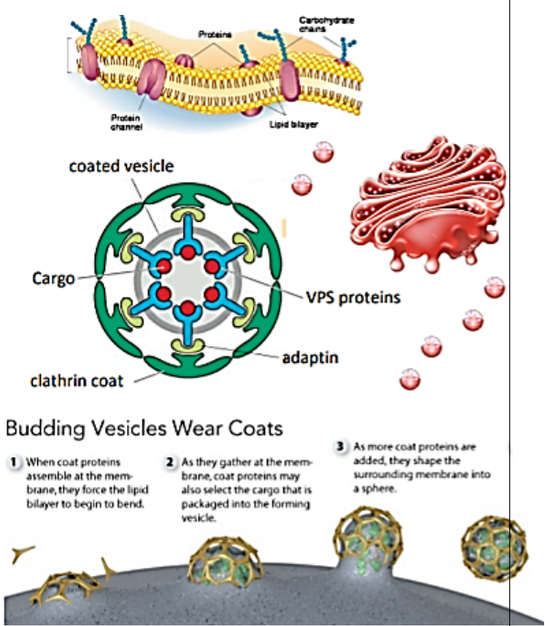


Is very interesting that homing endonuclease are used in gene therapy. All the gene editing technologies currently exploited by pharmaceutical industry are crispr/cas9 or Talen. These perform DNA recognition and cutting functions and differs in specificity, side, biochemistry, repair mechanism etc. actually homing endonucleases are the only monomeric and naturally occurring protein able to bind and cleave DNA.
Video da 23:25 a 25:29

GLYCOSYLATION

Sugar moieties can attach to the hydroxyl (forming o-glycosyl) or amine (forming n-glycosyl) of proteins, lipids and other molecules. Glycosylation is usually accomplished in the endoplasmic reticulum are completed in the Golgi Complex, which sort proteins to their destination (organelles, plasma membrane, secretion).
It has a significant effect on protein folding stability and recognition between proteins in immune system and intermolecular recognition in general. In the case of N-glyc, a precursor oligosaccharide is linked by a pyrophosphoryl residue to dolichol, a long-chain (75 – 95 carbon atoms) lipid molecule. Sorting in the cell of glycosylated proteins is mediated by budding vesicles which are made up of clathrin coat and Vps (Sortilins) proteins. Vesicle are composed by an outer layer of clathrin, another layer of adapting that mediate the attach of the Sortilins.

We can consider as an example targeting of proteins in lysosome which requires a particular glycosylation which is called mannose 6-P-glycosylation.



*Lysosomal enzymes in the slide are the target for this glycosylation.
Red blood cells are glycosylated on their external coat and this glycosylation determines the A, B,0 blood type. Indeed, all humans and many other primates have these blood groups which are of four kind: AB, A0, B0 or 00
There are two antigens and two antibodies mostly responsible for this AB0 types. The associated ant A and anti B antibodies are usually immunoglobulin M (IGM) antibodies and are produced in the first year of a new-born life. A mismatch of this serotypes after blood transfusion can cause a strong immune response.

Blood groups are inherited from both parents and is controlled by AB0 gene with three alleles. ABOO epitopes are conferred by different kind of sugars. N-acetyl galactosamine for A type and galactose for the B type. A specific combination of these 4 components determines an individual type in most cases.

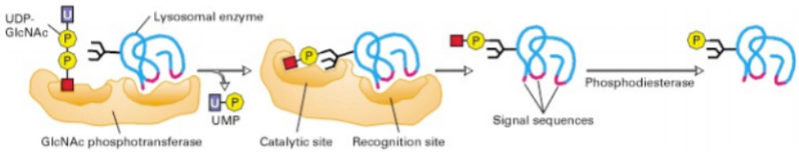
Glycosylation and ABO blood groups

	Group A	Group B	Group AB	Group O
Red blood cell type	A	B	AB	O
Antibodies in Plasma	Anti-B	Anti-A	None	Anti-A and Anti-B
Antigens in Red Blood Cell	A antigen	B antigen	A and B antigens	None

For a blood donor and recipient to be ABO-compatible for a transfusion, the recipient must not have Anti-A or Anti-B antibodies that correspond to the A or B antigens on the surface of the donor's red blood cells

Glycosylation

mannose 6-phosphate targets proteins to lysosomes



N-acetylglucosamine (**GlcNAc**) **phosphotransferase** in the cis-Golgi transfers an N-acetylglucosamine phosphate group to carbon atom 6 of one or more mannose residues.

This enzyme has a **recognition site** that binds to signal segments present only in cathepsin D and other lysosomal enzymes.

In the final reaction, a **phosphodiesterase** removes the GlcNAc group, leaving a **phosphorylated mannose** residue on the lysosomal enzyme.

For example, people with A blood type will have the A antigen on the surfaces of their erythrocytes and as a result, antibodies recognizing the A epitope will not be produced because they would target an endogenous target leading to autoimmune disease and anti-B will be produced. If B type blood is injected in the body with A blood type, antibodies will recognize the antigens as nonself, and this will trigger an immunity response.

A serious disease is the which maternal antibodies pass through the placenta in the fetal circulation and there cause the hemolysis of red blood cell of the new-born. Typically occurs in mother with 0 blood group because they can produce A and B antigen.

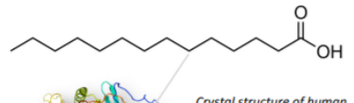
PROTEIN MODIFICATION WITH LIPDS

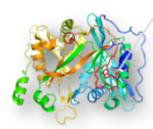
Some Proteins are post-translationally modified with lipids in ER and Golgi, which often anchor them to the inner face of the membranes. Depending on the lipid attached, we can have Myristoylation, Palmitoylation, Farnesylation, geranylgeranylation etc. all together all those processes are called prenylation.

1) For example, Myristoylation is a lipidation modification where a myristoyl group, derived from myristic acid, is covalently attached by an amide bond to the alpha-amino group of an N-terminal glycine residue.
2) Prenylation can regulate the localization and activity of proteins. For example, the protooncogene tyrosine kinase c-SRC plays an important role in embryonic development, apoptosis, signal transduction, cell proliferation, infectivity. An elevated activity of c-SRC is usually suggested to be employed in cancer progression. Myristoylation and membrane binding regulate the c-SRC stability and its kinase activity. Upon modification, a switch of conformation can detach protein from membrane.
3) Many viruses utilize Myristoylation of viral matrix protein to target the viral protein to the membrane for budding and viral maturation.

Protein Modification with Lipids

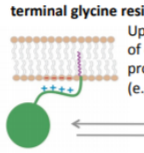
Some Proteins are post-translationally modified with **lipids** in ER and Golgi, which often anchor them to the **inner face** of the membranes. Depending on the lipid, we can have **Myristoylation**, **Palmitoylation**, **Farnesylation**, **geranylgeranylation** etc..



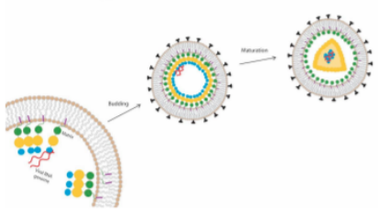


Crystal structure of human type-I N-myristoyl transferase with bound myristoyl-CoA. Myristoyl-CoA.
3IU1

For example, **Myristoylation** is a lipidation modification where a myristoyl group, derived from myristic acid, is covalently attached by an amide bond to the alpha-amino group of an **N-terminal glycine residue**.



Upon modification, a switch of conformation can detach protein from membrane (e.g. **Oncogene c-Src**)



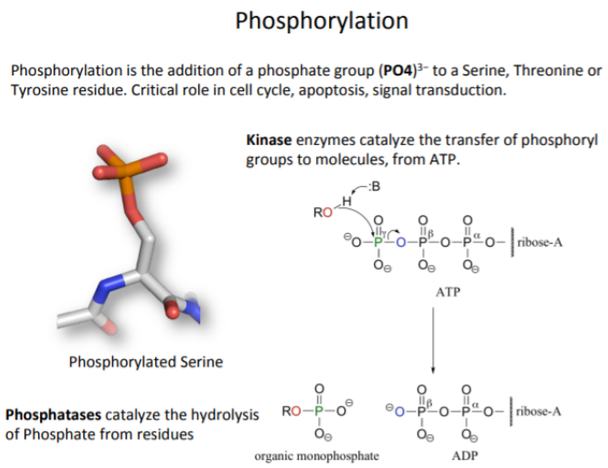
It is an important process in many cellular processes, e.g., **apoptosis**, **signal transduction**, **cell proliferation, infectivity.**

For example, HIV-1 utilizes myristoylation on the **Viral Matrix Protein** to target the viral proteins and viral genome to the membrane, for **budding** and **viral maturation**.

PHOSPHORYLATION

It consists in the addition of phosphate group to a serine threonine or tyrosine residue. Is the most important modification for cell signalling and is one of the few reversible modification and it provides the rational that phosphorylation is a regulatory process.

Kinases are used to transmit signal and regulate almost every complex processes. Kinases mediate the transfer of phosphate moiety from a high energy molecule such as ATP to a substrate molecule.



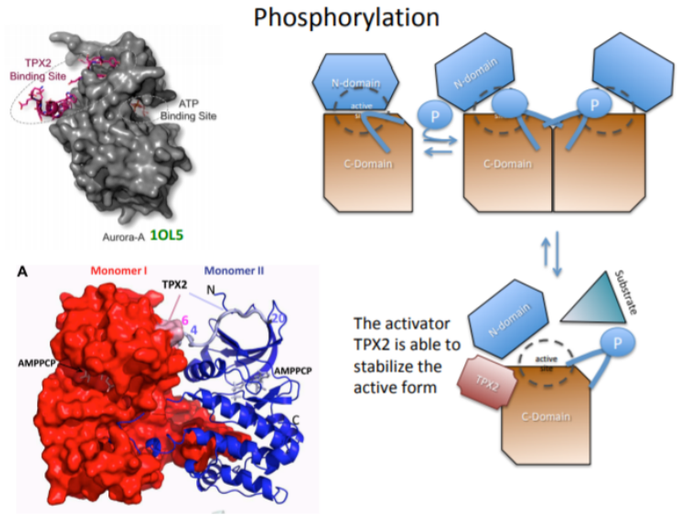
Kinases are indeed needed to stabilize this reaction because the phosphate contains high level of energy, kinases properly orient the substrate and the P group in the active site. In this way the rate if reaction in increase. they common use in the active site positively charged residues stabilizing the transition state interacting with the negatively charged phosphate of ATP. In some cases, other kinases utilize a metal cofactor in their active site to coordinate the P group and eliminate the repulsion between the negative charges.

Opposite reaction of kinases is the hydrolysis of the P group carried out by phosphatases.

Phosphorylation is so important as regulatory mechanism that even kinases re regulated by the so called autophosphorylation.

AURORA protein, which is a kinase involved in mitotic processes is activated by TPX2 but also by a phosphorylation at the level of the activation loop. This loop once phosphorylated undergoes a conformation transition leading to the active form of the kinase in which the active site is open. Actually, this autophosphorylation is a cross reaction between 2 kinases able to dimerize. The activator TPEX2 can stabilize the full active form of aurora.

Insulin to insulin receptor is another example. Insulin can bind to its receptor at surface of cell. Insulin receptor has a dimeric quaternary structure and in the absence of insulin its two intracellular domains are kept separated. Bind of insulin triggers a conformational change in the receptor that brings the inner part of the protein domains closer and able to dimerize. Each subunit in the intracellular domain of the receptor is a tyrosine kinase able to dimerize and phosphorylate each other. This autophosphorylation activates a signalling cascade involving other kinases that in turn activate the glucose importers and other molecules involved in glucose metabolism.



CDK are also involved in cell cycle control and are kinase. The activation of those is phospho-dependent but also cyclin dependent. In both cases teres a very important helix that is called PSTAIRE that interacts with ATP and is repositioned

In the case of phosphorylation there's an arginine switch. Tis evolutionarily conserved arginine interacts with a glutamate inside the PSTAIE helix and when a tyrosine residue of the kinase is phosphorylated this switch of arginine makes the

HYDROXYLATION

Another important modification hydroxylation. Hydroxylation is a chemical process that introduces a hydroxyl group (-OH) into an organic compound. Hydroxylation reactions are often facilitated by enzymes called hydroxylases.

An example is collagen which is the main component of connective tissue and the most abundant protein in mammals. Collagen is very rich in hydroxyproline which contacts water and is added post-transcriptionally by an enzyme called proline hydroxylase and require alpha-ketoglutarate and reduce form of iron and vitamin C.

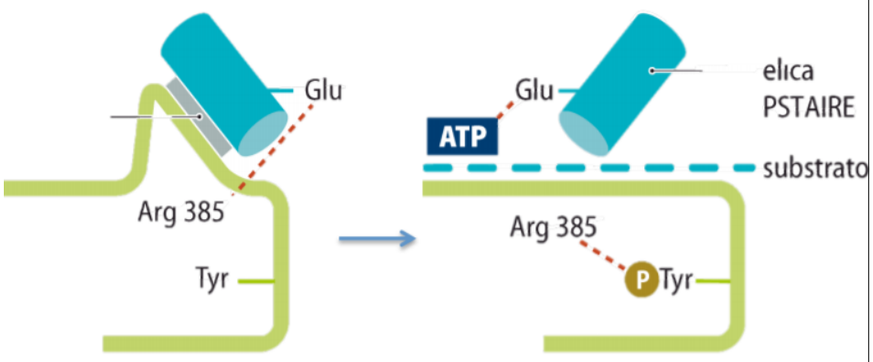
Vitamin c is an important cofactor in many reactions. These reactions involve the oxidation of iron.

Scurvy is a disease resulting from a lack of vitamin C. Early symptoms include weakness, feeling tired, curly hair, and sore arms and legs.

In other kinases of the mitotic cycle called Cyclin-dependent kinases (CDKs) they are involved in cell cycle control the activation of kinase function is both phospho-dependent and Cyclin dependent. In both cases there is a very important helix that Is called PSTAIRE interacting with ATP that is repositioned upon phosphorylation and cyclin binding.:

-In the first case (phosphorylation) there is arginine switch an evolutionary conserved arginine residue which interact with a glutamate inside the PSTAIRE helix, a tyrosine residue of the kinase is phosphorylated this switch of arginine makes the glutamate ready to react with ATP so in this way opening the active site of the substrate

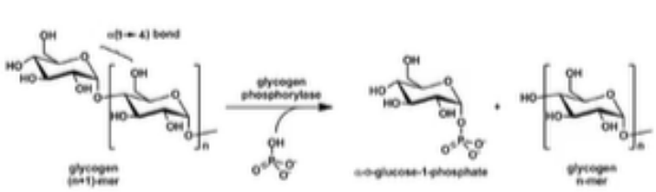
-in Second mechanism CDK also require the presence of cyclin to become active and cyclin have no enzymatic activity on their own but activate CDK in a way very similar to aurora. When cyclin is bound ATP is properly oriented



Cyclin bound
ATP properly oriented via interaction with repositioned T loop and PSTAIRE helix.
Substrate binding cleft suboptimal.
Tyr14 site in roof of ATP binding cleft is available for Wee1 phosphorylation (not shown)

Phosphorylation also regulates the activity of many metabolic enzymes for example Glycogen phosphorylase breaks Glycogen into glucose subunits by releasing Glucose 1 phosphate. Glycogen phosphorylase has got a pyridoxal phosphate (PLP) cofactor and this is covalently bound to lysine 680 forming the so called shift base in this case PLP act differently with respect than other situation in which PLP is used because in this case the lysine-PLP the so called internal aldimine bond is not broken but PLP is holding the active site and the catalytic centre is the phosphate group of PLP that donate a proton to an inorganic phosphate molecule allowing in this way the inorganic phosphate to be deprotonated by the oxygen atom forming the alpha 1 for glycosidic linkage. PLP can be readily deprotonated because the negative charge that is formed is stabilized not only within the phosphate group but also by rezoning in the pyridine ring and in this way the conjugate base resulting from the deprotonation of PLP is very stable

After phosphorylation on Ser14, Glycogen phosphorylase undergoes a deep conformational change that opens the active site entrance and phosphorylase A and Phosphorylase B exist in an active state (T) and a relaxed state (R)



Phosphorylase B is in t state due to presence of glucose 6 phosphate while phosphorylase A is normally in R state, phosphorylation shifts the equilibrium between two states the helix rotates of 45 degree one to the other.

Another kind of post transcriptional modification is methylation and acetylation:

It is the transfer of a methyl group or an acetyl group to a positively charged, usually to the nitrogen of a lysine or arginine. The main enzyme responsible for this reaction is the acetyl transferase which require Coenzyme CoA (vitamin B5) for the reaction to occur.

In this reaction Glutamate of the active site (negatively charged) act as a catalytic base deprotonating the epsilon cazz '?' group of lysine which in turn can carry out nucleophilic attack to the acetyl carbonyl group bound to Coenzyme A in this way the coenzyme A is the living group in this acetylation

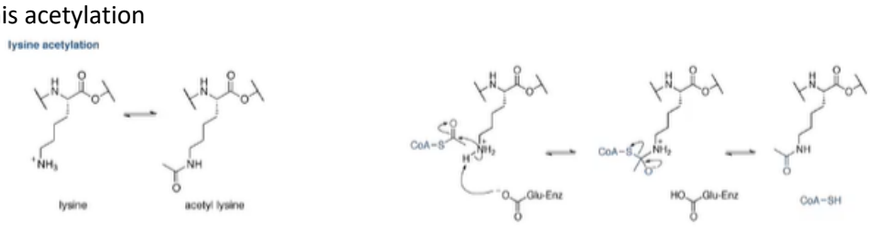
Lysine methylation:

You have different grades of methylation:

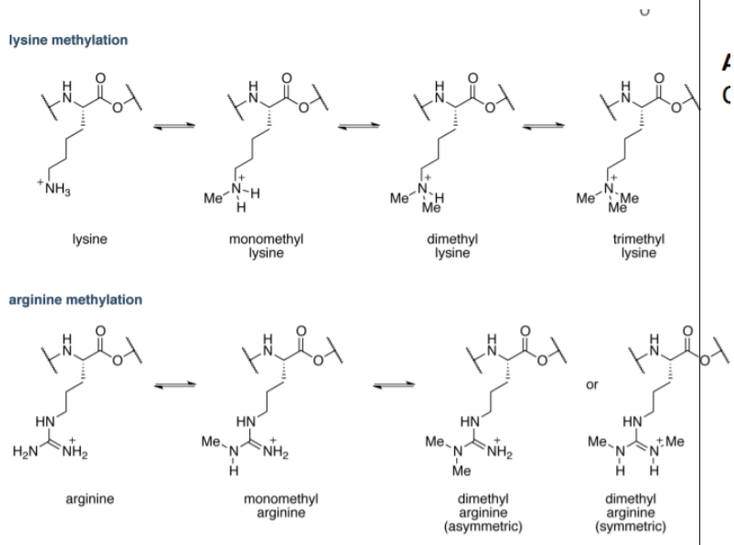
-monomethyl

-dimethyl

-trimethyl



For Arginine you have only monomethyl or dimethyl you can have an asymmetric (both methyl present on the same nitrogen atom) or symmetric methylation (methyl on different nitrogen atoms)



Methylation is carried out by Methyl Transferase and dependent on another cofactor that is S-Adenosyl methionine (SAM). Methyl Transferases can act on nucleotide and that is important for epigenetics process 5-methilcytosine is important for regulation of transcription. In Cancer the DNA is hypermethylated in articular in CpG islands preventing transcription (mask the activity of a promoter). DNA methylation requires a process that is called base flipping: when cytosine swing out and rotate of 180° toward the enzyme outside from DNA groove. This flipping is made possible by the structure itself of met transferases in which some catalytic groups in a loop can replace the fifth base in DNA groove.

Methylation and acetylation together with other kind modification at nucleosomes are known as epigenetic histone modification and can regulate finely the gene expression.

Methylation and Acetylation remove positive charge of histones decreasing the strength of interactions between tail of histones with the negatively charged phosphate group of DNA and the chromatin goes to a more relaxed conformation and allow transcription.

Another post translation modification is Poly-ADP-ribosylation, poly (ADP-ribose) polymerases also called PARPs catalyse the synthesis of ADP-ribose polymers and attach them to specific target proteins.

NAD is needed as a substrate for generating monomers nicotinamide is the living group while the ADP ribose is covalently attached to the protein.

Main role of PARPs, usually find in the nucleus, is to detect and initiate an immediate cellular response to single strand DNA break coming from metabolic, chemical or radiation induced stress.

When DNA damage is identified, enzyme can bind to this region and modify himself through ribosylation and so is able to produce a large and branched chains of poly (ADP-ribose) and those chains are able to recruit other enzyme involved in DNA repair.

PARPs carry out surveillance activity. We have data accumulated on efficacy of PARPs inhibitors in different kind of cancers cause without PARPS the DNA damages is not repaired since is not recognised. In cancer cells Lynparza(drug) take advantage of the fact that normal cells have to way of repairing while the cancer cells have only the PARPs pathway (sta in un video che lui ha fatto vedere e no non si è capito qual è l'altro modo per le normal cell)

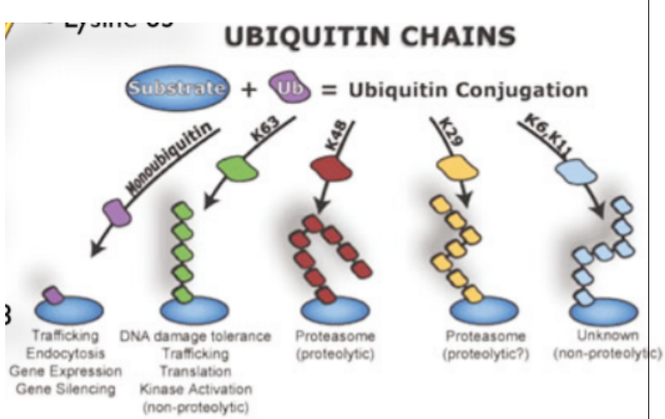
The end of protein Ubiquitination.

Covalent bounding of protein and ubiquitin is the first step for protein degradation in a proteosome-dependent manner. Is catalysed by a set of 3 different enzymes (E1, E2, E3) that activate, conjugate and ligate respectively ubiquitin protein to substrate proteins. The covalent bond is an isopeptide bond between a lysine of the target protein and the C-terminal glycine of ubiquitin.

Ubiquitin is very small (about 76 amino acids) and very highly conserved and widely expressed in all eukaryotic cells.

Ubiquitination involves one or more covalent additions to lysine residues of target proteins and lysine are used for subsequent elongation of ubiquitin chains Two most important lysine residue are lysine 63 and 48 depending on the lysine used we can have different fate for the ubiquitinated protein in fact ubiquitin has a variety of effect in target function and fate. This effect largely depends on type, location and dynamics of modification. (8 kind of chains) the most understood is the proteolytic K48.

Cascade of reactions:

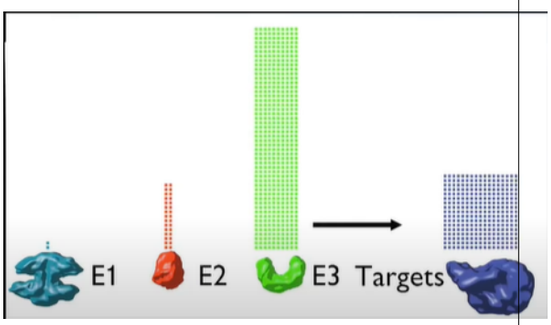


In the first step ubiquitin C terminus is covalently attached to thioester bond to the active site cysteine of E1 or Ubiquitin Activating Enzyme, this reaction is powered by ATP hydrolysis.

E1 than binds ton E2 or Ubiquitin Conjugating Enzyme and catalyse the transfer of ubiquitin onto the E2 active site cysteine this step is a trans thiolation reaction since the final product is also a thioester linked ubiquitin.

In the final step ubiquitin is transferred from E2 to a lysine in the target protein this step is orchestrated by the E3 ligase which brings together ubiquitin loaded E2 and the substrate and catalyse the transfer forming a covalent bond. We have multiple classes of E3 that catalyse this step differently the most common type is the ring E3 ligase that contain a ring domain

binding both E2 and substrate and stimulate direct attach of the substrate lysine on the E2 ubiquitin Thioester and so we have c-terminus substrate lysine bond.



Another class is HECT E3 ligase that contain an active site cysteine, in this case bind E2 and substrate but first stimulate transfer of ubiquitin to E3 cysteine having one trans thiolation reaction more.

We have few E1 enzyme, more e 2 and a lot of E3 that deals with target specificity

Ubiquitination revolutionized the world of pharmacology for the induced degradation this is made possible by molecules called PROTACS, small molecules composed by two active moieties and a linker. PROTAC induce proteolysis and is composed by two covalently linked binding molecules:

-one recruit E3 ligase

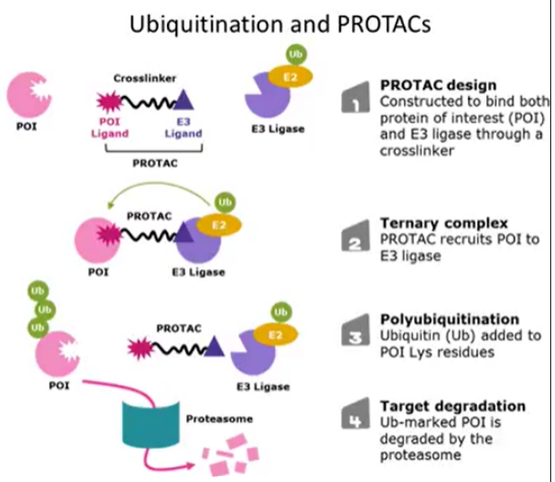
-another one bind target protein

This led to degradation by proteasome hijacking ubiquitin proteosome system (UPS), we have many products example:

(HIF-1 peptide fragment) → FKB is a protein target in treating patient after organ transplant or suffering autoimmune disorders.

(Nutlin) → MDM2 is E3 ubiquitin ligase of p53

Thalidomide → E3 Cerebron a linker and ligand of bromodomain (boh non mi chiedere non ho capito)



The main advantages of PROTAC is modularity, linker can be of different kind and length depending on the needs, we have many E3 and many readymade templates to attach to our protein of interest

Very similar post transcriptional modification is SUMOylation and SUMO stand for Small ubiquitin like modifier, those are a family of small proteins that are covalently attached to and detached from other proteins to modify their function this process is one of the few reversible together with phosphorylation this simulation is involved in a plethora of cellular process the process such as nuclear cytosolic transport, apoptosis, protein stability, response to stress, and progression through the cell cycle. Sumo proteins are usually expressed in their precursor form, cleavage of the residue after the diglycine (GG) region (a particular motif of this precursor) by SUMO specific proteases (SENP) during maturation is needed for activation of SUMO proteins and subsequent SUMOylation. The activated Sumo protein is bound to an E1 called SAE (SUMO activating enzyme) and is a heterodimer, so we have SAE1 and SAE2, Sumo than is passed to an E2 (UBC9) than E3 attach the sumo to his substrate the process is reversible in a way that is known as desumoylation.

The fate of SUMO substrate could be:

-Alter of conformation

-regulation of substrate stability through recruitment of ubiquitin ligases

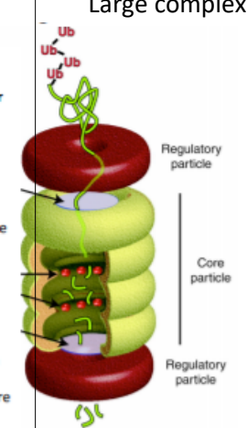
-sumo mediated interaction with other proteins

-block the interaction with other protein

Ubiquitin and SUMO are part of a family that is UBL (Ubiquitin like protein) those are involved in function like autophagy, protein trafficking, inflammation, immune responses, transcription, DNA repair and cellular differentiation.

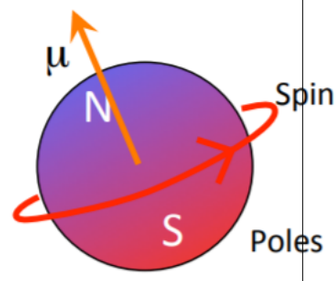
The Proteosome:

Large complex like a sort of basket able to recognize ubiquitin in a very specific way, it contains a "core" of four stacked rings forming a central pore. The inner two rings are made of seven β subunits that contain proteases active sites. These sites are located on the interior surface of the rings. The outer two rings each contain seven α subunits whose function is to maintain a "gate" through which proteins enter the barrel. These α subunits are controlled by binding to "cap" structures that recognize polyubiquitin tags MHC allow the immune system to bind to peptides … tutte cose che sappiamo da immunology in più ha detto solo per anticipare che ci farà un intera lezione sopra (come no)



NUCLEAR MAGNETIC RESONANCE

Pymol from min 1:30 to 8:00



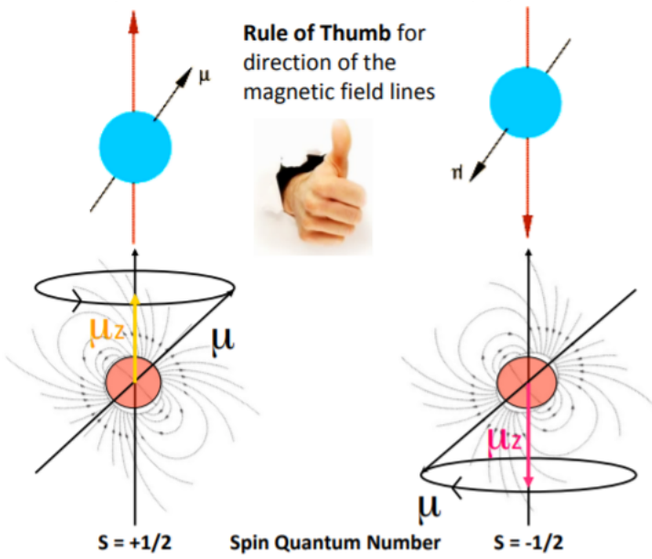
NMR is a spectroscopy technique based on the measurement of the absorption of an electron magnetic wave with a frequency between 4 and 900 MHz (radio wave) by the nuclei of some types of atoms. We measure the ability of nuclei of some particular atoms to absorb and to release waves with a wavelength in the range of radio waves. Is a technique able to determine the dynamical properties of molecules, the movements in time and space. The disadvantage is that is able to solve the structure only of very small macromolecules (max 100 residues).

The molecule is represented as a network of atoms, the data is obtained with the so called contact map (indicates which atoms of the system are in proximity in space) -> from the contact map we derive a 3D network and from the network we build the final 3D model of the protein. One of the most important properties of NMR is the SPIN of atoms (spin: rotation of an object around an axis which is at the center of the rotating object). All rotating objects have a spin. NMR focuses on the atoms' nuclei spin. Since nuclei are charged, when they rotate, they create an electromagnetic field.

NUCLEAR MAGNETIC MOMENTUM is the vector that describes the strength of the electromagnetic field (the longer the vector, the higher the field's intensity). The vector's direction is always perpendicular to the object's plane of rotation and the verse of the vector is given by the thumb's rule (the fingers follow the rotation while the thumb indicates the vector's verse).

NMR can be used only with atoms that have a nuclear magnetic momentum. Such atoms have: an odd number of protons (like H), odd n of neutrons (like C) or both (like N) -> must be present an asymmetry in the geometry of the nuclei to have a nuclear magnetic momentum.

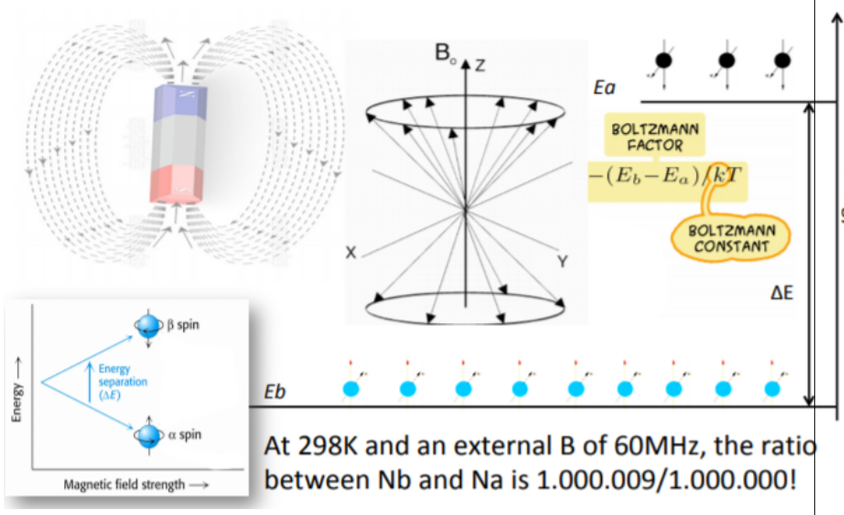
During rotation, also the rotational axis rotates in a process called PRECESSION (= change in the rotation of the rotational axis of a rotating body). The rotational axis is rotating about a second axis, the Z axis (the nucleus is said to be rotating about the Z axis). The rotation is a kind of wave because there is a rotation frequency. The higher is strength of the external magnetic field, the higher is the speed of rotation of the atom. The frequency (time needed to complete the circle) is the quantity describing that speed. For atoms the frequency is called LARMOR FREQUENCY (frequency of precession of atoms). Another property of the nuclear magnetic momentum is the PHASE -> is the starting of the rotation at time 0. Atoms can have same frequency but different phase. We can consider also the projection of the vector on the Z axis:



There can be 2 different kind of projections on the Z axis. Since this is a quantity naturally associated to the atoms' nuclei, this is associated to a quantum number: Spin quantum number (for ex H atoms can take +1/2 or -1/2 values).
In the absence of any external magnetic field, the nuclei of a molecule can orient in any direction with different precession frequencies; they are completely free. In the presence of an external magnetic field, since the nuclei' vectors are very small magnets, they orient themselves in the same direction of the external field, assuming also the same Larmor frequency -> the speed (frequency) at which the nuclei precess is the Larmor frequency, which is proportional to the strength of B:

Frequency ν (Larmor frequency) = γ * B γ = constant

In NMR, in the presence of an external magnetic field, we want that at least some nuclear magnetic momentums present in our population of atoms, point in the opposite direction of the external B momentum. To do that we have to apply energy. There will be 2 energy states: one corresponding to the vectors aligned to B (most favourable and stable energy state) and one corresponding to the vectors that have the opposite direction. Energy is quantified because we are in quantum mechanics -> there's nothing in between the 2 states:



The hourglass represents all the nuclear magnetic momentums of the atoms in our population. They are all rotating with the same speed but different phase. The energy difference between the 2 states is proportional to the strength of the external B -> if B = 0, all the elements in the population have the same energy level. As B increases, also the energy separation between the 2 states increases.

The Boltzmann law states that the ratio of the number of nuclei in the favourable state (Eb) and the number of those in the other state (Ea) is proportional to the energy gap between the 2 states.

To change the ratio, we can decrease the energy gap or apply energy in the form of radio waves -> the only way for atoms, nuclei and electrons to absorb and release energy is through waves. Applying radio waves some atoms absorb the energy and reach the unfavourable state. After a while, they release again the energy and come back to the favourable state. Are applied radio waves because they have the exact amount of energy needed by the system to absorb/release energy (that's the meaning of resonance in NMR) -> RESONANCE = transfer of energy at the maximum efficacy.

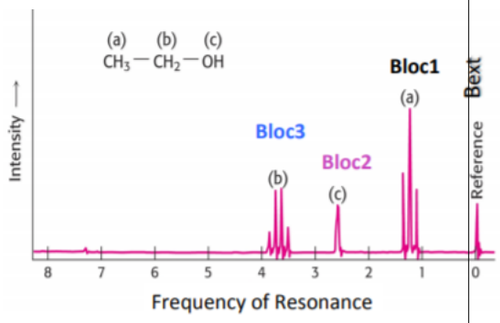
The radio waves frequency exactly corresponds to the Larmor frequency => the basic idea of NMR is to provide to the system radio waves, having a frequency equal to the Larmor freq, because in this way the system enters into resonance and can absorb energy. Each wavelength has an associated energy:

$$\Delta E = h\nu_l$$

h= Planck constant; v = frequency

If we use a radio wave with a freq corresponding to the Larmor freq, we are giving an energy quantity that corresponds exactly to the energy gap between the 2 energy states and the system is in resonance. (like the swing and a guitar's cord) In NMR we register the absorbed and reemitted radio wave.

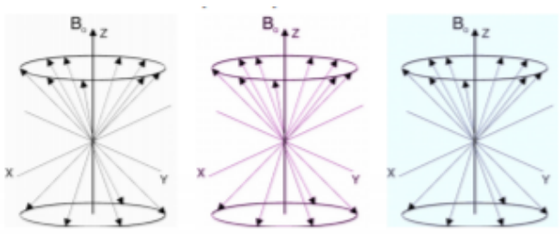
NMR is also used to obtain the fingerprint of a molecule. We put the sample in a tube and then we put the tube in a very huge magnet to create a very strong magnetic field in a way that the stronger is the magnetic field, the deeper is the difference between the energy states and in this way, we obtain a stronger signal of absorption or release of energy.
We take ethanol (CH3-CH2-OH) as our sample and we apply the external magnetic field (Bext). We obtain this graph:



We consider NMR only on H atoms as an example. We have 3 different H populations in the molecule (a, b and c) and we obtain 3 different signals in the plot. Since is present the external B, the H atoms in the molecule start to behave as magnets (because start precessing), creating local magnetic fields => the real magnetic field, sensed by each atom, is not the external one but is the external one + (or -) the environmental one created by the other atoms: $Bloc = Bext - Benv$

There is a sort of interference of the Bext caused by the other local B created by the other atoms. Larmor freq is proportional to the local B. Each member of the H populations is sensing a different magnetic field => we have 3 different H populations with 3 different Larmor frequencies and 3 different local B. Since they have 3 Larmor freq, they have also 3 different resonance frequencies => the radio waves that we must apply to the system to obtain resonance, must be different for the 3 kind of atoms. By comparing the plot to others present in the web, we can say that the plot corresponds to ethanol. This process can be done only in analytical chemistry.

The 3 different atoms populations correspond to 3 different hourglasses:



The scale of the shift in the resonance frequency is called CHEMICAL SHIFT (in ppm: parts per million). The chemical shift is the difference in the absorption compared to the external B. We measure the shift from the reference value (in this case the reference is Bext and corresponds to 0) => in this case we measure the shift from 0

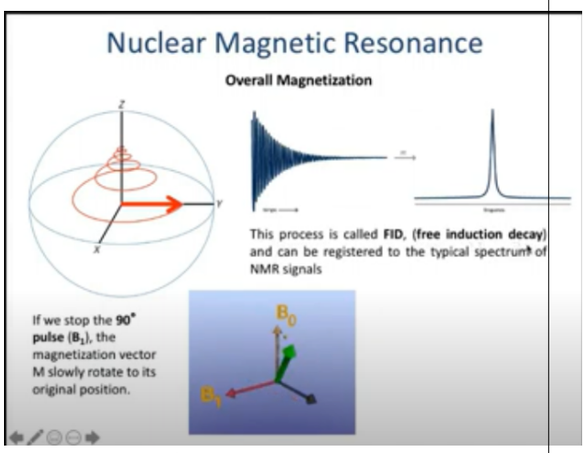
The summitry of the local magnetic fields can be called the environmental magnetic field, so each nucleus is sensing not only the external magnetic field but actually the external magnetic field shielded (minus) the environmental magnetic field and that's the real magnetic field that is sensed by each nucleus. Since the frequency of resonance of each nucleus is related to the intensity of the magnetic field and since each nucleus is sensing each population of nuclei is sensing different magnetic field we obtain 3 slightly different resonance peaks indicating that there is, compared to the external magnetic field, different shielding of the environmental magnetic field, moreover we can associate each of the peaks of this plot to a particular moiety of this molecule and in this way we can pinpoint the resonance peaks of each proton of this moiety, in this case we have 3 different hourglasses (populations of nuclei in this molecule)(each arrow 1 molecule).

Overall Magnetization (M0)

opposite vectors are called counter phase and their overall magnetic field is 0, by using the same criteria with the other vectors in the hourglass(lower energy state), since the number of vectors on the upper side of the hourglass is greater the resulting vector is pointing towards the positive side of the z axis. Overall magnetization (M0) is the sum of all the nuclear magnetic momentum of a population of similar nuclei (nuclei with the same Larmor frequency=same resonance)

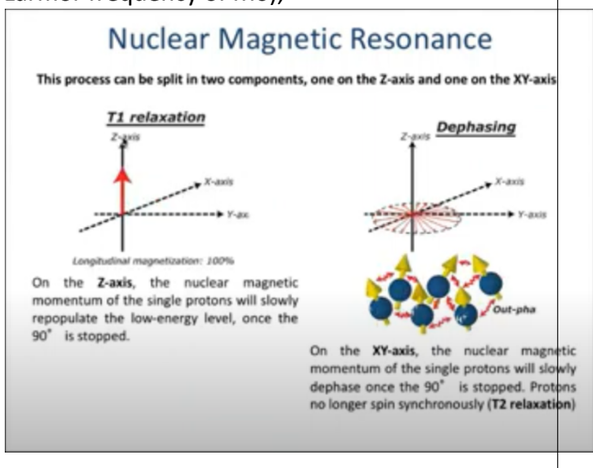
Pulsed Nuclear Magnetic Resonance

The direction of M0 is the same of the external magnetic field, by applying a 90° circularly polarized radio wave (a particular radio wave that is circularly polarized), when the nuclei of these atoms enter in resonance with this radio wave the results in the redistribution of the vectors on the side of the hourglass on a way that makes the vector shift on the X/Y axes, this process s called torque(the reason for the vector being on both axis is because by spinning it occupies both axis), when you stop the pulse the vector will slowly rotate back to it's original direction by drawing a sort of spiral. This process s called FID (free induction decay)



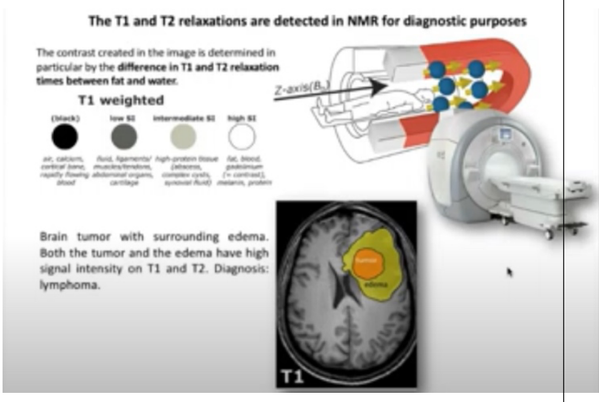
Another kind of pulse is the 180° circularly polarized radio wave and the overall effect results in the inversion of M0. Using the Fourier transform we can switch from the time domain to the frequency, by doing so we can register the frequency of this wave(= to the Larmor frequency of M0),

The FID can be observed in two phases, the one on the z axis and one in the X/Y plane, if we look at them from a geometrical POV, on the Z-axis, the vector, starting from 0 you will see it increase, while on the X/Y plane we observe a vector that while is rotating it's slowly disappearing due to the dephasing of the nuclei; the first one (z-axis) is called **T1 relaxation** , and the second one (X/Y plane) is called **dephasing** or **T2 relaxation**
They are the same process only from 2 different point of view in reality is a spiral.



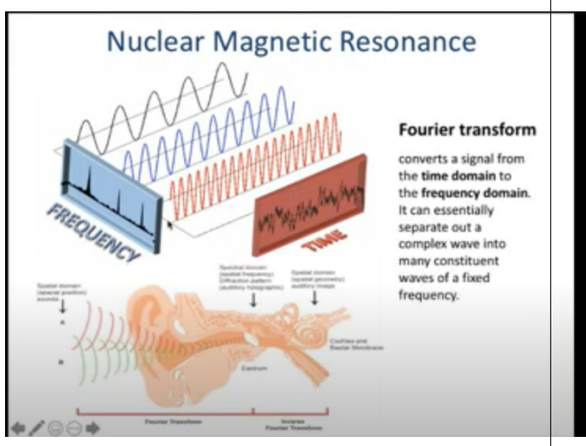
In MRI (magnetic resonance imaging) we measure the T1 and T2 relaxation of the protons of your body, when you enter the machine(that works as a magnet) and your body is oriented on the z axis (same direction of M0), then a full spectrum of 90° circularized radio waves are applied on your body (no radiation= safer) and then we register the FID on your body, the image derived because the T1 and T2 relaxation of the protons of your body is different for the kind of tissue origin of your body

The speed on T1 relaxation is higher for the protons in fat tissues compared to the increase of protons coming form the wet tissues, by doing so we are able to scan the body in 2D and obtain (through software's) pictures of "slices" of the body, in picture (since brain mostly fat tissue) the fact that we have wet tissue we are able to identify tumor and the inflammation (edema) around it.



The same principles of Nuclear Magnetic Resonance to obtain a finger print of the molecule can be applied to identify problems(tumors, inflammation , etc.) in your body.

Coming back to the FID of M0 after the applying of the circularly polarized wave, if we have more than a population of atoms (almost always) we get a more noisy signal composed of many FIDs superposed all together just as the case of the ethanol, this complex signal is simply composed by a number of waves.
An example of that can be sound(music), still measured in intensity over time, still the sum of all the waves together, we can analyze the song also in a frequency domain, the difference being in the first one we are analyzing the evolving of different waves in time while in the second we are analyzing the domain of frequencies, the trick to switch from one to another is Fourier transformation (switch from time domain to frequency spectrum, spectral domain=NMR plot)



NMR for Proteins

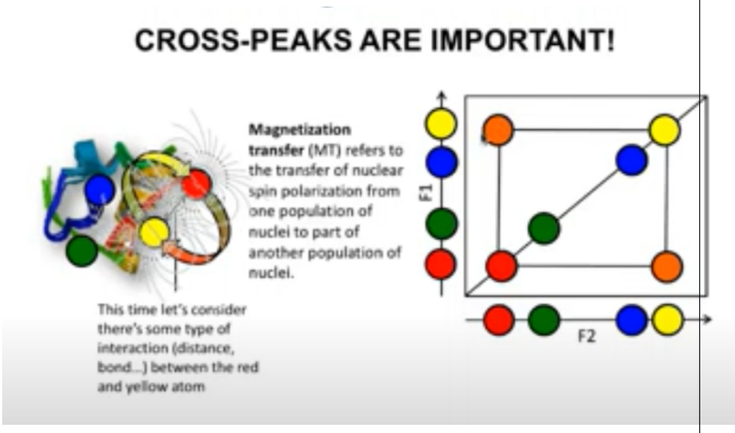
The problem of doing a NMR spectrum even of a small protein we get a very complicated plot due to many different populations of nuclei (each peak= Lamour frequency of a population of nuclei) so it's impossible to assign each resonance frequency to a particular kind of proton.

A way to try to solve the problem of superposed peaks is to use a technique called 2 dimensional NMR, that is like making the experiment twice, the first time with one circularly polarized radio wave(90°) and a second time with a different circularly polarized radio wave(180°) so in 1D we register the spectrum of resonance for the first frequency and on the other dimension for the other one. The final result by combining them is a 2D graph and the signal that we register is our peak, if the Larmor frequency is the same for both experiments we expect fo find the signal along the main diagonal. On NMR 1 we said that we have to pay attention to two different things, one is the protein is considered as a network of interactions, and the second one is that we have to consider the importance of crosspeaks (signal that we register that are nota long the main diagonal) meaning that the signal is changed from the first to the second experiment.

Cross peaks

The meaning of Cross peaks in a plot is related to the fact that 2 nuclei are interacting together, meaning that or connectivity (covalently bound) or distance (interacting because close in space less then 5A),
when 2 atoms are close in space when the first pulse is applied the two atoms start behaving like magnets and this behavior is sensed by the other atom and so during the second pulse the second atom s changing its Larmor frequency and that's what cross peaks are.
From a technical POV we perform the 1D experiment the first process is called preparation that means to apply the external magnetic field(so you have M0), then you have the application of the radio wave called Pulse and then the detection phase in which you register the FID and then we use the Fourier transform to switch from time to frequency domain(only one variable used=Time of decay then switched to frequency), in 2D experiment we have 1D+1D, so we still have preparation (to have M0), then the first radio wave(90°pulse) then we have a lag time called Evolution phase between the first radio wave and the application of the second radio wave, and finally the detection, 2 variables(1 called Time2 that is the same as the 1D experiment and the second is the evolution time called time1(lag time that changes each time from one pulse to another)also called indirect variable because each time we change the time between the first and the second pulse , the final result is the variation of the timing between the two pulses that we apply, if you put such FID all together you obtain a 2D wave and from that if we apply the 2D Fourier transform you obtain the signal, just like in Xray we had the 3D wave and from that we were able to obtain the single scattered Xray and their frequency to obtain the electron density while in this case from this one you obtain the signal.

Diagonal peaks correspond to the peaks in 1D experiment that in this case we don't pay attention to, what's important are cross peaks because they indicate interaction(coupling) between the nuclei.



Cross peaks happen because of magnetization transfer(MT) that is a transfer of nuclear spin polarization from one population of nuclei to part of another population of nuclei and since this process s true for both populations of nuclei we will have 2 cross peaks in opposite sides of the graph.

The reason why the Evolution Time is so important is because you need to catch the exact moment when there is resonance, in the example with pool is focusing on the fact that if the two balls are not interacting they will fall in their F1 and F2 holes no matter the delay between the shots, but if he two balls where interacting and so if the white was to go in the F1 hole and the red one in F2 they would cross paths so if you shoot at the right moment they will switch frequency and the same applies to magnetization transfer.
In a particular kind of experiment called NOESY(Nuclear Over Hauser Effect Spectroscopy NMR) when you observe a cross peaks it means that the atoms are less than 5 A apart and so you can build a contact map and from it you can derive a network of interaction between the atoms and from it you can derived the structure of your model.
Another experiment is called is called COSY(Correlation Spectroscopy NMR) that gives us a fingerprint of each base of RNA/DNA and information's about the connectivity of the bases in the spectrum and also about amino acids, NOESY and COSY work together to get the exact position and composition of the atoms in the sequence.