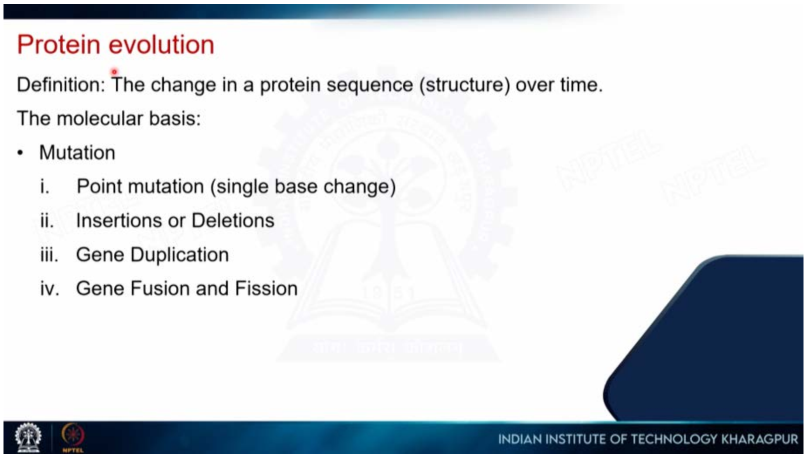
# Introduction to Complex Biological Systems

**Professor Dibyendu Samanta and Professor Soumya De**
**Department of Bioscience and Biotechnology**
**Indian Institute of Technology, Kharagpur**

## Lecture 38
## Mechanisms of evolution

Welcome to Lecture 38 of Introduction to Complex Biological Systems. Today, I am going to talk about the mechanisms of evolution. So this week, I am discussing evolution, the history of life. So, I started discussing protein evolution.

Today, I am going to talk about the major molecular basis of protein evolution. So, what is this? The change in a protein sequence or structure over time is protein evolution and the molecular basis of this protein evolution is mutation. Now, remember that this mutation happens at the DNA level, and the change is reflected at the amino acid level.



Protein evolution

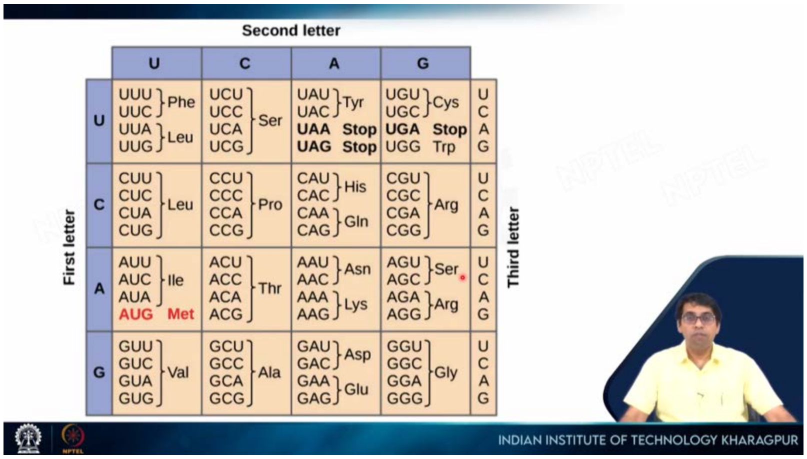
Definition: The change in a protein sequence (structure) over time.

The molecular basis:

- Mutation
  i. Point mutation (single base change)
  ii. Insertions or Deletions
  iii. Gene Duplication
  iv. Gene Fusion and Fission

So, if the change in DNA results in a change in amino acid, then you have a mutation in the protein. So, this mutation can be a point mutation where only a single nucleotide is mutated to something else. It can be an insertion or deletion. Meaning a big chunk of sequence is inserted into a DNA sequence. So, a small chunk is inserted or a small chunk is deleted.

Gene duplication occurs when the same gene is copied again, resulting in two copies of the same gene. Then, one of the copies may mutate because it is not under the correct promoter, and it becomes a different protein that may be with a related function or maybe a completely different one and then there is gene fusion and fission. Two genes can combine to form a chimeric protein with a completely new function, or a single gene can split into two different genes, resulting in different functions. These are the basic mechanisms by which protein evolution occurs. Since evolution happens or these mutations occur at the genetic level, we must examine the genetic table or the genetic code.



|  | Second letter |  |  |  |  |
|---|---|---|---|---|---|
| First letter | U | C | A | G | Third letter |
| U | UUU, UUC Phe<br>UUA, UUG Leu | UCU, UCC, UCA, UCG Ser | UAU, UAC Tyr<br>**UAA Stop**<br>**UAG Stop** | UGU, UGC Cys<br>**UGA Stop**<br>UGG Trp | U<br>C<br>A<br>G |
| C | CUU, CUC, CUA, CUG Leu | CCU, CCC, CCA, CCG Pro | CAU, CAC His<br>CAA, CAG Gln | CGU, CGC, CGA, CGG Arg | U<br>C<br>A<br>G |
| A | AUU, AUC, AUA Ile<br>AUG Met | ACU, ACC, ACA, ACG Thr | AAU, AAC Asn<br>AAA, AAG Lys | AGU, AGC Ser<br>AGA, AGG Arg | U<br>C<br>A<br>G |
| G | GUU, GUC, GUA, GUG Val | GCU, GCC, GCA, GCG Ala | GAU, GAC Asp<br>GAA, GAG Glu | GGU, GGC, GGA, GGG Gly | U<br>C<br>A<br>G |

INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

Here, each amino acids three-letter code is given. It is important to pay attention to this genetic code. Methionine has only one codon. Along with methionine, tryptophan is the only other amino acid that has only one codon. Thus, methionine and tryptophan are the only two amino acids with just one codon dedicated to them.

This means any single mutation in methionine or tryptophan will result in an amino acid change. For example, AUG codes for methionine. If G changes to A, it will code for isoleucine. If G changes to U, it will code for valine and so forth.

If this U changes to C, then it will code for threonine. If this A changes to C, then it will code for leucine. So, a change in any one of these three nucleotides will result in a mutation because methionine has only one codon dedicated to it. On the other hand, if you see valine, valine has four codons dedicated to it, which means that any change in the third base, the wobble base, will not result in a mutation.

So, if this U changes to C, A, or G and vice versa, it will still remain valine. So it turns out that there are amino acids, for example, isoleucine has three, many have two, and some have up to six. For example, leucine has six codons dedicated to it, and arginine has four here and two here. They have six codons dedicated to them.

Another important aspect of this genetic codon table is that similar types of amino acids tend to be nearby. For example, if you look at aspartic acid and glutamic acid. So, if aspartic acid is encoded by GAC, and the C changes to U, it will remain aspartic acid so no change in amino acid. However, if the C changes to A, then it will become glutamic acid.

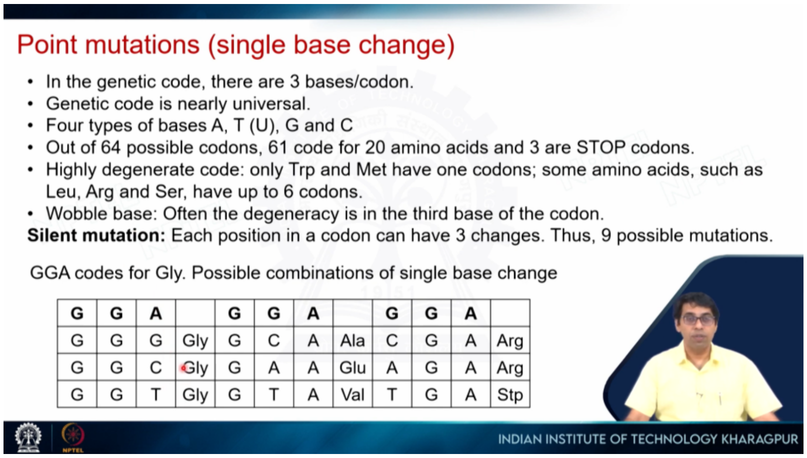
Now, both have acidic side chains. So, functionally, it might not be a very bad mutation. Similarly, if it changes to G, then again it becomes glutamic acid. So, a negatively charged residue remains a negatively charged residue. So, these are similar residues and this genetic code or the plan of the design of the genetic code determines what type of mutations you will see more often and what type of mutations you will see less often. For example, if you want to mutate a glycine to let us say, proline, then you will have to change at least two nucleotides. If you want to change glycine to serine, then you will have to most probably change two nucleotides.

So mutating two nucleotides, the chances, two consecutive nucleotides, the chances are much less compared to a point mutation. So here is a summary of what I just discussed. In the genetic code, there are three bases per codon. So each amino acid is encoded by a codon and in one codon, there are three bases. Genetic code is nearly universal.

It is not absolutely universal, but it is nearly universal among all organisms. There are four types of bases, A, T, G, and C. If you consider RNA, then it will be A, U, G, and C. Out of the 64 possible codons, 61 codes for 20 amino acids, and 3 are stop codons. Highly degenerate codon, the only tryptophan and the methionine have one codon. Some amino acids, such as leucine, arginine, and serine, have up to six codons.

Isoleucine has three codons, and there are other amino acids that have two or four codons so it is a highly degenerate codon. The wobble base, which is the third base of a codon, is often degenerate. For example, we saw the examples of serine and glycine. If you change the third base, it will not result in any change of the amino acid.

Silent mutations occur when a codon is changed into another codon but does not result in a change in the amino acid then that will be a silent mutation. Now, if we consider point mutations, there are nine possible point mutations. So, let us take the example of glycine. Let us say glycine is encoded by a particular protein. The codon we are looking for here is coded by GGA. If you consider GGA and if I mutate the third position, I can change it to G, C, or T. Now GGA codes for glycine, but if I change this third base or the wobble base to G, C or T, it will still remain glycine. So these are all silent mutations. Instead of this, if I mutate the second base so this G becomes C, A, or T then GCA will code for alanine, GAA will code for glutamic acid, and GTA will code for valine. If I mutate the first one, then this will code for arginine, arginine, and a stop codon. So these are the nine possible mutations for this particular codon, which is GGA.



## Point mutations (single base change)

- In the genetic code, there are 3 bases/codon.
- Genetic code is nearly universal.
- Four types of bases A, T (U), G and C
- Out of 64 possible codons, 61 code for 20 amino acids and 3 are STOP codons.
- Highly degenerate code: only Trp and Met have one codons; some amino acids, such as Leu, Arg and Ser, have up to 6 codons.
- Wobble base: Often the degeneracy is in the third base of the codon.

**Silent mutation:** Each position in a codon can have 3 changes. Thus, 9 possible mutations.

GGA codes for Gly. Possible combinations of single base change

| G | G | A |  | G | G | A |  | G | G | A |  |
|---|---|---|---|---|---|---|---|---|---|---|---|
| G | G | G | Gly | G | C | A | Ala | C | G | A | Arg |
| G | G | C | Gly | G | A | A | Glu | A | G | A | Arg |
| G | G | T | Gly | G | T | A | Val | T | G | A | Stp |

INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

It turns out that three of them are silent mutations. Glycine to alanine might not be a very drastic mutation. So the protein might be able to tolerate that because alanine is the next smallest amino acid after glycine. Glutamic acid might be a little problematic because we are introducing a negative charge. So if the glycine is buried, then this will result in denaturing of the protein.

Glycine to valine, again, if the glycine is buried then we are adding these extra carbons, and that can create problems. Glycine to arginine, the smallest residue to one of the largest amino acids, again, if it is buried, then this will create a problem and glycine to a stop codon. So this will result in a truncated protein.

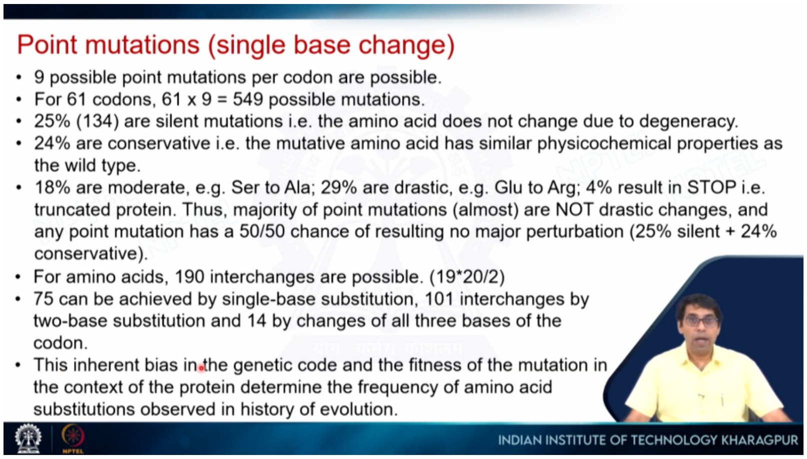
So, for example, let us say if my protein is 150 amino acids long, and if this glycine is at position 30, then this will be a stop codon. So only the first 29 amino acids will be encoded; the remaining will not be synthesized. So it will be a small peptide of only 29 amino acids. So it will be a non-functional peptide.

So this is a drastic mutation. This will definitely cause problems. Depending on where the glycine is, if it is in a loop, then any of these mutations might be tolerated. If it is something that is buried, then most of these mutations might not be tolerated so all of these factors will determine whether this particular mutation will be acceptable or detrimental to the structure and function of a protein. So, point mutations, 9 possible point mutations per codon are possible. So since there are 61 codons, 61 times 9, 549 possible mutations are possible. So, 549 point mutations are possible. 25% of these, that is 134, are silent mutations, which mean that a change in the nucleotide will not change the amino acid.

That is because of the degeneracy of the codons. 24% are conservative mutations which mean that the amino acid that it is mutated to will have very similar physicochemical properties for example arginine gets mutated to a lysine so both are basic or aspartic acid gets mutated to glutamic acid or phenylalanine gets mutated to tyrosine or tryptophan. So 24% are conservative so you will see that almost 50% of the mutations are most probably not going to create much problem, 18% are moderate. So moderate means, for example, serine to alanine.

So this serine side chain OH is polar, it can form hydrogen bonds, which will be absent in case of alanine. But again, both are small side chains. So this should not be a very big problem. 29% mutations will be drastic so these 29% mutations can create big problems, for example, glutamic acid to arginine where you are completely changing the charge on the side chain. 4% will result in stop codon as we saw in case of the glycine. So this will be a very bad mutation because if there is a stop codon, then the protein will become a truncated protein. So it turns out that the majority of these mutations are not going to be a problem.

So there is a 50-50 chance of resulting in no major perturbation because 25 plus 24 is 49% so almost 50% of the proteins will not create major problems. Now, if we think about amino acids, there are 190 interchanges that are possible. So, you have 20 amino acids, each amino acid can be mutated to 19 other amino acids and alanine to leucine or leucine to alanine, we can call this the same interchange. So, you divide it by 2. So, that gives us 190 interchanges. 75 of these interchanges can be achieved by a single substitution. So, these 75 interchanges will be seen more often because a single base substitution is more likely to happen. 101 changes require two base substitutions. So, this will be less likely, and 14 changes require mutating all three bases of the codon. So, this is the least likely interchange that you will see. So, this inherent bias in the genetic code and the fitness of mutation in the context of the protein determine the frequency of amino acid substitutions observed in the history of evolution. So, this is something and this frequency is something that I have discussed in the previous lecture.

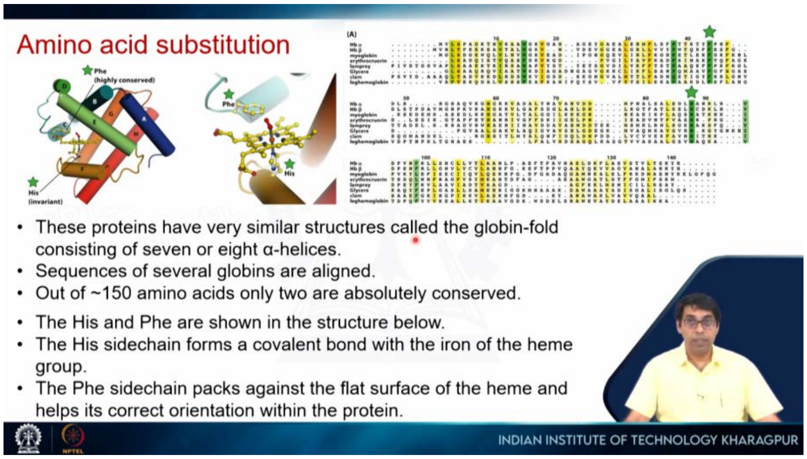


Point mutations (single base change)

- 9 possible point mutations per codon are possible.
- For 61 codons, 61 x 9 = 549 possible mutations.
- 25% (134) are silent mutations i.e. the amino acid does not change due to degeneracy.
- 24% are conservative i.e. the mutative amino acid has similar physicochemical properties as the wild type.
- 18% are moderate, e.g. Ser to Ala; 29% are drastic, e.g. Glu to Arg; 4% result in STOP i.e. truncated protein. Thus, majority of point mutations (almost) are NOT drastic changes, and any point mutation has a 50/50 chance of resulting no major perturbation (25% silent + 24% conservative).
- For amino acids, 190 interchanges are possible. (19*20/2)
- 75 can be achieved by single-base substitution, 101 interchanges by two-base substitution and 14 by changes of all three bases of the codon.
- This inherent bias in the genetic code and the fitness of the mutation in the context of the protein determine the frequency of amino acid substitutions observed in history of evolution.

So we have already seen in the previous lecture that there are certain amino acids, depending on the protein. Some amino acids will be absolutely conserved. Some amino acids which are in this loop can be changed to anything, whereas some amino acids in the core will be highly conserved. So again, for globin, which has 150 amino acids, only two are absolutely conserved. These are histidine and phenylalanine.

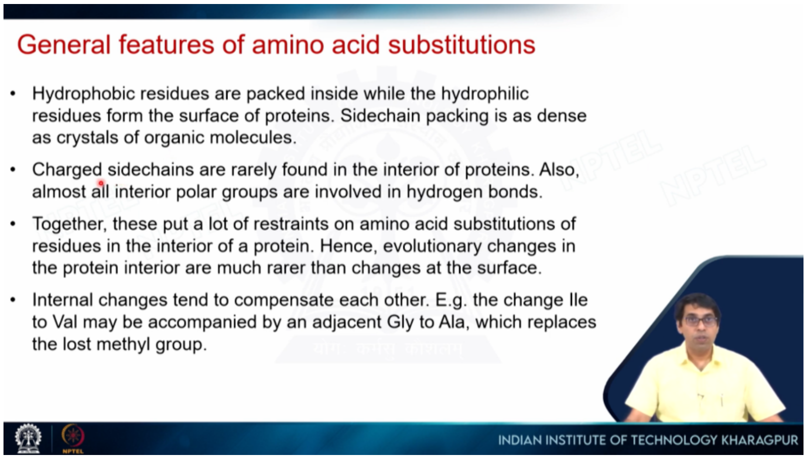
So what are the general features of amino acid substitution? Hydrophobic residues, which are typically inside the core of the protein, and hydrophilic ones are on the surface of the protein. So side chain packing is as dense as crystals of organic molecules. This packing of the side chains in the core of the protein is very dense.



## Amino acid substitution

- These proteins have very similar structures called the globin-fold consisting of seven or eight α-helices.
- Sequences of several globins are aligned.
- Out of ~150 amino acids only two are absolutely conserved.
- The His and Phe are shown in the structure below.
- The His sidechain forms a covalent bond with the iron of the heme group.
- The Phe sidechain packs against the flat surface of the heme and helps its correct orientation within the protein.

So there is no gap in the core of the protein, which means that mutating those amino acids in the core will be difficult because whatever mutation you make, you have to compensate for that change with some other mutation. Whereas charged side chains are on the surface. So charged side chains are rarely found in the interior of the protein. Also, almost all interior polar groups are involved in hydrogen bonds.



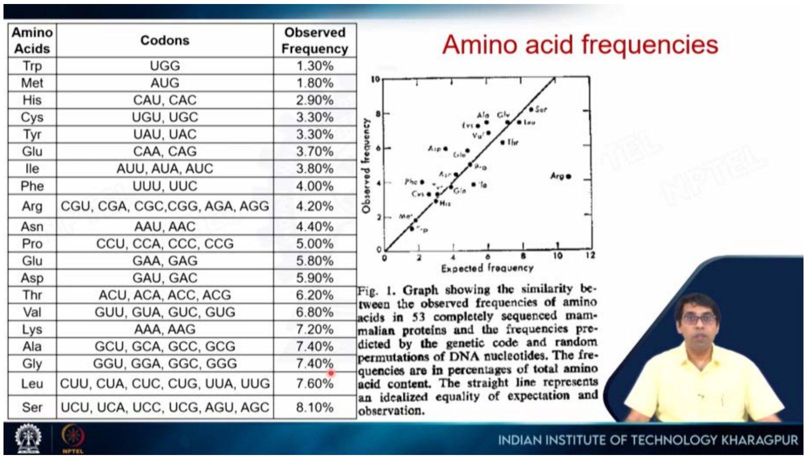
## General features of amino acid substitutions

- Hydrophobic residues are packed inside while the hydrophilic residues form the surface of proteins. Sidechain packing is as dense as crystals of organic molecules.
- Charged sidechains are rarely found in the interior of proteins. Also, almost all interior polar groups are involved in hydrogen bonds.
- Together, these put a lot of restraints on amino acid substitutions of residues in the interior of a protein. Hence, evolutionary changes in the protein interior are much rarer than changes at the surface.
- Internal changes tend to compensate each other. E.g. the change Ile to Val may be accompanied by an adjacent Gly to Ala, which replaces the lost methyl group.

So these are some of the restrictions that we have for a protein structure. So together, this puts a lot of restraints on amino acid substitution of residues in the interior of a protein. So proteins which are in the interior, mutating them will be harder than proteins which are on the surface. For example, if I have a protein like this, if there is an arginine or lysine that is sticking out from the surface, let us say this is arginine or lysine, I can mutate it to any other amino acid, for example, serine or even aspartic acid, if it is not involved in any salt bridge or if it is not involved in the interaction with any other molecule. So, any charged group will be happy on the surface. So, surface residues will mutate more frequently compared to the core residues that are present in the interior of a protein so evolutionary changes in the protein interior are much rarer than changes at the surface.

Internal changes tend to compensate for each other. So if you mutate one residue and you lose a methyl group, then you will have to mutate another residue where you have to add a methyl group. Then only the packing will be maintained. So that is what is shown here: isoleucine to valine will result in the loss of a methyl group. This has to be accompanied by an adjacent glycine mutation to alanine.

So you add a methyl group to glycine to make it alanine so that the internal packing remains the same. So based on this, what people have observed is these are the frequencies of amino acids. These frequencies of amino acids mean that if you consider all different protein sequences and see how often alanine shows up, how often glycine shows up, and so on. So you will see that tryptophan and methionine, which have only one codon, have very low frequencies.



Amino acid frequencies

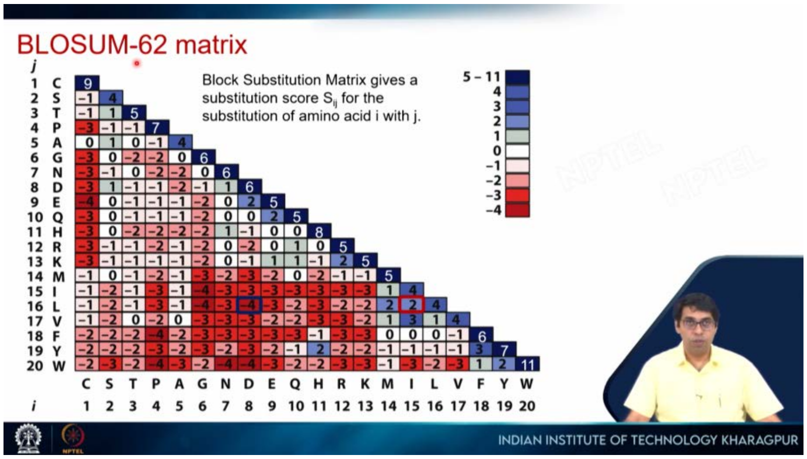
| Amino Acids | Codons | Observed Frequency |
|---|---|---|
| Trp | UGG | 1.30% |
| Met | AUG | 1.80% |
| His | CAU, CAC | 2.90% |
| Cys | UGU, UGC | 3.30% |
| Tyr | UAU, UAC | 3.30% |
| Glu | CAA, CAG | 3.70% |
| Ile | AUU, AUA, AUC | 3.80% |
| Phe | UUU, UUC | 4.00% |
| Arg | CGU, CGA, CGC,CGG, AGA, AGG | 4.20% |
| Asn | AAU, AAC | 4.40% |
| Pro | CCU, CCA, CCC, CCG | 5.00% |
| Glu | GAA, GAG | 5.80% |
| Asp | GAU, GAC | 5.90% |
| Thr | ACU, ACA, ACC, ACG | 6.20% |
| Val | GUU, GUA, GUC, GUG | 6.80% |
| Lys | AAA, AAG | 7.20% |
| Ala | GCU, GCA, GCC, GCG | 7.40% |
| Gly | GGU, GGA, GGC, GGG | 7.40% |
| Leu | CUU, CUA, CUC, CUG, UUA, UUG | 7.60% |
| Ser | UCU, UCA, UCC, UCG, AGU, AGC | 8.10% |

Fig. 1. Graph showing the similarity between the observed frequencies of amino acids in 53 completely sequenced mammalian proteins and the frequencies predicted by the genetic code and random permutations of DNA nucleotides. The frequencies are in percentages of total amino acid content. The straight line represents an idealized equality of expectation and observation.

INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

On the other hand, leucine and serine, which have six codons, have some of the highest frequencies. One oddball here is arginine because it has six codons. However, its frequency is lower than, not as high as leucine and serine. So the graph you see here is something that has been plotted. This is the expected frequency that is plotted for these amino acids. This expected frequency is based on how many codons there are and how often those codons are used. So based on that, and on the y-axis, we get the observed frequency, and you will see that there is a nice correlation between these two, which means that amino acids with more codons tend to be observed more often than those with fewer codons. However, arginine is the exception, and you can see that it is an outlier. So based on this frequency, the expected frequency, the observed frequency, how many substitutions are happening for a particular pair, whether it is the same as expected, more than expected, or less than expected, this type of Blosum-62 matrix has been generated.

So we saw how this matrix is generated in the last lecture. So, again just a brief summary numbers which are 0 indicate that this particular substitution. So, here 0 is for lysine so lysine to asparagine. This substitution is almost the same that you would expect.



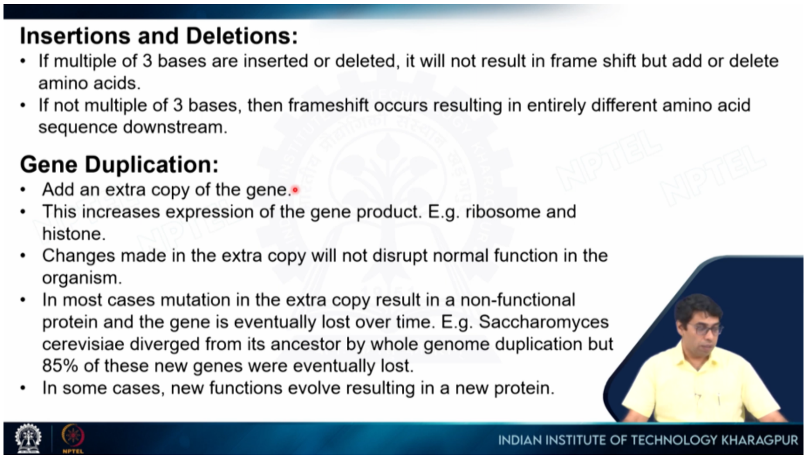
BLOSUM-62 matrix

Block Substitution Matrix gives a substitution score $S_{ij}$ for the substitution of amino acid i with j.

| j | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | C | 9 | | | | | | | | | | | | | | | | | | | |
| 2 | S | −1 | 4 | | | | | | | | | | | | | | | | | | |
| 3 | T | −1 | 1 | 5 | | | | | | | | | | | | | | | | | |
| 4 | P | −3 | −1 | −1 | 7 | | | | | | | | | | | | | | | | |
| 5 | A | 0 | 1 | 0 | −1 | 4 | | | | | | | | | | | | | | | |
| 6 | G | −3 | 0 | −2 | −2 | 0 | 6 | | | | | | | | | | | | | | |
| 7 | N | −3 | −1 | 0 | −2 | −2 | 0 | 6 | | | | | | | | | | | | | |
| 8 | D | −3 | 1 | −1 | −1 | −2 | −1 | 1 | 6 | | | | | | | | | | | | |
| 9 | E | −4 | 0 | −1 | −1 | −1 | −2 | 0 | 2 | 5 | | | | | | | | | | | |
| 10 | Q | −3 | 0 | −1 | −1 | −1 | −2 | 0 | 0 | 2 | 5 | | | | | | | | | | |
| 11 | H | −3 | 0 | −2 | −2 | −2 | −2 | 1 | −1 | 0 | 0 | 8 | | | | | | | | | |
| 12 | R | −3 | −1 | −1 | −2 | −1 | −2 | 0 | −2 | 0 | 1 | 0 | 5 | | | | | | | | |
| 13 | K | −3 | −1 | −1 | −1 | −1 | −2 | 0 | −1 | 1 | 1 | −1 | 2 | 5 | | | | | | | |
| 14 | M | −1 | 0 | −1 | −2 | −1 | −3 | −2 | −3 | −2 | 0 | −2 | −1 | −1 | 5 | | | | | | |
| 15 | I | −1 | −2 | −1 | −3 | −1 | −4 | −3 | −3 | −3 | −3 | −3 | −3 | −3 | 1 | 4 | | | | | |
| 16 | L | −1 | −2 | −1 | −3 | −1 | −4 | −3 | −4 | −3 | −2 | −3 | −2 | −2 | 2 | 2 | 4 | | | | |
| 17 | V | −1 | −2 | 0 | −2 | 0 | −3 | −3 | −3 | −2 | −2 | −3 | −3 | −2 | 1 | 3 | 1 | 4 | | | |
| 18 | F | −2 | −2 | −2 | −4 | −2 | −3 | −3 | −3 | −3 | −3 | −1 | −3 | −3 | 0 | 0 | 0 | −1 | 6 | | |
| 19 | Y | −2 | −2 | −2 | −3 | −2 | −3 | −2 | −3 | −2 | −1 | 2 | −2 | −2 | −1 | −1 | −1 | −1 | 3 | 7 | |
| 20 | W | −2 | −3 | −2 | −4 | −3 | −2 | −4 | −4 | −3 | −2 | −2 | −3 | −3 | −1 | −3 | −2 | −3 | 1 | 2 | 11 |
| | | C | S | T | P | A | G | N | D | E | Q | H | R | K | M | I | L | V | F | Y | W |
| i | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |

So, the observed frequency is the same as the expected frequency on the other hand, if I go to this lysine to glycine it is less than that is expected. So whatever frequency we expect, the observed frequency is less than that. So we have a negative number, on the other hand, if I go here, lysine to arginine, it is plus 2, which means that the observed frequency is more than expected.

So lysine to arginine or arginine to lysine, this exchange is observed more than you would expect based on the natural frequency of these amino acids. So this was all about point mutations. So apart from that there can also something that can happen that is insertion or deletion. Now I have a gene so that encodes for a particular protein if something is inserted now if this insertion is multiple of three then my reading frame is not changed, which means that if I insert let us say whatever my reading frame was, so I insert 6 amino acids, so it means that 2 amino acids will be inserted. But if it is not multiple of three, if I insert seven amino acids, then the first three will code for one amino acid.



**Insertions and Deletions:**

- If multiple of 3 bases are inserted or deleted, it will not result in frame shift but add or delete amino acids.
- If not multiple of 3 bases, then frameshift occurs resulting in entirely different amino acid sequence downstream.

**Gene Duplication:**

- Add an extra copy of the gene.
- This increases expression of the gene product. E.g. ribosome and histone.
- Changes made in the extra copy will not disrupt normal function in the organism.
- In most cases mutation in the extra copy result in a non-functional protein and the gene is eventually lost over time. E.g. Saccharomyces cerevisiae diverged from its ancestor by whole genome duplication but 85% of these new genes were eventually lost.
- In some cases, new functions evolve resulting in a new protein.

Second three will code for another amino acid. Then this last one, the seventh plus two more from next will code for a completely different amino acid. So it will change the sequences that will be downstream of this insertion. So this goes exactly the same for both insertion and deletion.

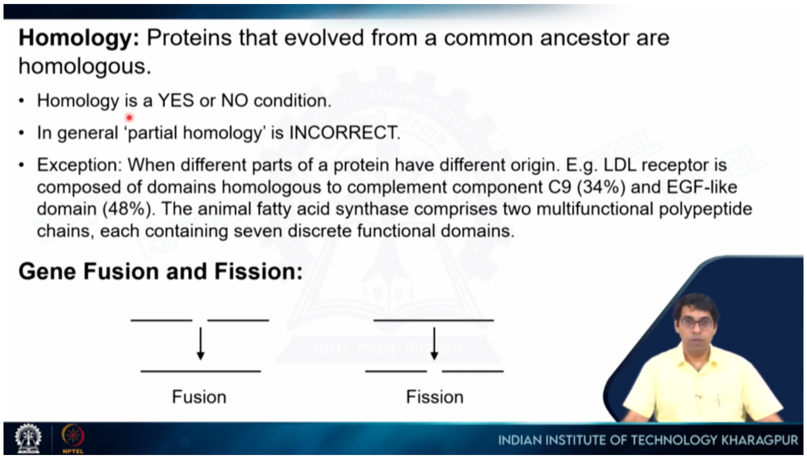
So any multiple of 3 will not cause much problem but if it is not multiple of 3 then it will create frame shift for the downstream sequence. Gene duplication, so this is something that happens quite often. So a particular gene gets duplicated means that it gets copied again.

Now this gene which is there it has a promoter on its upstream and that promoter will regulate its expression but if it gets copied somewhere else and if there is no promoter which results in its expression then that extra copy will start getting mutations and by chance it can show up in context of another promoter so that it gets expressed and it will become a different protein. In many cases what happens is that gene duplication happens with the promoters so that all the copies are actually expressed. So this is something that is observed for ribosomes and histone. So these proteins have multiple copies of the same gene.

There are several advantages to this. One is that your gene dosage is more so because you need a lot of ribosome and histone proteins. So since we have more genes, more mRNA will be produced and you will get more of this protein production. Second is that if one copy is corrupted by mutation, then the function will not be hampered because you have other functional copies. So in most cases, mutation in the extra copy results in non-functional protein and the gene is eventually lost forever.

But in these cases, the gene will still be there because you have extra functional copies. So, for example, Saccharomyces cerevisiae diverged from its ancestor by whole genome duplication, which means that this particular yeast, the whole genome was duplicated and 85% of these new genes were eventually lost, but 15% of the genes resulted in some new protein and that resulted in the evolution of this new organism. So in some cases, new functions evolve resulting in a new protein. So after gene duplication, the extra copy can evolve into a new protein with a new function.

So there is a particular term that is used a lot in this context, which is called homology. So homology means proteins that evolve from a common ancestor are homologous proteins. So this is a yes or no condition. There is no partial homology. So two proteins are either homologous or they are not homologous.



**Homology:** Proteins that evolved from a common ancestor are homologous.

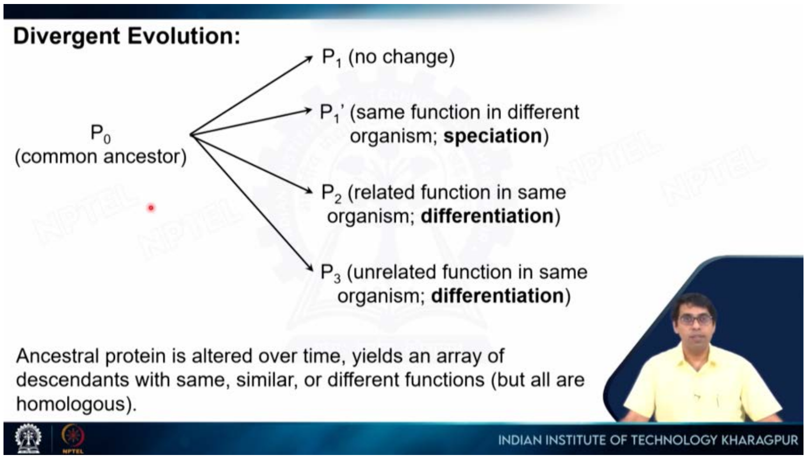
- Homology is a YES or NO condition.
- In general 'partial homology' is INCORRECT.
- Exception: When different parts of a protein have different origin. E.g. LDL receptor is composed of domains homologous to complement component C9 (34%) and EGF-like domain (48%). The animal fatty acid synthase comprises two multifunctional polypeptide chains, each containing seven discrete functional domains.

**Gene Fusion and Fission:**

However, there are problems that can occur if there are multiple domains in a protein. So one example is when different parts of a protein have different origins. So the LDL receptor is composed of domains that are homologous to the complement component C9 so 34% homology and an EGF-like domain. So it has 48% homology with this.

So it means that some sort of fusion occurred. So two different domains came from two different genes and they fused and now you have a protein that has homology with two different proteins. The animal fatty acid synthase comprises two multifunctional polypeptide chains, each containing seven discrete functional domains, and each of these domains might have come from different proteins. So these are all examples of gene duplication as well as fusion. So multiple genes come together to form a single gene;

however, they will fold as independent domains. Similarly, gene fission can also occur. If there is a protein with two different domains and the gene breaks here, then you will have two different proteins where each domain is now a separate protein. So evolution, in most cases, is divergent evolution. So what is divergent evolution? So let's say there is some ancestral species that has a particular protein and that is denoted by P0. So this is the common ancestor.



**Divergent Evolution:**

$P_0$ (common ancestor)

$P_1$ (no change)

$P_1'$ (same function in different organism; **speciation**)

$P_2$ (related function in same organism; **differentiation**)

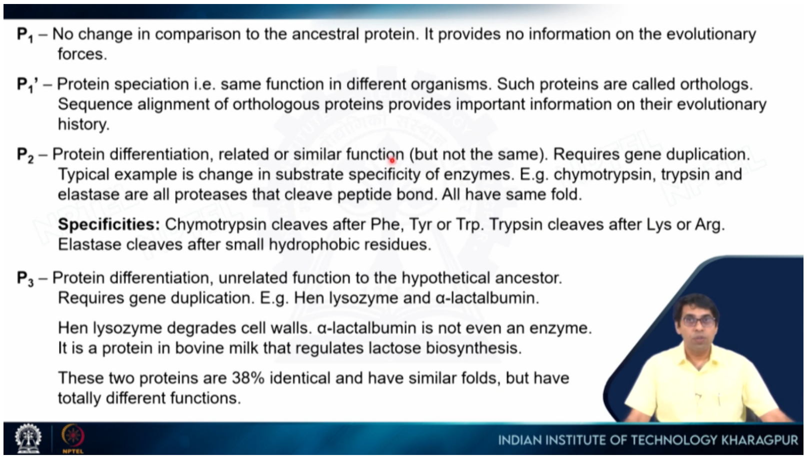
$P_3$ (unrelated function in same organism; **differentiation**)

Ancestral protein is altered over time, yields an array of descendants with same, similar, or different functions (but all are homologous).

INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

That species can survive. So it remains the same, and the protein also remains the same. So, let us say we are observing after, so this ancestor originated maybe a billion years back. We are, this is the present time. So the same ancestor is there. So there is no change in the protein. However, a new species can form where this protein is slightly changed. However, it performs the same function in a different organism. So this is called speciation. So it's the same protein; the sequence has changed, but the function remains the same.

The same protein or the same gene can get duplicated; the function changes may be very similar or slightly different, but in the same organism. So in this case, it is differentiation. So the sequences will be very similar. However, they have slightly different functions or they can be completely different functions so these two are in the same organism. So these two proteins are originating from the same ancestor. However, they have gained so many mutations that they can now perform either related functions or very different functions. So this will be differentiation. In this case, this sequence will be slightly different from this, but they perform the same function in different organisms. So this will be speciation. So the ancestral protein is altered over time. It yields an array of descendants with same, similar or different functions. So these are all homologous proteins. So these proteins will be homologous to each other and to this ancestor. So whatever I mentioned is written here. So $P_1$, there is no change in comparison to the ancestral protein. $P_{1'}$, this is speciation.



**$P_1$** – No change in comparison to the ancestral protein. It provides no information on the evolutionary forces.

**$P_1$'** – Protein speciation i.e. same function in different organisms. Such proteins are called orthologs. Sequence alignment of orthologous proteins provides important information on their evolutionary history.

**$P_2$** – Protein differentiation, related or similar function (but not the same). Requires gene duplication. Typical example is change in substrate specificity of enzymes. E.g. chymotrypsin, trypsin and elastase are all proteases that cleave peptide bond. All have same fold.

**Specificities:** Chymotrypsin cleaves after Phe, Tyr or Trp. Trypsin cleaves after Lys or Arg. Elastase cleaves after small hydrophobic residues.

**$P_3$** – Protein differentiation, unrelated function to the hypothetical ancestor. Requires gene duplication. E.g. Hen lysozyme and α-lactalbumin.

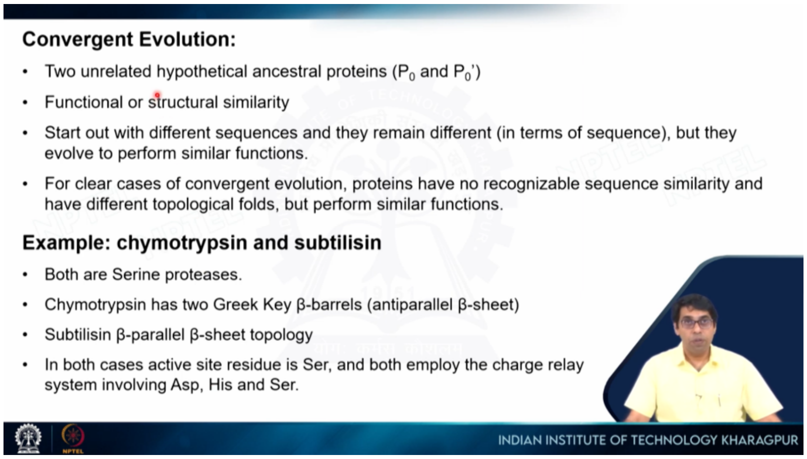
Hen lysozyme degrades cell walls. α-lactalbumin is not even an enzyme. It is a protein in bovine milk that regulates lactose biosynthesis.

These two proteins are 38% identical and have similar folds, but have totally different functions.

So such proteins are called orthologs. So sequence alignment of orthologs proteins provides important information on their evolutionary history because it's the same protein. So for example, we saw Cytochrome C in all these different organisms and then we used that to create this tree of life. So these types of orthologs are very useful. Differentiation, here the protein has related or similar function but not exactly the same function. So this requires gene duplication examples will be chymotrypsin trypsin and elastase these are all enzymes which cleave the peptide bond so these are called proteases they have very similar fold. So chymotrypsin cleaves after these bulky aromatic groups, trypsin cleaves after the basic amino acids and elastase cleaves after residues which have small hydrophobic side chains. So these are all proteases and they will be the result of differentiation and then the third one is where the function is completely unrelated.

For example, hen lysozyme which breaks the bacterial cell wall and alpha lactal bovine which is not even an enzyme. Alpha(α)-lactalbumin, it's a protein in bovine milk that regulates lactose biosynthesis. These two proteins are 38% identical and they have similar folds but have completely different functions. So this was a divergent evolution.

We also have something called convergent evolution, and this is something that is rare. We will see an example of convergent evolution in the next lecture, where I discuss the evolution of the eye. Convergent evolution means that two unrelated proteins, starting from different ancestors, converge to the same function. They start with different sequences and remain different, but they evolve to perform similar functions. In the case of the eye, we will see that eyes have evolved in different ways and perform the same function, which is vision, but their designs are completely different. In clear cases of convergent evolution, proteins have no recognizable sequence similarity or identity.



**Convergent Evolution:**

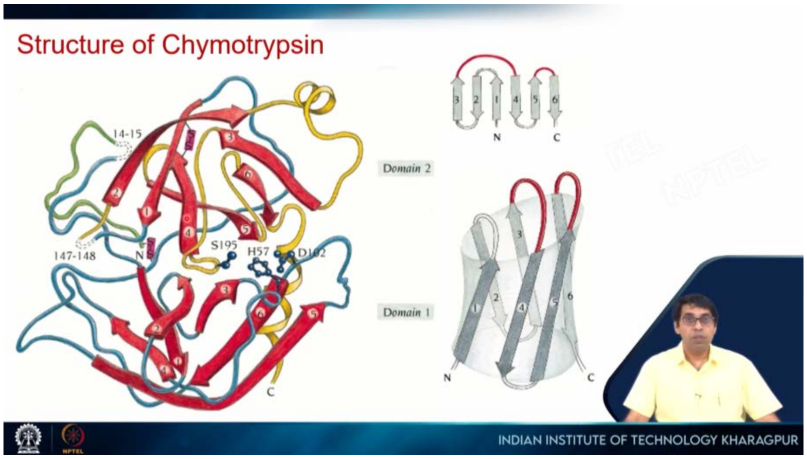
- Two unrelated hypothetical ancestral proteins ($P_0$ and $P_0$')
- Functional or structural similarity
- Start out with different sequences and they remain different (in terms of sequence), but they evolve to perform similar functions.
- For clear cases of convergent evolution, proteins have no recognizable sequence similarity and have different topological folds, but perform similar functions.

**Example: chymotrypsin and subtilisin**

- Both are Serine proteases.
- Chymotrypsin has two Greek Key β-barrels (antiparallel β-sheet)
- Subtilisin β-parallel β-sheet topology
- In both cases active site residue is Ser, and both employ the charge relay system involving Asp, His and Ser.

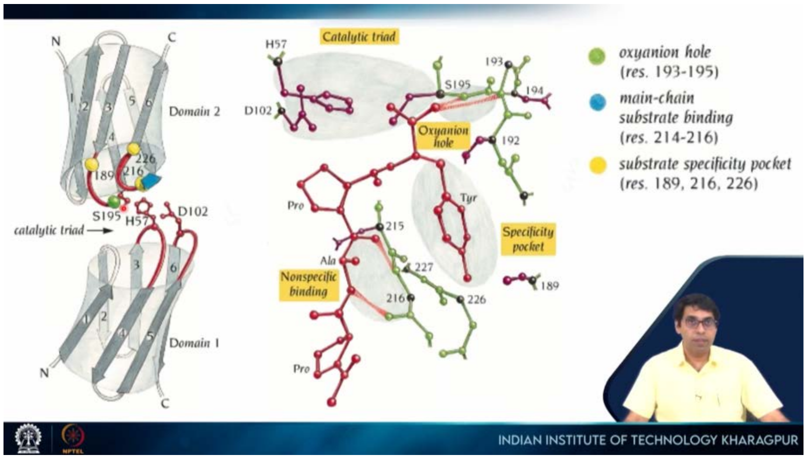
Their structures will also be different, but they will perform similar functions. At the protein level, a very good example is that of chymotrypsin and subtilisin. Both are proteases, specifically serine proteases. Chymotrypsin has two key beta barrels, Greek key beta barrels. That is its structural topology, whereas subtilisin has a beta-parallel, beta-sheet topology.

In both cases, the active site residue is serine. Serine is the one that performs the reaction, and both employ a charge relay system involving aspartic acid, histidine, and serine. The active site has the same residues, and the reaction mechanism is also very similar, but the structure is completely different, and the sequence is also very different. This is the structure of chymotrypsin. It has two domains. You can see there is one domain here. There is the other domain here and if we look through this domain from this direction, it will look like a beta barrel like this and the active site residues are in the interface of this beta barrel. So there is a serine, there is a histidine, there is a aspartic acid and together they form the catalytic triad.

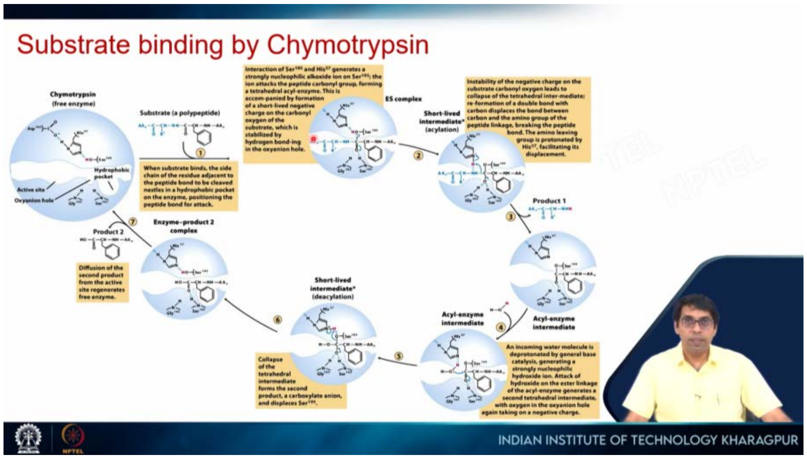


Structure of Chymotrypsin

So I have already discussed chymotrypsin in detail in a previous lecture. One important thing that I should point out here is you will see that these two domains look very similar and both have this beta barrel structure. So it is hypothesized that the original chymotrypsin had only one domain and then gene duplication happened where the same gene was copied just right next to each other to result in this complete protein and that resulted in this two beta barrel structure. However, it is interesting to note that serine is in one of the beta barrels and these two are in the other beta barrel.



This is the mechanism at a glance. I have already discussed this in a previous lecture. So serine chymotrypsin, it is formed as a pre-protein, pre-pro-protein and then it gets activated by other enzymes and then it forms this functional enzyme.



Substrate binding by Chymotrypsin

Interaction of Ser$^{195}$ and His$^{57}$ generates a strongly nucleophilic alkoxide ion on Ser$^{195}$; the ion attacks the peptide carbonyl group, forming a tetrahedral acyl-enzyme. This is accom-panied by formation of a short-lived negative charge on the carbonyl oxygen of the substrate, which is stabilized by hydrogen bond-ing in the oxyanion hole.

When substrate binds, the side chain of the residue adjacent to the peptide bond to be cleaved nestles in a hydrophobic pocket on the enzyme, positioning the peptide bond for attack.

Instability of the negative charge on the substrate carbonyl oxygen leads to collapse of the tetrahedral inter-mediate; re-formation of a double bond with carbon displaces the bond between carbon and the amino group of the peptide linkage, breaking the peptide bond. The amino leaving group is protonated by His$^{57}$, facilitating its displacement.

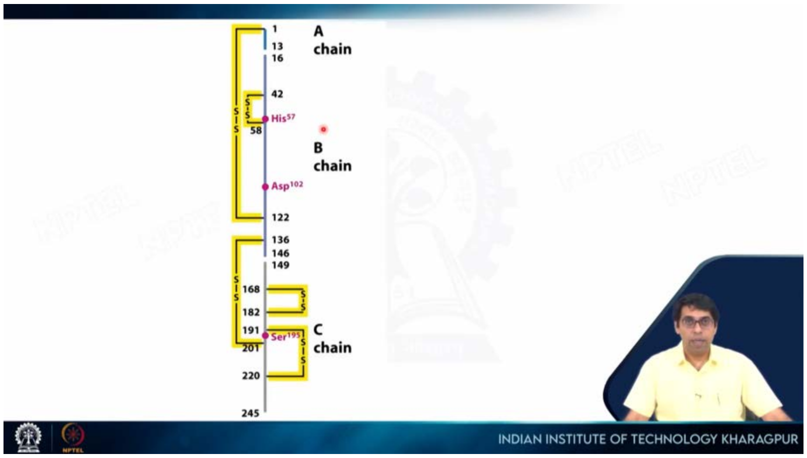
An incoming water molecule is deprotonated by general base catalysis, generating a strongly nucleophilic hydroxide ion. Attack of hydroxide on the ester linkage of the acyl-enzyme generates a second tetrahedral intermediate, with oxygen in the oxyanion hole again taking on a negative charge.

Collapse of the tetrahedral intermediate forms the second product, a carboxylate anion, and displaces Ser$^{195}$.

Diffusion of the second product from the active site regenerates free enzyme.

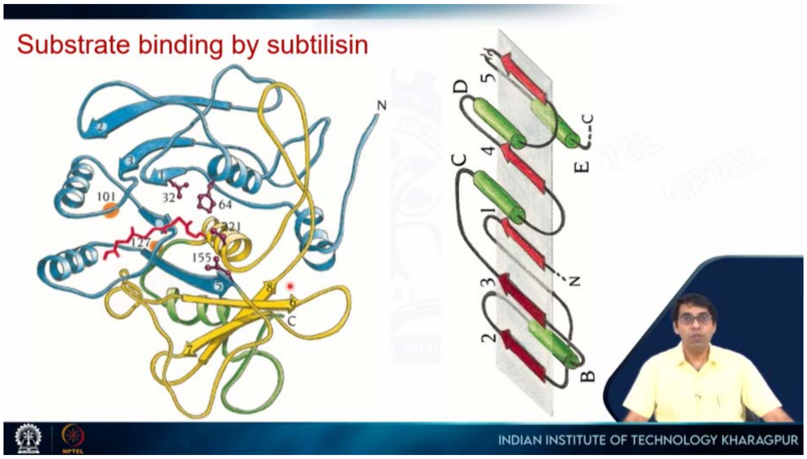
INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

So if we look at subtilisin, it looks like this and you can see its topology is completely different and also it looks very different in structure. However, it has the same three residues in its active site, very similar arrangement and the same residues in the active site.

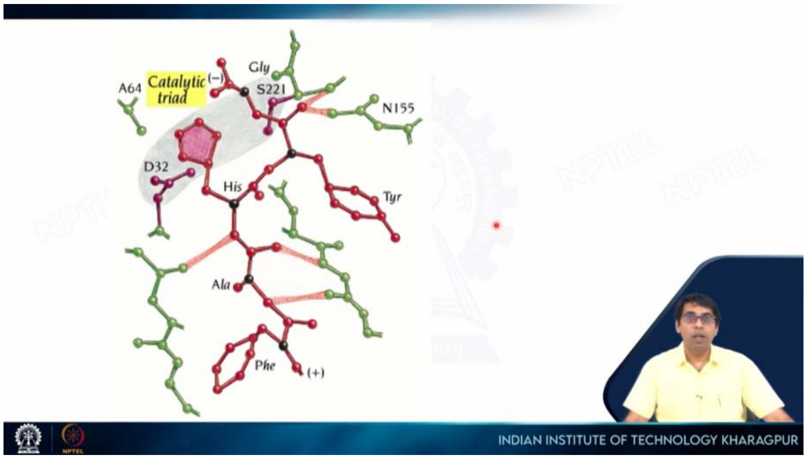


INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

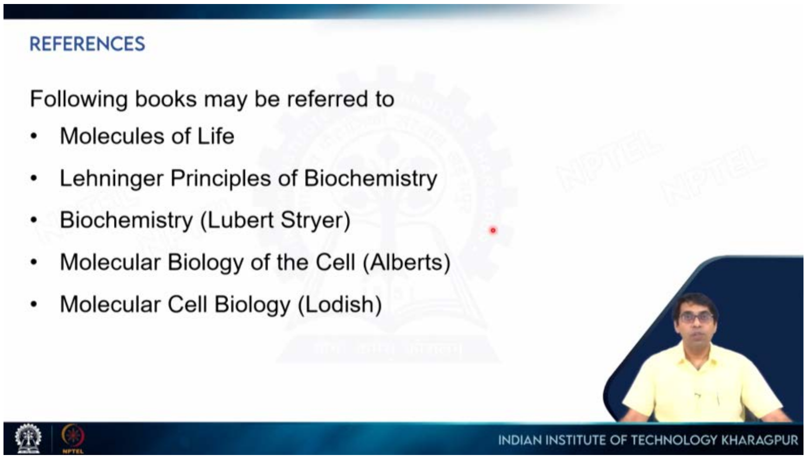
So it means that the same catalytic active site has evolved from two different ancestors. So they have converged into the same function. That is why this is an example of a convergent evolution.


Substrate binding by subtilisin

So this is the active site of the subtilisin. So we have a serine here, a histidine here, and we have an aspartic acid here and together they form the catalytic triad for subtilisin.



So for this lecture, you can go through these books, especially Lehninger Principles of Biochemistry and Molecules of Life, these two. Thank you.


REFERENCES

Following books may be referred to

- Molecules of Life
- Lehninger Principles of Biochemistry
- Biochemistry (Lubert Stryer)
- Molecular Biology of the Cell (Alberts)
- Molecular Cell Biology (Lodish)