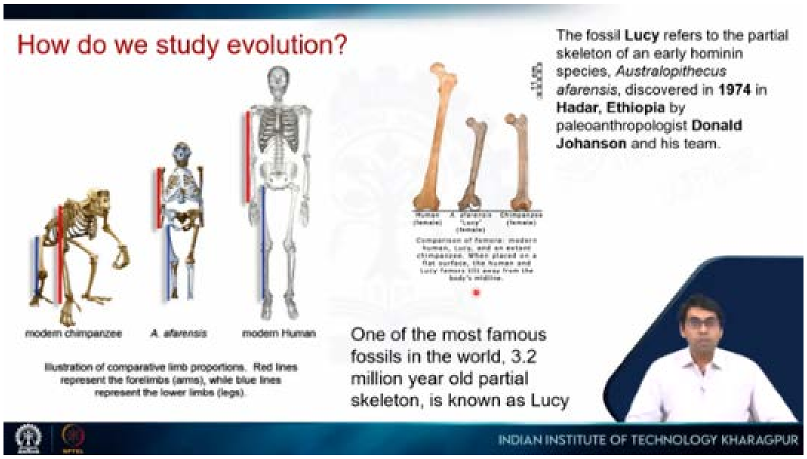
Introduction to Complex Biological Systems
Professor Dibyendu Samanta and Professor Soumya De
Department of Bioscience and Biotechnology
Indian Institute of Technology, Kharagpur

Lecture 37
Protein evolution

Welcome back to Week 8 of Introduction to Complex Biological Systems. So today, I am going to discuss protein evolution. So this week, I am discussing evolution, the history of life. So today, I am going to discuss protein evolution. How do we study evolution?

So let us look at one example. We normally study evolution based on fossil records that we obtain. Now here, you see three different bones are shown. So these are fossils. So this is from a female chimpanzee.



How do we study evolution?

Illustration of comparative limb proportions. Red lines represent the forelimbs (arms), while blue lines represent the lower limbs (legs).

Comparison of femora: modern human, Lucy, and an extant chimpanzee. When placed on a flat surface, the human and Lucy femora tilt away from the body's midline.

The fossil **Lucy** refers to the partial skeleton of an early hominin species, *Australopithecus afarensis*, discovered in **1974** in **Hadar, Ethiopia** by paleoanthropologist **Donald Johanson** and his team.

One of the most famous fossils in the world, 3.2 million year old partial skeleton, is known as Lucy

This is from a human female, and this is from some other species. So if you look at this particular bone, you will see that the length of this bone is similar to this. However, when you place this bone on a horizontal surface like this, it tilts to the right, just like the human bone. This is something that is difficult for us, but if you place a chimpanzee bone, this is the femur; it will stand vertically like this.

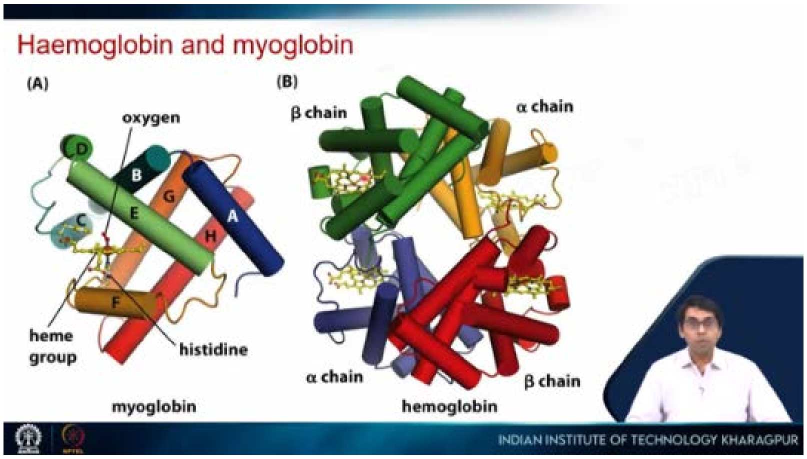
Now, this tells you that whatever this species is, it sort of shares a characteristic between this and this so these two species. We know when chimpanzees appeared and when the first human species appeared. So this is something that is in between and then we can do carbon dating to know when this particular, how old this fossil is.

So this is actually one of the most famous fossils in the world. It is a 3.2 million years old partial skeleton called Lucy. So a partial skeleton means that almost 40% of the skeleton, the complete skeleton was recovered in this particular dig. So the fossil Lucy refers to this partial skeleton of an early hominin species. So it is called Australopithecus afarensis.

A. afarensis was discovered in 1974 in Ethiopia by Donald Johnson and his team. So this is how we sort of study evolution, that we look at this fossil record based on the height of this. We can say that this species was not as tall as modern humans, but they were somewhere in the same range as these chimpanzees. However, this tilt tells you that they would, they can walk upright like us instead of the chimpanzees. So chimpanzees cannot walk upright like us, but this hominin species can actually walk upright. However, from the fossil record, from the skull, you can tell that the brain of this species was not very big. So it was not as large as that of modern humans. It was somewhat similar to that of modern chimpanzees.

So you can also conclude that the evolution of walking upright occurred earlier than the evolution of increased brain size. So this is how we can draw all these different conclusions. Now, protein evolution can be studied in a very similar manner. So we can compare two proteins which have the same structure.

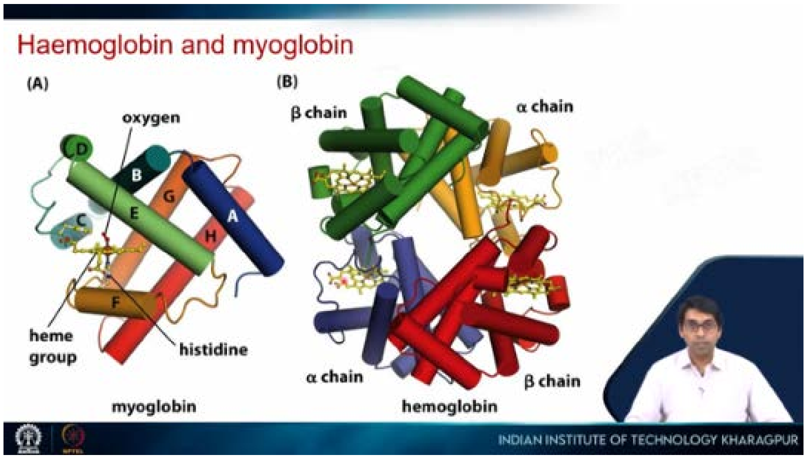
So we are comparing the same bone here. We are not comparing different bones. We are comparing the same bones here. Similarly, you have to compare the same proteins from different species and then by looking at how different they are, we can tell which one appeared first or if we have two different proteins and we find something that is in between, we can say that if this is modern, this is older, this will be somewhere in between because it shares traits with both proteins. So that is something that we do for proteins. So let us see how we do that. So I will take up the example of hemoglobin and myoglobin.



Haemoglobin and myoglobin

INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

So I have talked a lot about hemoglobin. Hemoglobin is the protein that carries oxygen. So it has this quaternary structure, four different polypeptide chains are there with the heme group. So there are two alpha chains and two beta chains.

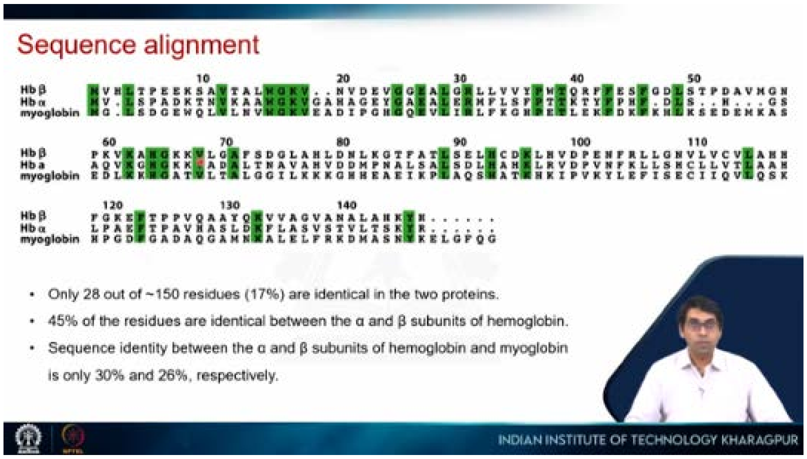


Haemoglobin and myoglobin

INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

The monomeric form of this protein is called myoglobin. So it has only one polypeptide chain. It has one heme group. It also binds oxygen. If we compare the amino acid sequence of hemoglobin and myoglobin, these are the two chains of hemoglobin: the alpha chain, the beta chain, and myoglobin.

So we have written all the amino acid sequences here. It turns out that only 28 out of 150 residues are identical in these proteins so identical means that this position is all methionine, all leucine, all valine, like so. So if you count, there are only 28 amino acids like this which are the same in all three sequences out of almost 150 amino acids so only 70% identical.



Sequence alignment

```
                      10                  20                  30                  40                  50
Hb β        M V H L T P E E K S A V T A L W G K V . . N V D E V G G E A L G R L L V V Y P W T Q R F F E S F G D L S T P D A V M G N
Hb α        M V . L S P A D K T N V K A A W G K V G A H A G E Y G A E A L E R M F L S F P T T K T Y F P H F . D L S . . H . . . G S
myoglobin   M G . L S D G E W Q L V L N V W G K V E A D I P G H G Q E V L I R L F K G H P E T L E K F D K F K H L K S E D E M K A S

               60                  70                  80                  90                  100                 110
Hb β        P K V K A H G K K V L G A F S D G L A H L D N L K G T F A T L S E L H C D K L H V D P E N F R L L G N V L V C V L A H H
Hb α        A Q V K G H G K K V A D A L T N A V A H V D D M P N A L S A L S D L H A H K L R V D P V N F K L L S H C L L V T L A A H
myoglobin   E D L K K H G A T V L T A L G G I L K K K G H H E A E I K P L A Q S H A T K H K I P V K Y L E F I S E C I I Q V L Q S K

               120                 130                 140
Hb β        F G K E F T P P V Q A A Y Q K V V A G V A N A L A H K Y H . . . . . .
Hb α        L P A E F T P A V H A S L D K F L A S V S T V L T S K Y R . . . . . .
myoglobin   H P G D F G A D A Q G A M N K A L E L F R K D M A S N Y K E L G F Q G
```

- Only 28 out of ~150 residues (17%) are identical in the two proteins.
- 45% of the residues are identical between the α and β subunits of hemoglobin.
- Sequence identity between the α and β subunits of hemoglobin and myoglobin is only 30% and 26%, respectively.

INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

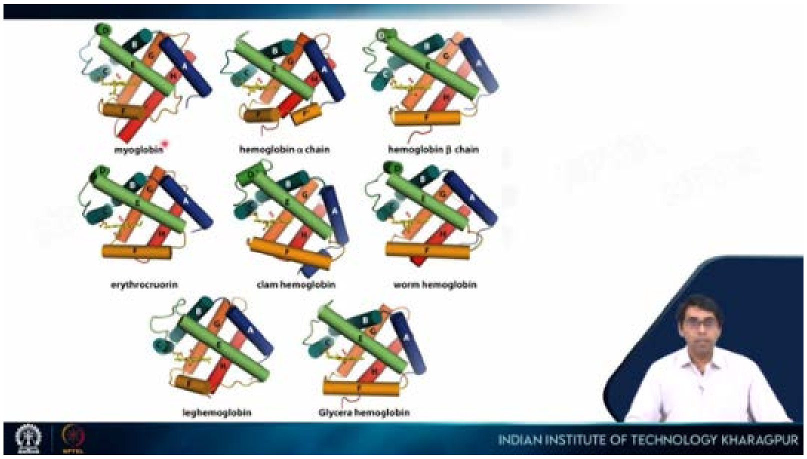
It means that all the other positions can change but if you look at the structure, they look very similar. So there are exactly the same numbers of helices. So I think there are seven helices. There are the same number of helices in them and if you superimpose the structures. So take one of these subunits and superimpose it with this then you will see they align quite nicely. So even with only 70% identical residues, the proteins fold in a similar manner. It means that we have to look for something beyond the identical residues, 45% of the residues are identical between the alpha and beta subunits of hemoglobin.

So, if we ignore myoglobin, if we take only these first two, then you will get this V, you will get this P. You will get this K. So, you will get some additional residues. So, these two are K, but here it is W. So, you will get these additional residues, and it turns out that it increases from 17% to 45%, but still it is less than 50%. Sequence identity between alpha and beta subunits of hemoglobin and myoglobin is only 30% and 26%, respectively, which means that if I remove beta, if I just look at alpha and myoglobin that will be 30%. If I remove alpha(α), if I look at beta(β) and myoglobin, that will be 26%. So the best that we see is 45% sequence identity. So, it means that we also have to look at something else apart from identical residues, and that is called similarity. So, what do I mean by similarity, for example, if I look here, the hemoglobin alpha chain has a valine, beta chain has a leucine, and myoglobin also has a leucine. Now, if you look at the chemical structures of valine and leucine, they are very similar, only one carbon difference.

So, both are hydrophobic, and both have two methyl groups. So, you can imagine that we can easily replace, or we can more often replace, a valine with a leucine or a leucine with a valine. So, valine and leucine will be residues which are similar. Similarly, if I think about this, here in hemoglobin, it is glutamic acid E, and in myoglobin, it is aspartic acid D. So, both are acidic groups.

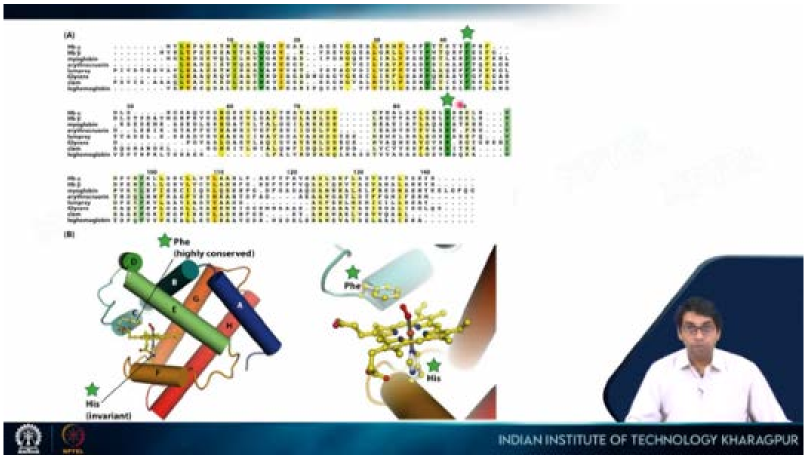
So if you replace one with another, you are not going to change the characteristics of the side chain. So that will also be, these will also be similar residues. So you can replace one with the other. Similarly, if we replace phenylalanine with tyrosine, then they will be similar. So, we can group residues like that according to the chemistry of the side chain as similar residues.

So, now let us look at myoglobin, hemoglobin alpha chain, and beta chain, again, similar proteins from other organisms. So there are a total of eight different protein chains from these different organisms. If you look at the structure, they look exactly the same.



Now we can align their amino acid sequences and see how similar or how different they are. So if I do an alignment of all these eight amino acid sequences, this is something that you will see so only two residues shown here are absolutely conserved.

So this is the histidine. So it is histidine in all the sequences, and this is phenylalanine in all the sequences. Everything else can change, and they are color-coded. So anything that is in darker green means it is more conserved. So you see that it is mostly proline except Q here. This is mostly tryptophan except in two positions where it becomes phenylalanine. Here it is either valine or isoleucine so almost 50-50 as you go lighter and lighter.



INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

So in this case, valine, isoleucine, nothing. Valine, isoleucine, phenylalanine. So hydrophobic residues, but still it changes from a methyl group to an aromatic group and then there are positions which are white.

So it can become anything: leucine, phenylalanine, histidine, threonine, aspartic acid, isoleucine, asparagine, valine. So all eight are different and that is why it is white. So from this, you can tell that there are many positions where you can completely change the amino acid, without any consequence on the structure of the protein.

In fact, there are regions where you can delete amino acids and still the protein folds in the same manner. So this poses an interesting question: how do they do this sequence alignment? How do you know what aligns with what? So here it is exactly identical. So I know that this is identical, but then you are putting a gap here.

So instead of aligning this aspartic acid here, I can put this aspartic acid here but I have not done that. I have put three gaps and then I have put the aspartic acid here. So this histidine I can easily put here because this histidine comes after this serine. So I have not done that.

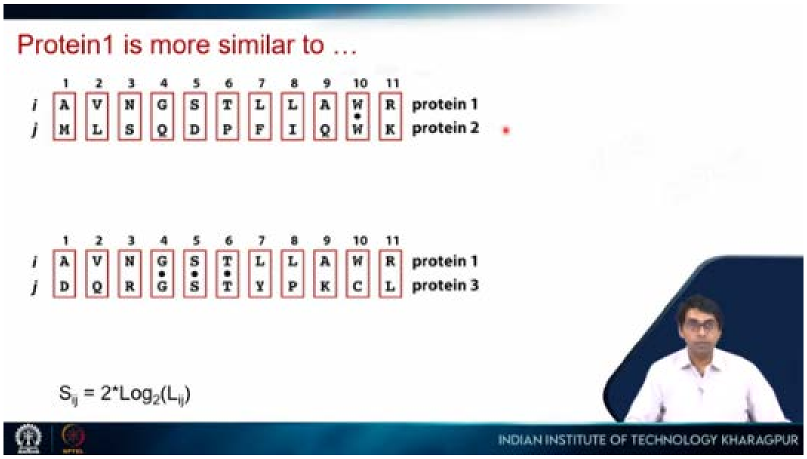
I have put this huge gap and I have put this histidine here. So how do you do that? How do you know which residues are aligning with each other, where you have to put these gaps? So I can align two residues like this or I can align them like this or I can align them like this.

So how do you align these residues? And can we measure this? Can we quantify how similar two sequences are? So this will be something that will be very useful for further studies. So now before I do that let's just point out why these two residues are conserved. So if you look at the structure this histidine is the one which directly coordinates with the iron and histidine is the only residue which can fit there and which can do this coordination. So you cannot histidine, you cannot change this particular histidine with any other amino acid. So that's why this histidine is absolutely conserved. On the other hand, this phenylalanine which is also conserved shows up here and this is also conserved because this phenylalanine allows the oxygen binding and binding in the correct orientation. If I change this phenylalanine with a tyrosine, which means I will put an oxygen group here, then there will be no place for this oxygen binding.

If I change it with tryptophan, the structure will change. So no other amino acid will be tolerated here because of this very strict requirement. That is why phenylalanine is absolutely conserved. So histidine and phenylalanine, these two amino acids are absolutely conserved. The remaining amino acids you can change, but again, that change cannot be random.

So there are certain characteristics that have to be preserved at certain positions. So let's look in more detail at how we can compare two sequences. So what I have done is I have given two sequences. So this top one is from protein 1 and this bottom one is from protein two.

Similarly, again, the top one is from Protein 1. So it is the exact same sequence shown up here. This bottom one is from protein three. Now, the question that you can ask yourself is whether Protein 2 is more similar to Protein 1 or Protein 3 is more similar to Protein 1. So which one is more similar to Protein 1, Protein 2 or Protein 3?



Protein1 is more similar to …

	1	2	3	4	5	6	7	8	9	10	11	
i	A	V	N	G	S	T	L	L	A	W	R	protein 1
j	M	L	S	Q	D	P	F	I	Q	W	K	protein 2

	1	2	3	4	5	6	7	8	9	10	11	
i	A	V	N	G	S	T	L	L	A	W	R	protein 1
j	D	Q	R	G	S	T	Y	P	K	C	L	protein 3

$S_{ij} = 2*\mathrm{Log}_2(L_{ij})$

INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

So let us say these three sequences come from protein, the same protein but from different species. So based on that, we can also answer whether Species 2 is more similar to Species 1 or Species 3 is more similar to Species 1. So if you look at this, here only one amino acid is identical, everything else is different and in this case, three amino acids are identical, everything else is different.

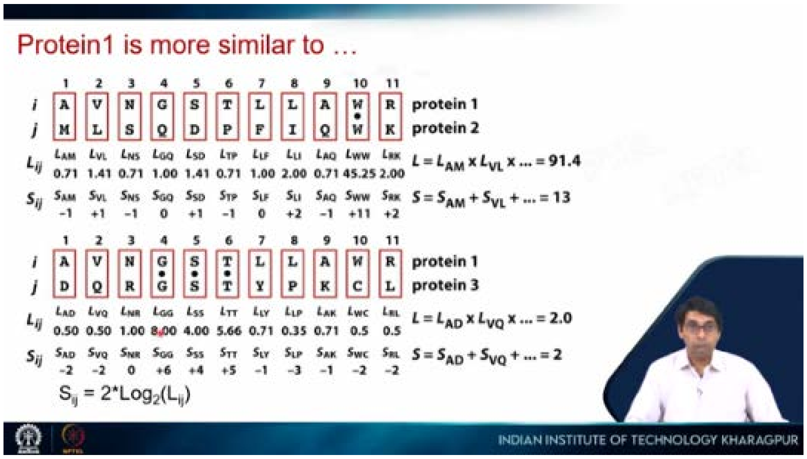
So, if I go by sequence identity, then only 1 is identical here and 3 are identical here. So, based on identity, I will say Protein 3 is more similar to Protein 1 than Protein 2. So, Protein 3 is more similar to Protein 1. But will that be correct? Because we also have to take into consideration something called similarity. So if i look at this let's say asparagine and serine both are polar residues but here asparagine is changed to an arginine which is a charged residue. Arginine to lysine both are positively charged arginine to leucine so positively charged becomes a hydrophobic residue leucine to isoleucine, leucine to proline. Proline is an amino acid with a very different structure. So, alanine to methionine, both are hydrophobic, alanine to aspartic acid, which is negatively charged.

So, we have to take into consideration these other substitutions also. So, if we do that, there are these numbers which are obtained for a pair of substitutions. So 0.71 for substituting alanine with methionine, 1.41 for substituting valine with leucine, so on and so forth. Now these same numbers are converted to, if you take the log, so log of this, so 2 log to the base 2 of this L gives you this S.

Now if it is this, you have to multiply all of them and you get something. But instead of multiplying numbers, if I take the log, then I can just add them. Addition is computationally easier compared to multiplication. So that is why people convert this into a log scale so that you can just add these coefficients and get some number.

So Protein 1 and Protein 2, if you do this, you get this number as 13. What do we get for protein 1 and 3? So if I do that for Protein 1 and 3, I get this number as 2. So clearly this is greater than this. So higher this number, the more similar these two proteins are.

So it means protein 1 is more similar to Protein 2 than Protein 3. So even though we have more identical residues here, if you take into consideration similarity, then Protein 1 is more similar to Protein 2 than Protein 3. So now the question is: how do we get these numbers? How can we tell what the weightage is of substituting one amino acid with another amino acid? This type of weightage is given by a matrix called the BLOSUM matrix, or block substitution matrix.



Protein1 is more similar to …

	1	2	3	4	5	6	7	8	9	10	11	
i	A	V	N	G	S	T	L	L	A	W	R	protein 1
										•		
j	M	L	S	Q	D	P	F	I	Q	W	K	protein 2
L_{ij}	L_{AM}	L_{VL}	L_{NS}	L_{GQ}	L_{SD}	L_{TP}	L_{LF}	L_{LI}	L_{AQ}	L_{WW}	L_{RK}	$L = L_{AM} \times L_{VL} \times \ldots = 91.4$
	0.71	1.41	0.71	1.00	1.41	0.71	1.00	2.00	0.71	45.25	2.00	
S_{ij}	S_{AM}	S_{VL}	S_{NS}	S_{GQ}	S_{SD}	S_{TP}	S_{LF}	S_{LI}	S_{AQ}	S_{WW}	S_{RK}	$S = S_{AM} + S_{VL} + \ldots = 13$
	-1	+1	-1	0	+1	-1	0	+2	-1	+11	+2	

	1	2	3	4	5	6	7	8	9	10	11	
i	A	V	N	G	S	T	L	L	A	W	R	protein 1
				•	•	•						
j	D	Q	R	G	S	T	Y	P	K	C	L	protein 3
L_{ij}	L_{AD}	L_{VQ}	L_{NR}	L_{GG}	L_{SS}	L_{TT}	L_{LY}	L_{LP}	L_{AK}	L_{WC}	L_{RL}	$L = L_{AD} \times L_{VQ} \times \ldots = 2.0$
	0.50	0.50	1.00	8.00	4.00	5.66	0.71	0.35	0.71	0.5	0.5	
S_{ij}	S_{AD}	S_{VQ}	S_{NR}	S_{GG}	S_{SS}	S_{TT}	S_{LY}	S_{LP}	S_{AK}	S_{WC}	S_{RL}	$S = S_{AD} + S_{VQ} + \ldots = 2$
	-2	-2	0	+6	+4	+5	-1	-3	-1	-2	-2	

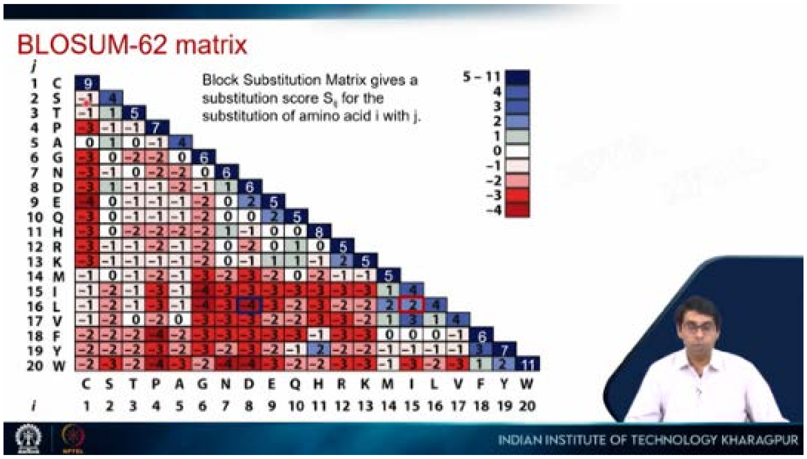
$S_{ij} = 2^{*}\mathrm{Log}_2(L_{ij})$

INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

So if you look at this matrix, it gives you the weightage of substituting one amino acid with another amino acid. So if I take serine, if I substitute a serine with a cysteine, it gives me minus one. Zero means there is no preference for this change. So it can happen.

So there is a random chance of mutation of a serine to cysteine. If you have this number as 0, it means that this substitution is observed, which equals this random chance. If the number is negative, it means that this substitution happens less often than what you would expect by random chance, which means that this substitution is unfavorable. Since it is unfavorable, we see it less often.

If it is positive, then this substitution is favorable, favorable in the sense that these two residues are very similar. So, the change between them happens more often than you would expect by random chance. So, let us look at some numbers, Glycine versus cysteine. You see, glycine versus cysteine is -3. So cysteine has this side chain SH, and glycine is the smallest amino acid. So glycine will not be changed to cysteine, nor will cysteine be changed to glycine. That would be detrimental to the protein.



BLOSUM-62 matrix

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

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

INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

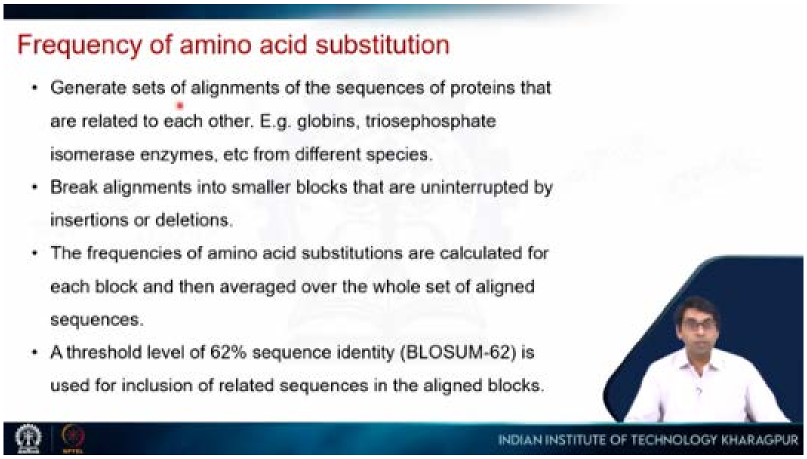
So glycine-cysteine or cysteine-glycine mutations are seen less often than you would expect by random chance. On the other hand, if you see this there are three. So valines to isoleucine, both are hydrophobic side chains. Both have two methyl groups so both are hydrophobic amino acids. Both have methyl groups as their side chains. So substitution of isoleucine with valine or substitution of valine with isoleucine is seen more often.

Then you would expect by random chance. So that is this positive number that signifies. So I have marked some of them here, for example, isoleucine with leucine and leucine with aspartic acid. So this is negative and this is positive. So you can explain to yourself that why this is negative and why this is positive.

You will also see a number in the diagonal. Now, this 11 means changing tryptophan with tryptophan. So, of course, you are not mutating tryptophan with tryptophan. So, why do we have this number in the diagonal? The number in the diagonal signifies how hard it is to mutate tryptophan into something else.

So, the higher this number, the more difficult it is to mutate this particular amino acid. So, tryptophan is the one which is the hardest. Then it is cysteine because it can form disulfide linkages, which no other amino acid can. Then it is histidine, which has a very unique pKa for its side chain.

So, the frequency of amino acid substitution is something that is given by this type of block substitution matrix. So, how do you determine these numbers? To do that, what is done is a set of alignments of sequences of proteins that are related to each other.



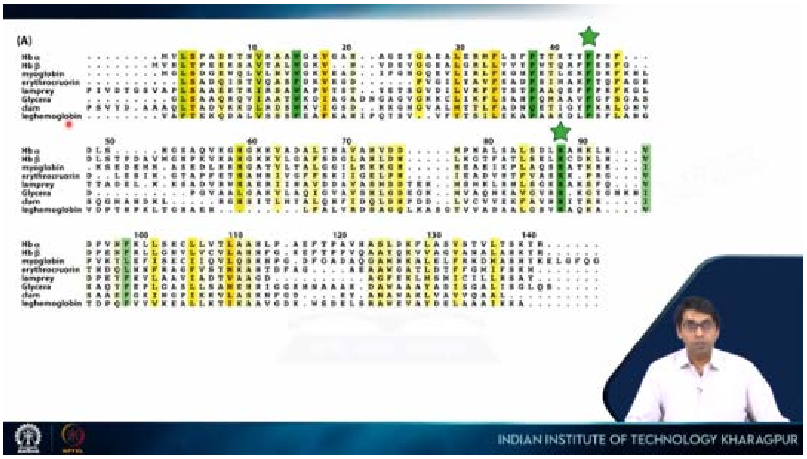
Frequency of amino acid substitution

- Generate sets of alignments of the sequences of proteins that are related to each other. E.g. globins, triosephosphate isomerase enzymes, etc from different species.
- Break alignments into smaller blocks that are uninterrupted by insertions or deletions.
- The frequencies of amino acid substitutions are calculated for each block and then averaged over the whole set of aligned sequences.
- A threshold level of 62% sequence identity (BLOSUM-62) is used for inclusion of related sequences in the aligned blocks.

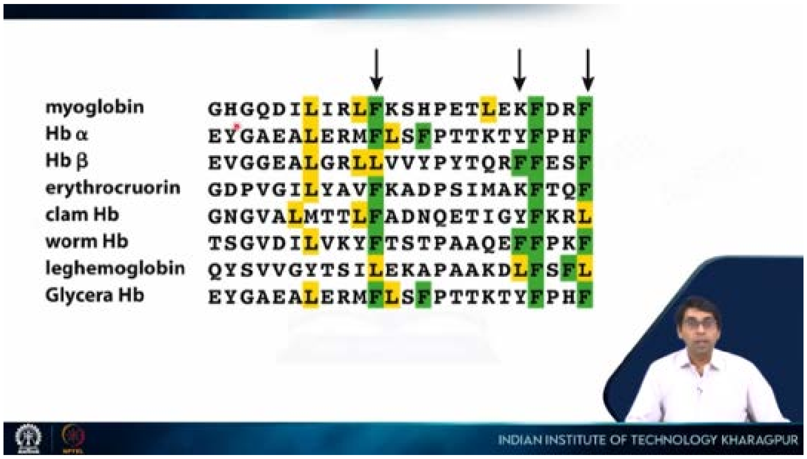
We saw the example of hemoglobin, myoglobin and all these other proteins so globins or one can take triose phosphate isomerase enzymes or some other enzyme from different species. So you get this bunch of amino acid sequences and you align them. You break these alignments into smaller blocks that are uninterrupted by insertions or deletions so there are no gaps so you take such blocks and there you measure the frequencies of this amino acid substitution which I will show in the next slide. Now, one of the factors that is done is a threshold level of 62% sequence identity is used, which means that no two sequences will have more than 62% identity because if there are sequences which have more sequence identity, then they will bias your numbers, your statistics. So, sequences are taken which do not have more than 62% identity and that is why this is called block summation 62. So this number was arrived at by trial and error and this was found to be something that works very well.

So let us look at this alignment again. So we have taken these 8 sequences and we have aligned them somehow. Now we want to look at places where there are no interruptions.

So I can look at this. So I can take this block. So up to this and this, I can take this block where there are no insertions or deletions. I cannot take this region because there are so many gaps here.

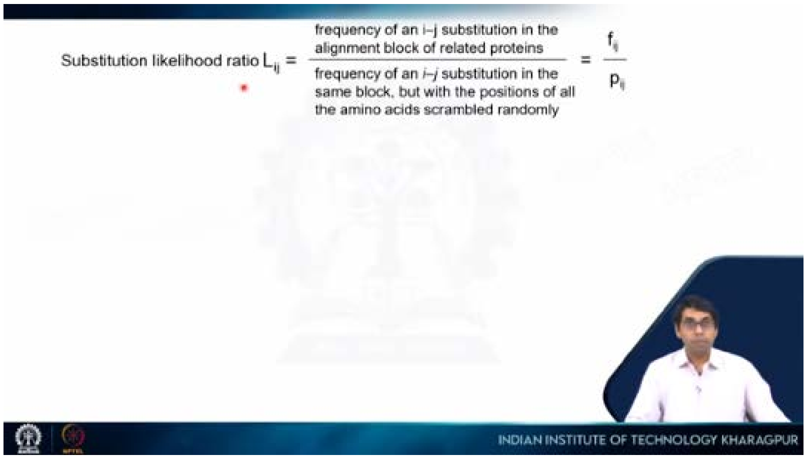


Similarly, I can take this region, so I can go from here to this phenylalanine up to this. I can take this block; there are no insertions or deletions. So let us say I take that. So it will look like this. Now we have to look; we want to determine the statistics for pairwise substitution. So let us say I want to look at the rate of substitution of phenylalanine with leucine or vice versa, leucine with phenylalanine. So I will look at all these places. So there is no leucine or phenylalanine here. No leucine or phenylalanine here.



myoglobin	GHGQDILIRLFKSHPETLEKFDRF
Hb α	EYGAEALERMFLSFPTTKTYFPHF
Hb β	EVGGEALGRLLVVYPYTQRFFESF
erythrocruorin	GDPVGILYAVFKADPSIMAKFTQF
clam Hb	GNGVALMTTLFADNQETIGYFKRL
worm Hb	TSGVDILVKYFTSTPAAQEFFPKF
leghemoglobin	QYSVVGYTSILEKAPAAKDLFSFL
Glycera Hb	EYGAEALERMFLSFPTTKTYFPHF

So in this place, I do see phenylalanine and leucine. Here, I do see leucine, but it is not substituted by phenylalanine. In this case, I do see phenylalanine, but there is no leucine. So here, I have leucine but no phenylalanine. So I can take only those columns, which have both leucine and phenylalanine. That is why I can get the pairwise substitution. For example, if I look at this, there are other residues, but I do have two phenylalanines and one leucine. In this case, everything is phenylalanine, no leucine, so I cannot take this. So, these are the three columns that I will take, which have both phenylalanine and leucine and from here, I will calculate the frequency of substitution of these two amino acids and then compare it with random chance to see whether it is more or less than that. So the substitution likelihood ratio is that L_{ij} are the two amino acids. So it will be phenylalanine and leucine.



$$\text{Substitution likelihood ratio } L_{ij} = \frac{\text{frequency of an } i\text{–}j \text{ substitution in the alignment block of related proteins}}{\text{frequency of an } i\text{–}j \text{ substitution in the same block, but with the positions of all the amino acids scrambled randomly}} = \frac{f_{ij}}{p_{ij}}$$

So, it is given as the ratio of the frequency of an ij substitution, substituting L with f or f with L in the alignment block of related proteins. So, this is my alignment block. Frequency of this is divided by the frequency of this ij substitution in the same block, but with the positions of all the amino acids scrambled randomly. If I take all these amino acid sequences and then scramble them randomly, and then I calculate the L and f substitution.

So, that will give me the denominator. So, these are the two numbers. How do I get this f, the frequency of substitution in this alignment block? So, I will do this by counting the number of times I-type and J-type occur in the same column. So, how many times phenylalanine and leucine occur in this column?

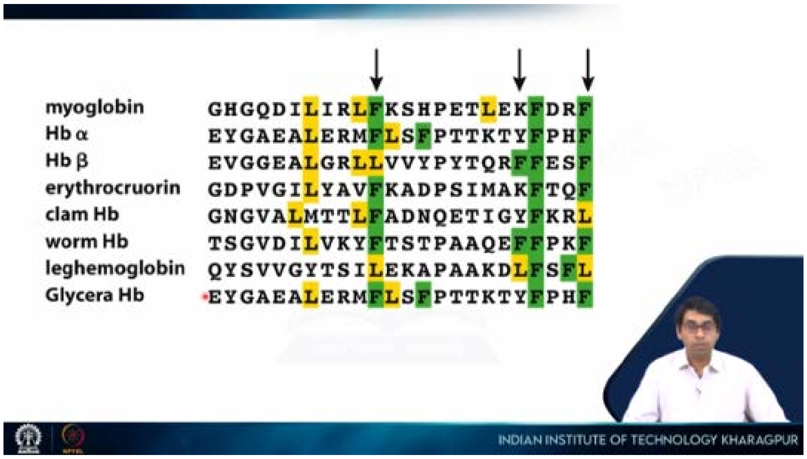
Then the total number of amino acid pairs in the sequence block. This P_{ij}, I will calculate like this. So, I will calculate it as 2 times P_i times P_j. So, what is this P_i or P_j? This I will calculate as the number of occurrences of the i^{th} amino acid divided by the number of positions in the alignment.

So, how many times the amino acid shows up is divided by the total number of positions in the alignment. So, let us look at the substitution of phenylalanine with leucine or leucine with phenylalanine. So, let us calculate this first. P_L is 18 divided by 8 times 24.

So, there are 8 rows and 24 columns. So, that is the total number of amino acids there. How many times does leucine show up if I count here? You will see it shows up 18 times. What about phenylalanine? It shows up 25 times. So, these numbers are given here, 18 divided by 8 times 24 and 25 divided by 8 times 24. So, the denominator is exactly the same and you get these two numbers. So those two multiplied by two, you get this P_{LF}, which is the denominator. Now we have to calculate the numerator, which is this f. Number of times the i^{th} type and the j-th type occur in the same column. So if I look at this column, 1, 2, 3, 4, 5, 6, there are 6 phenylalanines and 2 leucines. So if I calculate the number of pairs, 6 times 2, that will be 12.

Similarly, here, 6 times 2, that will be 12 and here, 2 phenylalanines and 1 leucine. So, this is one pair and this is another pair so, 12 plus 12 plus 2. That is what I get. 12 plus 2 plus 12, that's the numerator. Now, I have to scramble the sequence. So, what is the total number of scrambles? If I do that, what is the total number of pairs that are present? So the total number of pairs present will be 8 rows and 24 columns. So, 8 rows, now, in each column, there are 8. So if I take this, this will be one pair. This will be another pair.

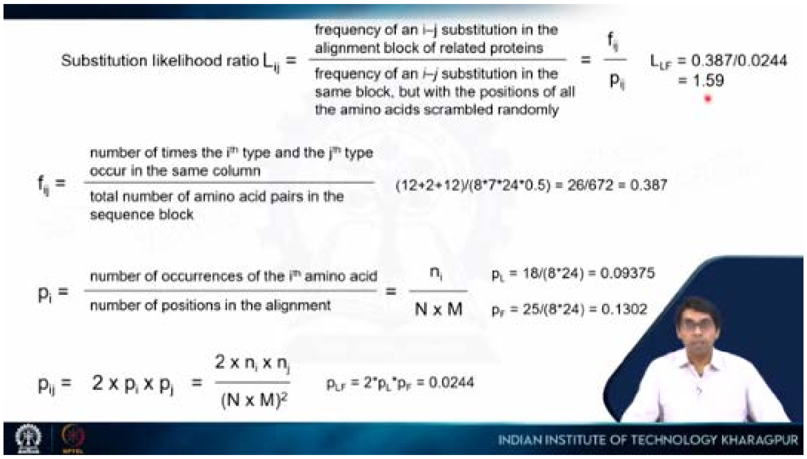
This will be another pair like that. So, 8 times 7, which is 56, will be the number of pairs in each column, and 56 times 24 will be the total number of pairs in this whole block. Now, I can calculate QE and EQ, but they are the exact same pair. So, I am counting them twice. So, I have to divide it by 2, and that is exactly what is done here. 8 times 7, that is 56 so, that is the number of pairs in each column, and you divide it by 2, and then multiply by the total number of columns. So, that becomes 672. So, 26 by 672, that is 0.387.



myoglobin	GHGQDILIRLFKSHPETLEKFDRF
Hb α	EYGAEALERMFLSFPTTKTYFPHF
Hb β	EVGGEALGRLLVVYPYTQRFFESF
erythrocruorin	GDPVGILYAVFKADPSIMAKFTQF
clam Hb	GNGVALMTTLFADNQETIGYFKRL
worm Hb	TSGVDILVKYFTSTPAAQEFFPKF
leghemoglobin	QYSVVGYTSILEKAPAAKDLFSFL
Glycera Hb	EYGAEALERMFLSFPTTKTYFPHF

INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

So, this is the frequency which is the numerator, and this is the frequency, which is the denominator. So, if you take the ratio of these two, you get this LF, which is 1.59, which is greater than 1 and if we take the log to the base 2 of this, multiply it by 2, then we will get S. So, it turns out that when we align two sequences. If the sequence is significantly long, more than, let us say 100 residues and we end up with 10 or 20% similarity. The chances of these two sequences being structurally similar is very likely because you have two sequences which are 100 amino acids long, and their sequence similarity is 20%.



$$\text{Substitution likelihood ratio } L_{ij} = \frac{\text{frequency of an i–j substitution in the alignment block of related proteins}}{\text{frequency of an } i\text{–}j \text{ substitution in the same block, but with the positions of all the amino acids scrambled randomly}} = \frac{f_{ij}}{p_{ij}}$$

$L_{LF} = 0.387/0.0244$
$= 1.59$

$$f_{ij} = \frac{\text{number of times the } i^{th} \text{ type and the } j^{th} \text{ type occur in the same column}}{\text{total number of amino acid pairs in the sequence block}}$$

(12+2+12)/(8*7*24*0.5) = 26/672 = 0.387

$$p_i = \frac{\text{number of occurrences of the } i^{th} \text{ amino acid}}{\text{number of positions in the alignment}} = \frac{n_i}{N \times M}$$

$p_L = 18/(8*24) = 0.09375$

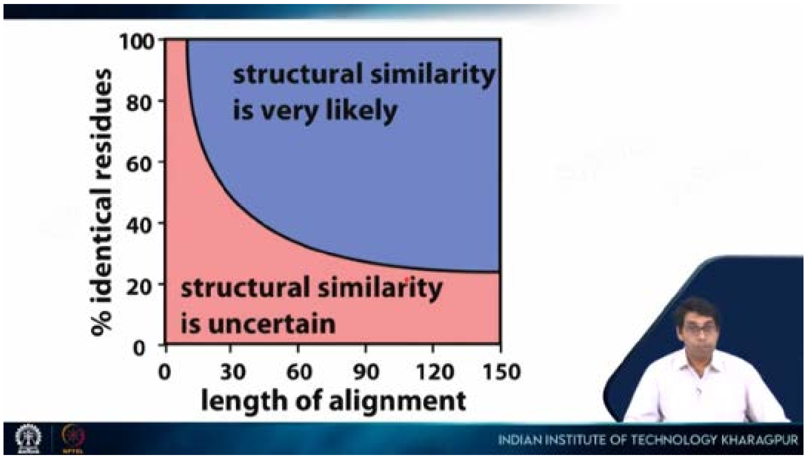
$p_F = 25/(8*24) = 0.1302$

$$p_{ij} = 2 \times p_i \times p_j = \frac{2 \times n_i \times n_j}{(N \times M)^2}$$

$p_{LF} = 2*p_L*p_F = 0.0244$

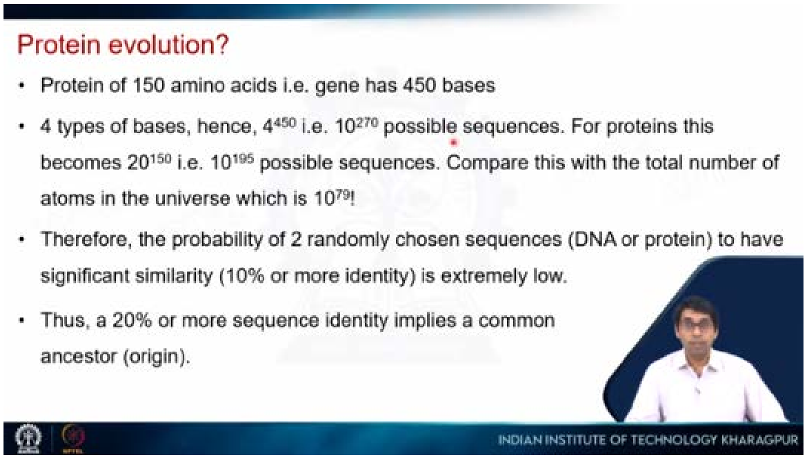
INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

That happening by random chance is very, very, very low. So, it means that they most likely are going to have very similar structures. So, this is the basis of something called homology modeling that people do for newly identified sequences. They try to predict the structure for this new sequence based on the structures of proteins whose structures are already known. So we can do this type of calculation to show that when two sequences have 20% or more identity, then the chance of them being different is very, very low.



So, let us consider a protein with 150 amino acids. So, if you consider its gene, it will be 450 bases. Now, there are four types of bases. So, what are all the possible sequences which will have 450 bases?

That will be 4^{450}, which are 10^{270} possible sequences. That's a huge number. If we just consider the amino acids, 150 amino acids then a protein or how many different 150 amino acid sequences we can generate. That will be 20^{150}, which are 10^{195} possible sequences. Now, compare this 10^{195} numbers with the total estimated number of atoms in the universe, which are 10^{79}. So you can imagine that this is a huge number. Now, if two random sequences are taken and they have 10% or more identity, then that identity arising from random chance is going to be extremely low, which means that these two sequences originated from a common ancestor.

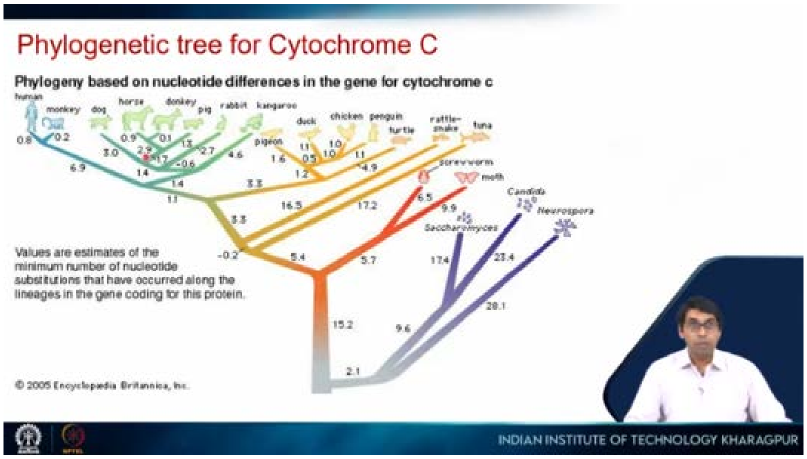


Protein evolution?

- Protein of 150 amino acids i.e. gene has 450 bases
- 4 types of bases, hence, 4^{450} i.e. 10^{270} possible sequences. For proteins this becomes 20^{150} i.e. 10^{195} possible sequences. Compare this with the total number of atoms in the universe which is 10^{79}!
- Therefore, the probability of 2 randomly chosen sequences (DNA or protein) to have significant similarity (10% or more identity) is extremely low.
- Thus, a 20% or more sequence identity implies a common ancestor (origin).

So there was some common protein in the past, and then it diverged, accumulated mutations, and then after several million years, we are looking at these two proteins, and the sequence identity between those two proteins is 10% or 20%. Then you know that these two proteins originated from the same common ancestor. So this is something that we use for creating something called a phylogenetic tree and also measuring time in the past so evolutionary time.

So, a phylogenetic tree will look something like this. This tree has been constructed using a particular protein called Cytochrome C. So we have already seen Cytochrome C in photosynthesis and respiration. So the tree, the way it looks like that is the ancestor at different points in time, this branching happened. Now, anything that is close, they are much related.



Phylogenetic tree for Cytochrome C

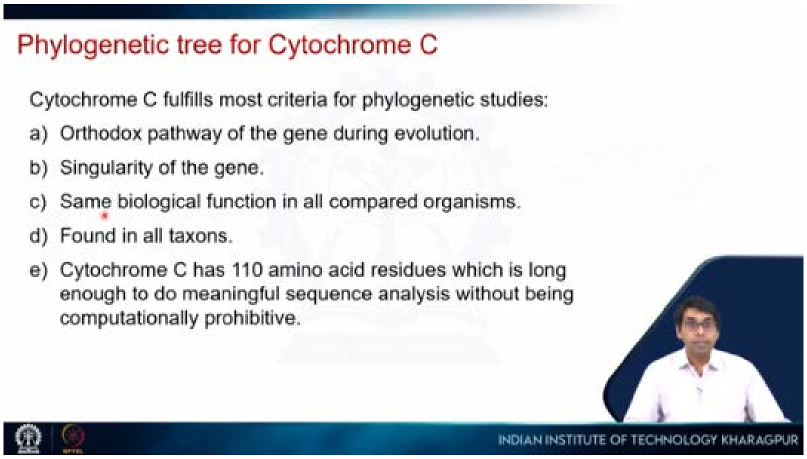
Phylogeny based on nucleotide differences in the gene for cytochrome c

Values are estimates of the minimum number of nucleotide substitutions that have occurred along the lineages in the gene coding for this protein.

© 2005 Encyclopædia Britannica, Inc.

INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

So, they are diversed at this point, this and this diversed much more in the past. So, things which are more separated diversed more and more in the past things which are close together diversed in the near past and these numbers are normally they are some weighted numbers you can also put them in terms of millions of years. So phylogenetic tree are created using proteins which have certain criteria so for example cytochrome c it fulfills most of the criteria that are needed for creating a phylogenetic tree. So it has this orthodox pathway of gene during evolution. So nothing fancy nothing funny happened. So we can trace this protein in different species. Singularity of a gene so there is only one gene for this protein not multiple genes like ribosomes. Same biological function in all compared organisms. It is not doing something different in different organisms. It is found in all these different taxons of life. Cytochrome C is small.



Phylogenetic tree for Cytochrome C

Cytochrome C fulfills most criteria for phylogenetic studies:

a) Orthodox pathway of the gene during evolution.
b) Singularity of the gene.
c) Same biological function in all compared organisms.
d) Found in all taxons.
e) Cytochrome C has 110 amino acid residues which is long enough to do meaningful sequence analysis without being computationally prohibitive.

It has only 110 amino acids, which means that the sequence alignment will not be computationally problematic. So you can actually do that meaningfully. So this type of phylogenetic tree helps us to determine how long back in the past two species diversed. So whenever we talk about this time, it means that we are talking about some clock.

So where is this clock coming from? This clock is coming from the rate of change or rate of evolution or rate of mutations that happen in certain genes. If we know the rate of mutation, if that rate of mutation is not changing much over time, then if you compare two sequences and see how many changes have happened, you can get an estimate of time that it took this many times, this much time to gather this many mutations. So this is something that is referred to as a molecular clock. So the molecular clock is based on the rate of molecular evolution, which can be remarkably constant over time.

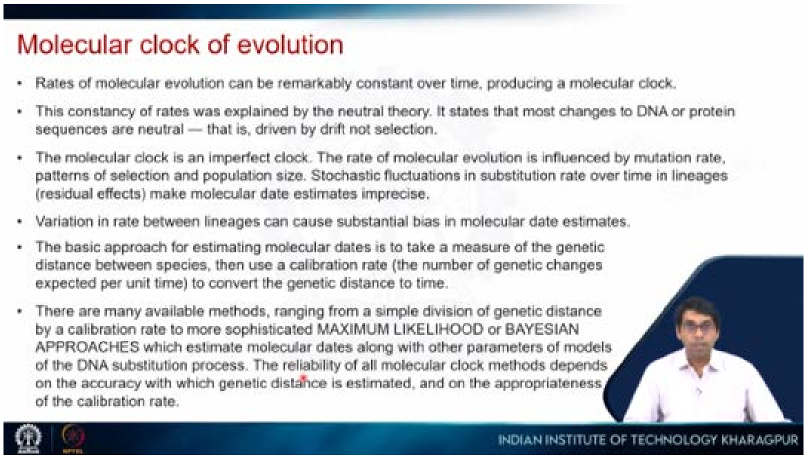
So this is one of the major hypotheses of a molecular clock. This constancy of rate was explained by the neutral theory. So what is a neutral theory? It states that most changes to DNA or protein sequences are neutral. That is, they are driven by drift, not by selection.

Molecular clock of evolution

- Rates of molecular evolution can be remarkably constant over time, producing a molecular clock.

So now there are, of course, problems with this, but it is something that helps us in determining some of these time periods. So, of course, this molecular clock is an imperfect clock because the rate of molecular evolution is influenced by mutation rate, which can be influenced by so many different things, pattern of selection, population size, stochastic fluctuations in this substitution rate. So just by chance, there can be accumulation of more mutations in a small period of time so all these things can make this clock really imperfect and the estimation of the dates imprecise. Variation in rate between different lineages can cause substantial bias in molecular date estimates. For example, if you consider unicellular organisms, evolution or mutation happens faster in unicellular organisms compared to multicellular organisms like us. So the rate of mutation will not be the same, and this variation can create substantial bias in the molecular date estimates that we are doing using this molecular clock. The basic approach for estimating molecular dates is to take a measure of the genetic distances between species, then use a calibration rate to convert the genetic distance to time.

So this is the basic premise. However, of course, as pointed out here, there are several problems with this. So there are many statistical methods by which people do this. So there are sophisticated methods like maximum likelihood or Bayesian approaches, which estimate molecular dates along with other parameters of the model of the DNA substitution process. Now, the reliability of all these molecular clock methods depends on the accuracy with which genetic distance is estimated and on the appropriateness of the calibration rate.

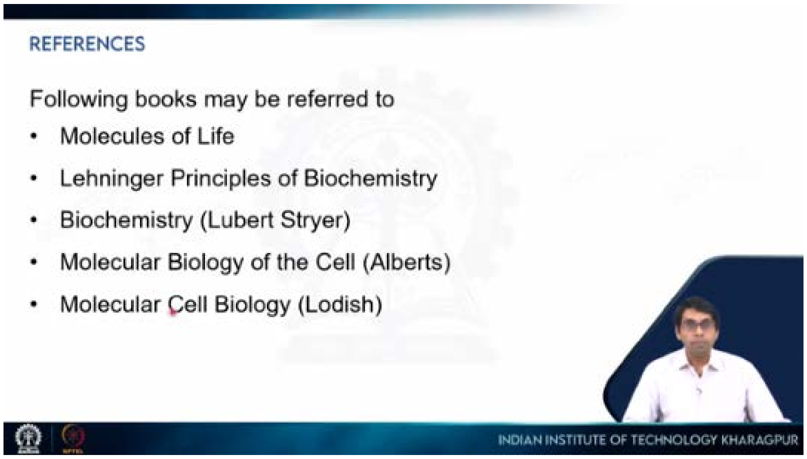


Molecular clock of evolution

- Rates of molecular evolution can be remarkably constant over time, producing a molecular clock.
- This constancy of rates was explained by the neutral theory. It states that most changes to DNA or protein sequences are neutral — that is, driven by drift not selection.
- The molecular clock is an imperfect clock. The rate of molecular evolution is influenced by mutation rate, patterns of selection and population size. Stochastic fluctuations in substitution rate over time in lineages (residual effects) make molecular date estimates imprecise.
- Variation in rate between lineages can cause substantial bias in molecular date estimates.
- The basic approach for estimating molecular dates is to take a measure of the genetic distance between species, then use a calibration rate (the number of genetic changes expected per unit time) to convert the genetic distance to time.
- There are many available methods, ranging from a simple division of genetic distance by a calibration rate to more sophisticated MAXIMUM LIKELIHOOD or BAYESIAN APPROACHES which estimate molecular dates along with other parameters of models of the DNA substitution process. The reliability of all molecular clock methods depends on the accuracy with which genetic distance is estimated, and on the appropriateness of the calibration rate.

However, this is still a challenging problem, but it still gives us some really meaningful and interesting insight into the evolution of these different organisms and species over this period of several billion years.

So, for this lecture, you can follow these books as I have listed here. 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)