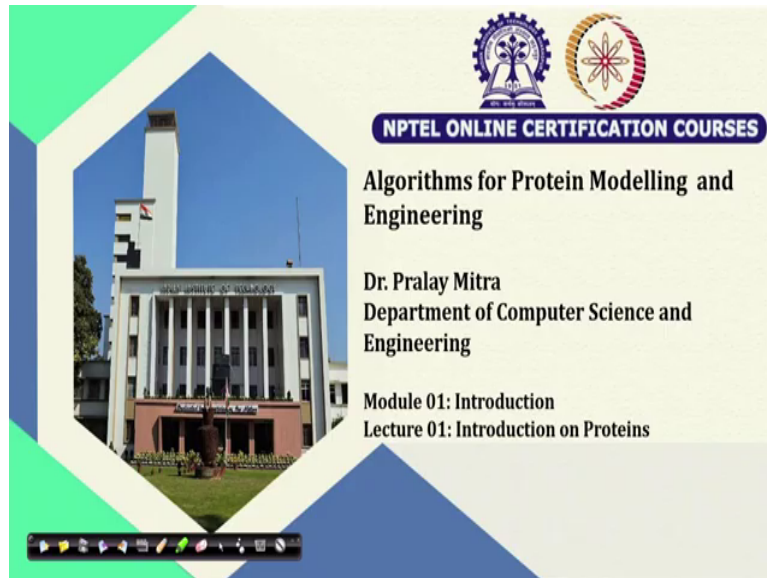


Algorithms for Protein Modelling and Engineering
Professor Pralay Mitra
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur
Lecture: 01
Introduction on Proteins

(Refer Slide Time: 00:16)




Welcome to the course on Algorithms for Protein Modelling and Engineering. We are starting today. Today in the first module, we shall go with an introduction and since it is all about modelling and engineering the protein molecules. So, in this introduction, mostly we shall discuss briefly the protein biomolecules. We shall not go into details about that, but we shall discuss the protein molecule which will be relevant in the context of its modelling and engineering.

And in this course, we are planning to discuss several problems and the challenges which are there in modelling and engineering and designing protein molecules and their relevance in industry, in healthcare, etc. We shall discuss the relevant algorithms in the context of protein modelling and engineering specifically, which are used to solve this particular problem. So, let us start!



(Refer Slide Time: 01:30)

CONCEPTS COVERED

- Proteins
- Essential Amino Acids
- Protein Primary Structure
- Protein Tertiary Structure
- Protein Quaternary Structure




Pralay Mitra





KEYWORDS

- Protein
- Amino Acids
- Protein Sequence
- Protein Structure



Pralay Mitra

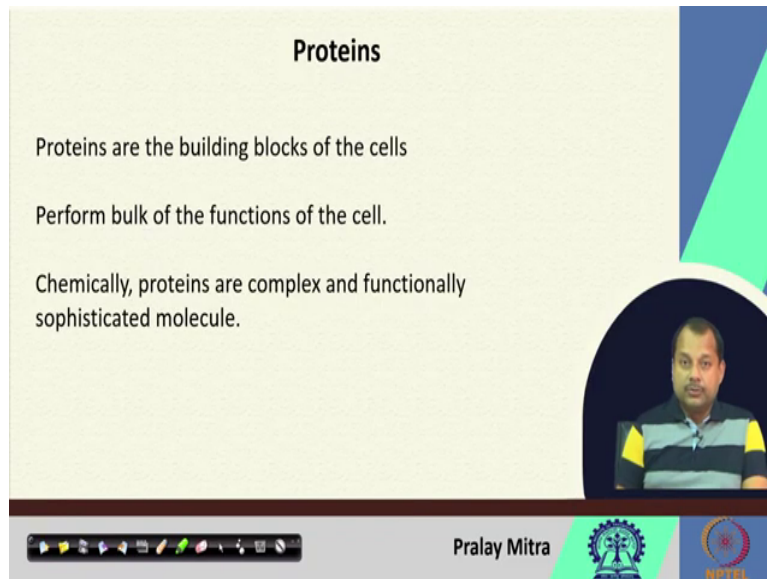


Proteins

Proteins are the building blocks of the cells

Perform bulk of the functions of the cell.

Chemically, proteins are complex and functionally sophisticated molecule.

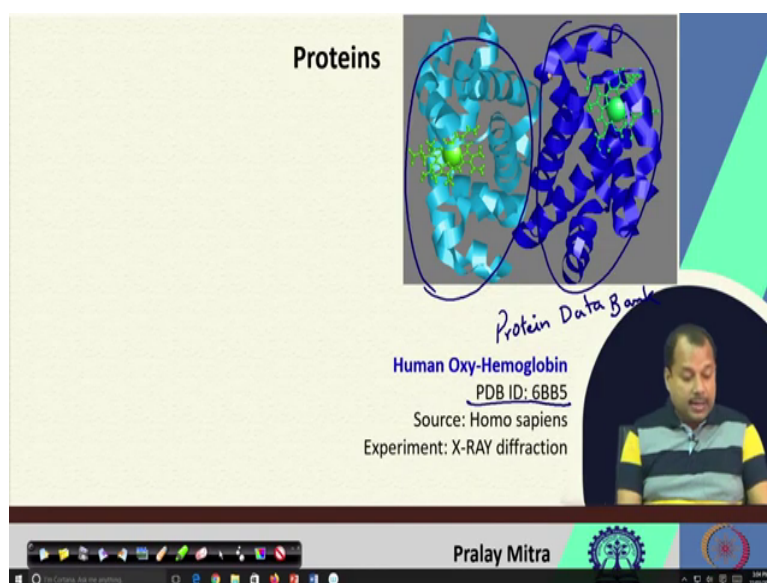


In this lecture, I am planning to cover the proteins and the composition of protein - amino acids, next protein primary structure which is also called the protein sequences, protein tertiary structures, protein quaternary structures and that's it. As keywords, I am providing you protein and amino acids. So, let us start with the protein by looking at the structure of a protein.

Proteins are the building blocks of the cells and it performs bulk of the functions of the cell. It performs almost all the functions of the cell by interacting with other protein molecule or any small molecule or with any ligand or with a DNA or RNA. Chemically, proteins are complex and functionally they are sophisticated molecules. Let us start with an example.

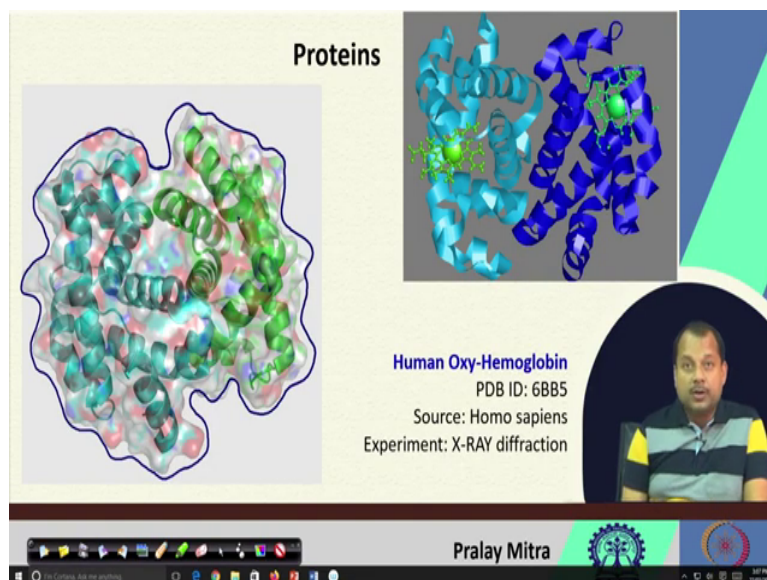
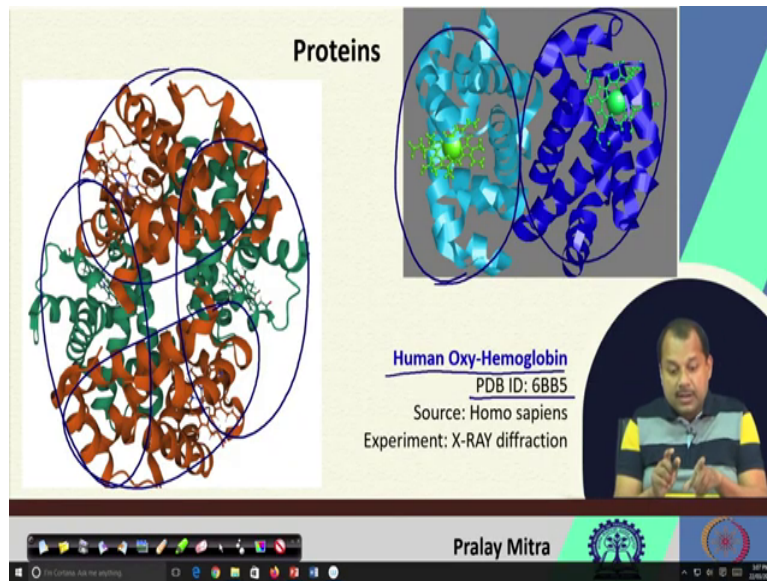
(Refer Slide Time: 02:44)

Proteins



Human Oxy-Hemoglobin
PDB ID: 6BB5
Source: Homo sapiens
Experiment: X-RAY diffraction

Protein Data Bank



Here is an example I am going to show you - the protein of a human Oxy-Haemoglobin. We are aware of the fact that Haemoglobin among all the other functions is the oxygen carrier. It transports the oxygen and the structure I am going to show you is composed of only two chains.

Although, we know that there are four chains, I mean four individual components of a Haemoglobin alpha chain, beta chain, alpha chain, and beta chain. But in the structure which is deposited in the Protein Data Bank (PDB) there are two chains. We shall discuss in detail what is this PDB and what is this PDB ID.

In details we shall discuss all those. For this particular human Oxy-Haemoglobin with PDB ID 6BB5, its source is Homo sapiens (human) and this particular structure has been identified experimentally using the X-ray diffraction technique.

You can see that there are two different components. One is with the light blue colour which is on the left-hand side, another with the blue colour which is on the right-hand side. These two are two separate components or chains. When these two along with a copy of the other two will be combined then the complete structure will be created which is on the left-hand side. Thus, these two, along with other two, I mean, four different subunits, will constitute a complete quadratic structure.

Finally, combining all those four will have some functions. As you can understand that if there is only this blue colour on the right-hand side, or say this blue colour on the right-hand side this part, so, this part is one connected part and that do not have a function by itself. It will have function only when it will interact with its left-hand side that is the sky-blue part.

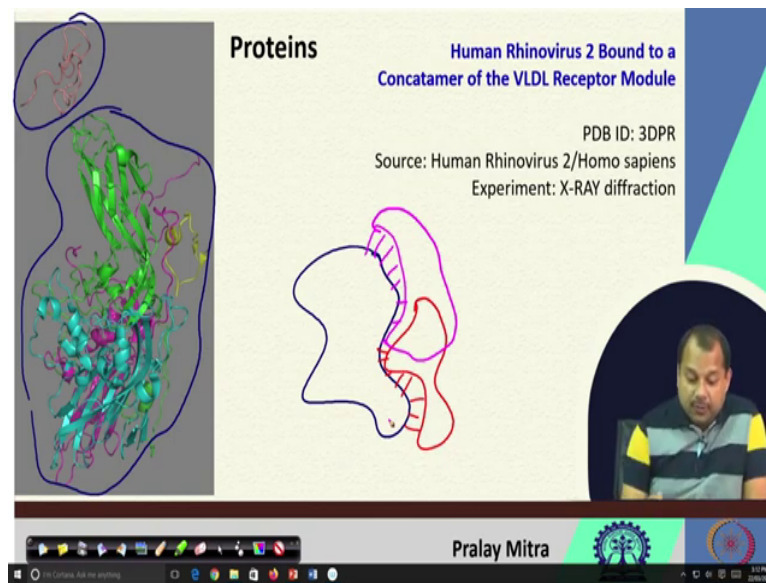
Assuming these two will interact, they may have some functions, but all the functions of the Haemoglobin will occur when these two structures will again have another two structures and that will have something like this. On the left-hand side, you can see this is one part, this is another part, on the back of this there is one part and another is this one. So, these four subunits when they interact then only it will have the functions of the Oxy-Haemoglobin.

With this structure, PDB ID is 6BB5 and X-ray diffraction method is used to get the structure. Using X-ray diffraction, we got only the two structures, but the rest of the two we have to get by some software or by some computational techniques.

What is the software? How can I get that one? Those challenges and the algorithm behind that one we shall discuss throughout this class. Now, if you look at these, generally this is called the cartoon structure. There is software that also I shall discuss as part of the introduction section. But if you think that Haemoglobin is a biomolecule, then it will have some surface.

And regarding the surface, if you wish to look at then on the left-hand side you will see that the surface structure. This is the surface of the biomolecule and inside that one this cartoon structure that I am showing you. This is a protein, but not only has this, for a good reason like the Oxy-Haemoglobin; this particular protein also had some other functions.

(Refer Slide Time: 08:00)



Let us consider another protein - Human Rhinovirus 2 bound to a Concatamer of the VLDL Receptor Module, its PDB ID is 3DPR, its source is Human Rhinovirus 2 and Homo sapiens and technique is X-ray diffraction again. There are a lot of experimental techniques, X-ray diffraction is one of them, another is nuclear magnetic resonance (NMR), then is cryo-EM.

Despite many techniques, most widely used technique is X-ray diffraction and NMR. And if I look at the protein data bank, then I will see that most of the structures which are available in protein data bank is using X-ray diffraction. For this one the available structure looks like this. Here on top you see that this is one protein and this is another protein.

These two proteins interact, I mean, they go to close proximity with each other and perform some function. If you remember, at the beginning I mentioned protein performs almost all the functions of our body. If it is true, then when I am performing something, it is because of the functionality of some protein. Because of some reason, when that particular functionality is blocked. For example, one reason may be that protein is engineered or maybe it is designed or during the protein synthesis process, some part of the protein is not synthesized properly, I mean, it got mutated or maybe because of some attack by some, say, virus or say from bacteria. And because of that one, it does not allow our human body proteins to perform their own job, then that is also another interaction but that is not for the benefit of our body. Because of that interaction, the proper functionality of the protein is blocked.

This can also be considered as a situation for human rhinovirus. When we got infected, we have some disease condition. the disease condition is because some of the proteins are not

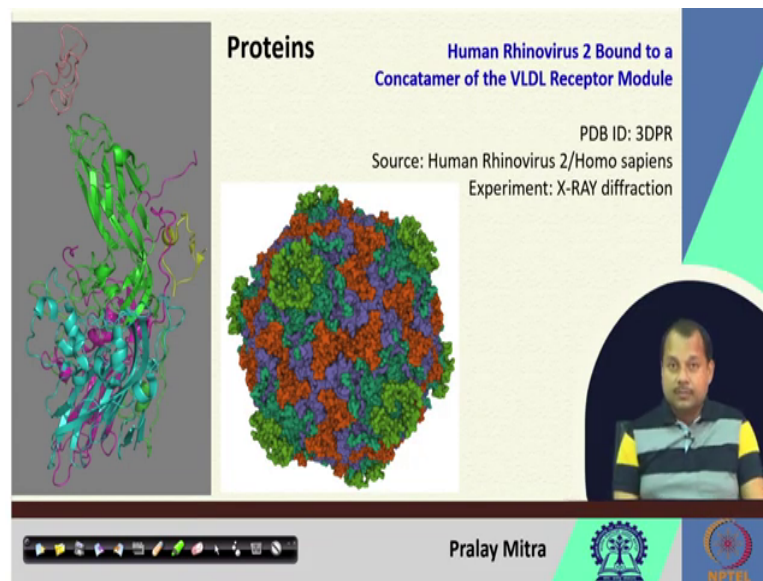
able to do it's a proper function because some disease protein prompts another external or foreign body has come in our system or our body which are not allowing us to perform its function.

For example, if I say draw one protein structure like this human protein and it has some function which allows it to interact with this particular protein and then it will have some function. I am just drawing the surface of the protein so that it can have another function because of the interaction it has at this region.

Now, if I assume that suddenly some other, so this pink colour protein has come and that blocks this part. You see, if this pink colour protein comes into the system and interact with a blue colour protein at this position, then it will not allow the red colour protein to interact and if the affinity of this pink colour with the blue colour, I mean the interaction energy between blue colour and pink colour is higher compared to the red colour then it will not allow the red colour to interact.

If I say that blue and red interaction is essential for our normal body function, then definitely blue and pink, when it will be interacting with each other, then that particular function will be affected. What are the conditions or what are the regions? If it is occluded by other proteins then red will not able to interact. So, is it possible that computationally we can identify that one? Those problems we are going to discuss. This is another structure of the protein.

(Refer Slide Time: 12:40)



Looking at this protein structure (PDB ID: 3DPR) that we got using X-ray diffraction technique, we find that this is not the complete structure of the protein.

It is only part or fraction of the protein, may be because of limitations by the experimental technique like, the crystallization ability of the full protein structure, size of the molecule - if it is a very large size molecule then crystallizing is not that much easy.

Considering all those facts, it is not always possible that the functional form of a protein or the protein which is in action will be crystallized properly and we will get their structure, maybe it is not possible. If not, then only part of the information will be provided along with the information like, what was the crystallization state, etcetera. Based upon that is it possible that we can generate the total structure using the fact that protein always prefers symmetry?

Yes, it is possible. Here I am showing you one such structure. This structure is in the complete functional form. As of now, what we have discussed is that protein is one biomolecule; it performs almost all the functions in your body or any living organism, when it interacts with another protein molecule or one DNA or RNA or say, any small molecule or another biomolecule.

To know the functionality of the protein, there are experimental techniques, but most of the time experimental techniques are limited by the fact that it is time-consuming, it is resource consuming. So, is it possible that there is guide for those experimental techniques, so that not all the possibilities, but only a subset of the possibilities will be given to the experimentalist or the clinicians for exploring instead of all the different possibilities?

We will work in that area. I mean in that particular area, we will explore the computational problems which are there, what the challenges of those problems are and what are algorithms we can deal with.

(Refer Slide Time: 15:35)

Essential Amino Acids					
Amino Acid	3-letter code	1-letter code	Amino Acid	3-letter code	1-letter code
Alanine	Ala	A	Methionine	Met	M
Cysteine	Cys	C	Asparagine	Asn	N
Aspartic Acid	Asp	D	Proline	Pro	P
Glutamic Acid	Glu	E	Glutamine	Gln	Q
Phenylalanine	Phe	F	Arginine	Arg	R
Glycine	Gly	G	Serine	Ser	S
Histidine	His	H	Threonine	Thr	T
Isoleucine	Ile	I	Valine	Val	V
Lysine	Lys	K	Tryptophan	Trp	W
Leucine	Leu	L	Tyrosine	Tyr	Y

Let us start with the building blocks of the protein. Proteins are composed of amino acids. There are n numbers of amino acids, but we shall restrict ourselves only within 20 amino acids. Those 20 amino acids are listed here: alanine, cysteine, aspartic acid, glutamic acid, phenylalanine, glycine, histidine, leucine, isoleucine, lysine, methionine, asparagine, proline, glutamine, arginine, serine, threonine, valine, tryptophan, and tyrosine.

Corresponding to each amino acid, there is a three-letter code. The first column if I say gives one vertical line from here. So, on the left-hand side, it indicates one table, and on the right-hand side, it is another table. For the brevity of the space, I concatenated them together. First column indicates the name of the amino acids. They can be represented using three-letter codes.

For example, alanine can be represented by Ala, cysteine by Cys, aspartic acid by Asp, glutamic acid by Glu, etc. Mostly you can see that the first three letters are being used for this purpose but not for say isoleucine, not for say asparagine, not for glutamine, not for tryptophan.

But for the rest except these four, you can see that the first three letters of the amino acids are used as the three-letter code. But even the simple way is that corresponding to each amino

acid, there is one single letter code. Now, there are 20 amino acids and in the English alphabet, there are 26 characters.

So, English alphabet characters are used for this purpose. For example, alanine is represented by A, cysteine is represented by C, aspartic acid is represented by D, glutamic acid is represented by E, phenylalanine is represented by F, glycine is represented by G, like that way it will go. To remember this one, I can tell you that if you take all the English alphabets who are missing. So, I am writing the missing part in red here.

B is missing, then J is missing, O is missing, U is missing, X is missing, and Z is missing. So, 6 missing, 20 amino acids, and hence there are in total 26. For this U, X, etcetera some other amino acids are being used, but we will not consider those amino acids. We will restrict only to these 20 essential amino acids, that is A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y. Now, these amino acids form the basis of the protein. When I say protein, at the core of the protein there are these amino acids.

(Refer Slide Time: 19:35)

Protein Sequence

His Ser His Val Lys Gly Ala Lys Ala

H S H V K G A K A

.....HSHVKGAKA.....

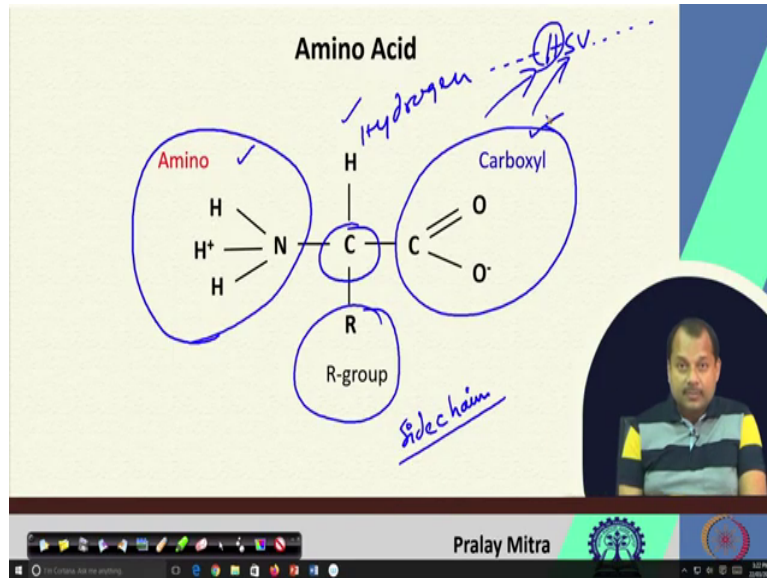
Pralay Mitra

Now, if I place these amino acids one after another, then I shall get a protein sequence, which is also called the protein primary structure that I will come later. Now, in this protein sequence, this is three-letter code that you can see, this His, Serine, histidine, valine, lysine, glycine, alanine, lysine, alanine, all these are three-letter codes.

If I translate that to single letter codes then this His will be H, Serine will be S, His will be H, Val will be V, Lys will be K, GLY will be G, ALA will be A, LYS will be K, ALA will be A. So, H, S, H, V, K, G, A, K, A, dot, dot, dot, on the left-hand side as well as on the right-hand

side indicates that it is only a part of the whole protein sequence that I am considering but it has a long stretch. Thus, this is my protein sequence. These amino acid molecules placed one after another are form one protein molecule.

(Refer Slide Time: 20:48)

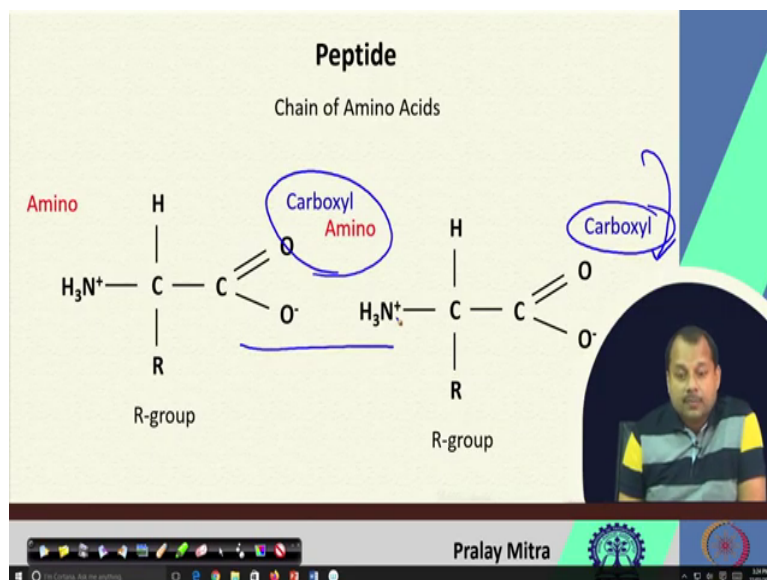
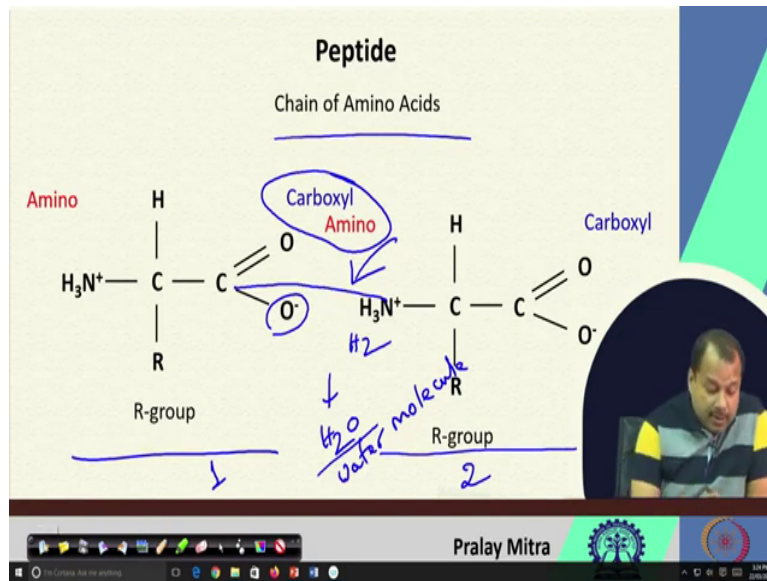


Those 20 amino acids are molecules. If I look at the structure of that molecule, then at the core of that molecule, you will find one carbon atom. It is achiral carbon at the centre position. This carbon is connected on one side by hydrogen, another side is by carboxylic group, and another side is with the amino group. And this amino group, carboxylic group, and the hydrogen are common to all the 20 amino acids. Thus it is a piece of good news, and it is easy to remember. Only difference between those 20 amino acids is in the R part. You can consider as a fourth hand. Sometimes we also call this as side chain. We will discuss those in detail later in the relevant context.

Now, this is one amino acid and if in this case one particular sidechain is placed, then this will represent one of the 20 amino acids. If it is true, then you can consider in the previous slide where I have shown you, H dot dot dot dot then H S V dot dot dot dot. Corresponding to H the histidine, if I replace this sidechain with its proper sidechain of the histidine, then I can place that histidine here.

Corresponding to serine, if I change the side chain, I can place it here that way if I keep on placing them and then their connections because of the placement of the side, it will form one protein sequence. How? That I am coming.

(Refer Slide Time: 23:05)



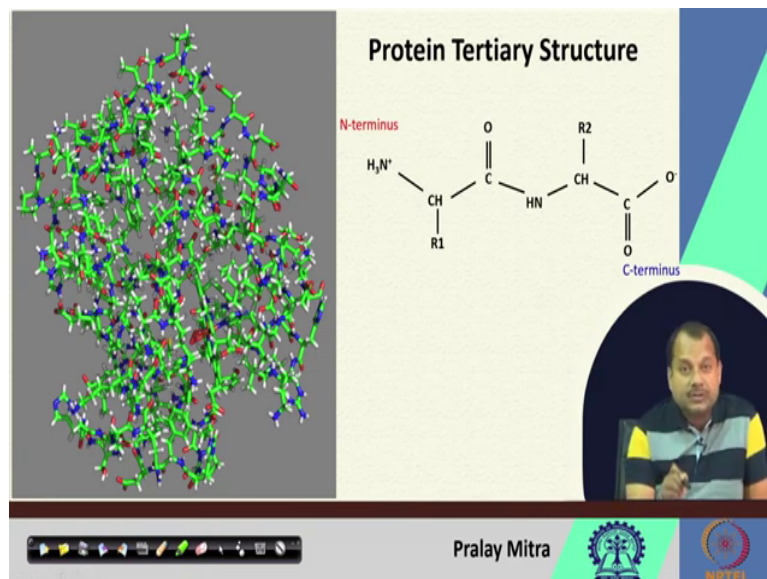
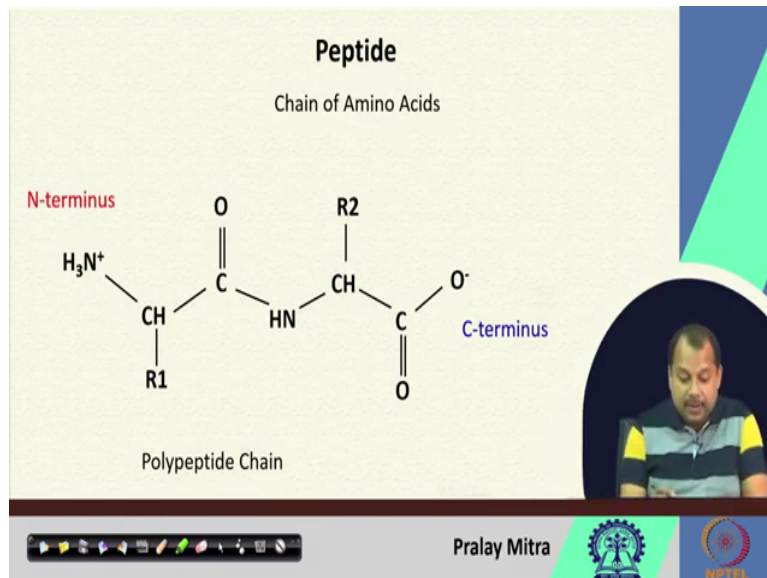
Here you see that on the left-hand side, I placed one amino acid and on the right-hand side, I placed another amino acid, one and two. Now, this carboxyl group and amino group will get one oxygen atom from here, two hydrogens from here will come out in the form of H₂O.

So, one water molecule will go out. You know this is water molecule. This water molecule will go out and then one connection will be established between this C and this N - one covalent bond, and because of that one of these two amino acids, one and two. Based upon its side chain its particular existence will appear whether it is histidine, serine or any other amino acid - those will be connected. The amino acids will be bonded through their covalent bond. This is called a peptide which is a chain of amino acids.

If I keep on going so carboxyl group and this amino group will combine and they will form the peptide bond connection between these two. In the long run, after this one, if I plus one

more amino acid here then this carboxyl group will go, and that way you will see on the leftmost side that is one amino group on the rightmost side that is another carboxyl group. And in between these two, the peptide bonds are there.

(Refer Slide Time: 25:05)



Likewise, what I shall get is a list of amino acids placed one after another and that is the chain of amino acids. Now, you think about the situation that a protein composed of amino acids, those amino acids in the sequence format is written as one of the alphabets taken from the English alphabets, so, A, C, D, E, F like that way - you place and that is the sequence.

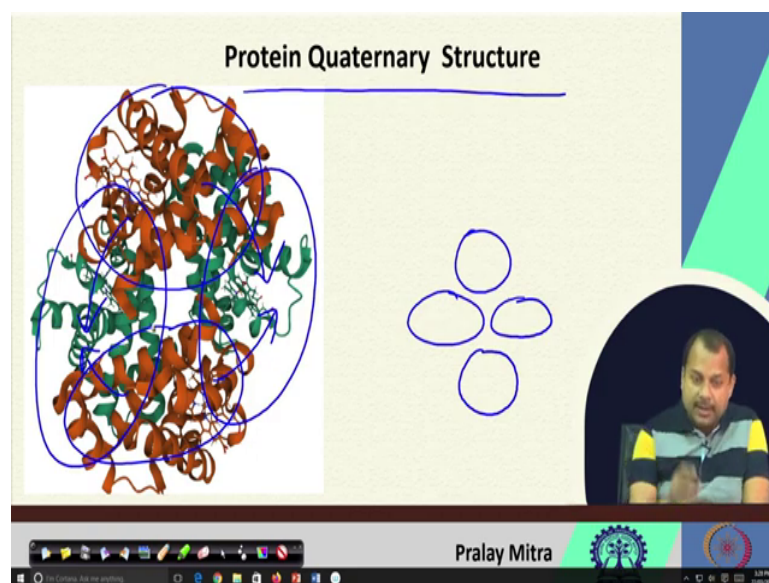
Now, in that sequence each character is replaced by one such molecule and between two consecutive molecules, there is a peptide bond. You think about the situation like - those atoms will be placed one after another to form a string.

That string may not be a stable one. If you allow the string to move freely, then it will fold and it will take a structure which is on the left-hand side. This particular structure if you see - you will see that there are different colours white, red, blue, green - representing whether the atom is a carbon or nitrogen or oxygen or say hydrogen, but they will take some structure like this.

This structure you already have seen, at the beginning using the secondary structure element or those coils, etcetera the cartoon diagram, I have shown that to you. Now, this is the stick presentation using PyMol. That also I shall discuss later, but here each say white indicates one atom, each red indicates one atom, the green indicates the junction of this green point indicates one atom.

Atoms are there in the structure. This particular structure is called a protein tertiary structure. N-terminus and C-terminus are there. It is difficult to identify that one right now in this figure. After that one, all the atoms are there with its coordinate or position. So, that position is important that we will discuss later.

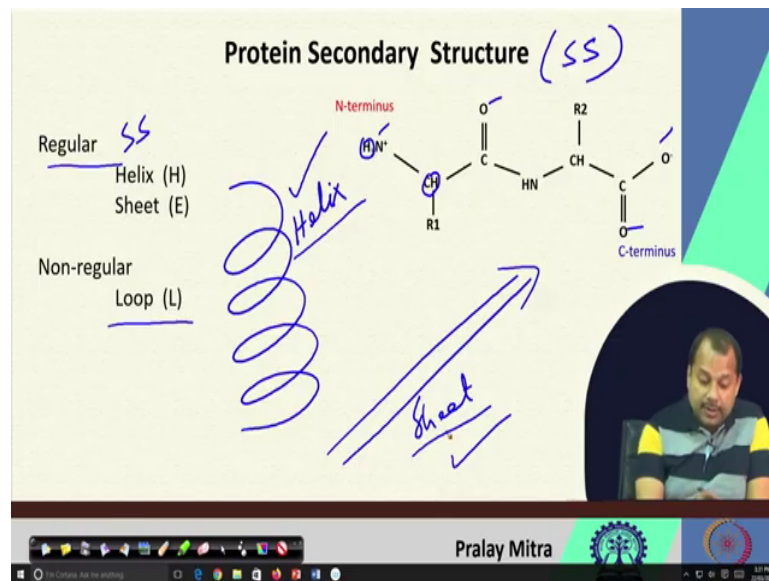
(Refer Slide Time: 27:45)



Now, as I mentioned, when say one protein molecule is there, then that protein molecule or one single connected part cannot have a function by itself. To have its function, it must interact with somebody that somebody may be another protein molecule or say another small molecule or any other biomolecule. If one such molecule interacts with another molecule or a set of another molecule, then what I shall get is a quaternary structure.

In this case, there are four molecules 1, 2, 3, 4. Although, you cannot see, but their structure is like this. This is called a D₂ symmetry that we shall discuss later in the context of the symmetry. This will form the quaternary structure where this molecule will interact possibly with this, this will interact with this, this will interact, this will interact with this, this will interact with this, because of these interactions some function may happen, will happen not may, definitely will happen. This is the quaternary structure.

(Refer Slide Time: 29:08)



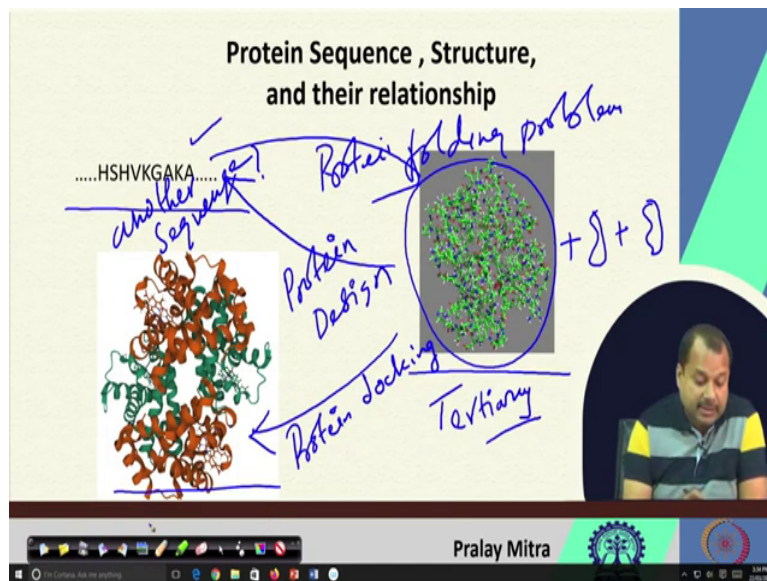
Now, one thing I am missing because we do not have many purposes for this one. It is the secondary structure. As, I mentioned - primary, tertiary, and quaternary. What about the secondary structure? Secondary structure indicates that when there is a pattern - what pattern? You remember that tertiary structure I have shown you.

In this structure, it is taking some fold or shape while taking some shape then if the hydrogen molecule presents here and present here forms some hydrogen bond with some nitrogen or oxygen present somewhere say here or say here or say here or say here. If that happens then it may take some regularity in the structure, something like this may be one regularity or maybe this is one regularity.

If that regularity is there then we call that as a regular secondary structure, secondary structure in short SS, so regular SS. Otherwise, if it is not that then we will call that as a non-regular secondary structure. Mostly we shall restrict ourselves to helix, you see this is kind of a helix and this is called as the sheet. So, there are different variations of the helix and sheet based upon the pattern of the hydrogen bonds, etcetera.

We shall not go into details of that one if it is not required for us. If it is required, then in the context, it will be discussed but grossly it is helix and sheet and for the rest also regarding the non-regulatory there are different names like turn, bridge, and etcetera. We shall not go into details of those. What we shall say if it is not if it is like this then it is a helix, if it is not like this but like this then it is a sheet, if it is not like this or this then it is a loop. That is my definition of non-regularity for our course purpose only.

(Refer Slide Time: 31:26)



To summarize what we have discussed is the protein sequence which is called a primary structure, protein three-dimensional structure which is called a protein quaternary structure, sorry protein tertiary structure, and protein quaternary structure. Regarding the protein sequence, when I placed amino acids one after another then the amino acids single-letter character will give me what is the protein sequence.

If each amino acid is replaced by its corresponding molecule that overview structure I have shown you by its corresponding molecule and the molecule contains the atoms, atom has its own 3D coordinate positions. And if a particular sequence is allowed to fold and take some space in three-dimensional space then that I am calling as a protein tertiary structure that is here - tertiary structure.

When more than one tertiary structure interacts with each other, then I shall have the quaternary structure. If this protein sequence is given to you, is it possible that you can have one computational algorithm so that you can tell what will be the structure of this without the need for an experimental technique? That is called the protein folding problem. If you have a number of these three-dimensional tertiary structures, is it possible that you can give me this quaternary structure? That is called the protein docking problem.

Is it possible that one such structure is given to you, if it is, then what is the sequence, but is it possible that you can come up with another sequence that is not this sequence but that particular sequence will also fold to this structure? That means, given this as an input I wish to have another sequence not this one, another sequence as output - is it possible? That is

called the protein design problem. Those things we shall discuss one after another along with other problems which will arise in the context of this protein modelling and engineering.

(Refer Slide Time: 34:22)

REFERENCES

- The Anatomy and Taxonomy of Protein Structure
by Jane S. Richardson
- Introduction to Protein Structure
by Carl Ivar Branden, John Tooze

Pralay Mitra

IIT Bombay NPTEL

Here are some of the references that you may use. Thank you very much.