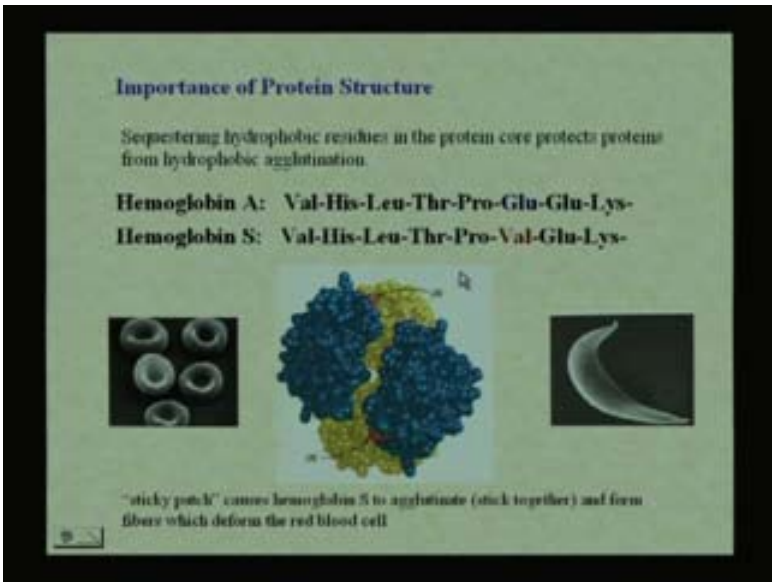
Biochemistry - I
Prof. S. Dasgupta
Department of Chemistry
Indian Institute of Technology, Kharagpur
Lecture –5
Protein Structure III

Welcome, this is lecture number five on Protein structure. Previously we discussed about the importance of protein structure.

(Refer Slide Time 0:56 min)



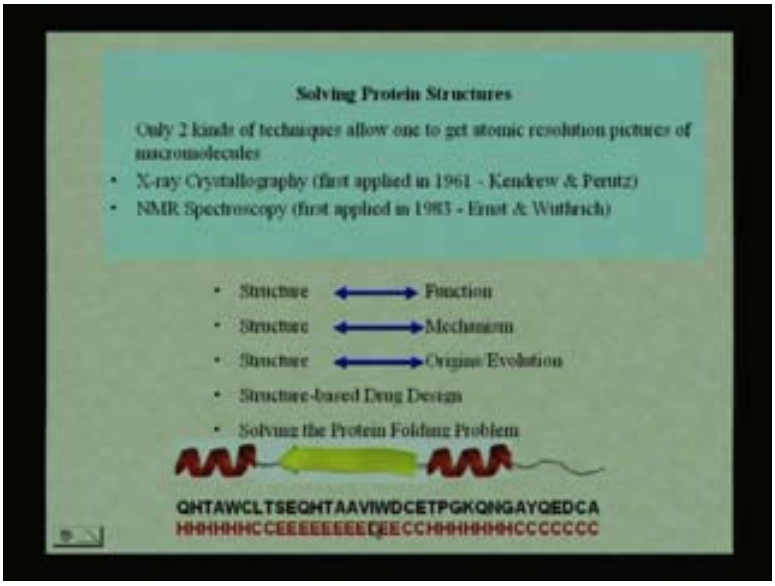
So here you have difference in the folding of the protein because of a specific genetic mutation. An amino acid mutation is where we have Glutamic acid goes to Valine. Now we have a hydrophobic residue on the surface because of this mutation. This is entirely different from the Glutamic acid because of this property. We have these sequestering together or sticking together to form a fiber which will gives disease.

The folding of the protein is extremely important and there are certain forces associated with a folding of a protein. Now we will understand more of how we can actually get to solving protein structures because we already know that we have this amino acid sequence, we know that we can get a secondary structure out of it from that we can get a tertiary structure finally to a quaternary structure.

So, if we want to look at the process to get a protein structure, there are only two kinds of techniques are available to get atomic resolution pictures of macromolecules because you want to know exactly where the atoms are. These will give the information about the protein structure. It is very difficult to pin point where the atoms are if you were to take a snap shot of an extremely large system which is a macro molecule. Now there are two techniques available to get atomic resolution pictures. One is X-ray Crystallography and the other is NMR Spectroscopy.

X-ray Crystallography is the only method that is very useful in solving a protein structure. But NMR Spectroscopy is actually very fast catching up. Now the difficulty of the X-ray Crystallography is getting a crystal because you should have a crystal of the protein to do the Crystallography study. But getting a protein crystal is very difficult because of its large size. It actually either forms a powder or just sticks together and does not form crystal at all. And that is why it is very difficult to get protein crystals. So you cannot do Crystallography without having a protein crystal. Due to this reason people will do a protein structure prediction because it is easy to get the sequence of the protein. We will be studying a method to find out the amino acid sequence of the protein very easily.

(Refer Slide Time 3:55 min)

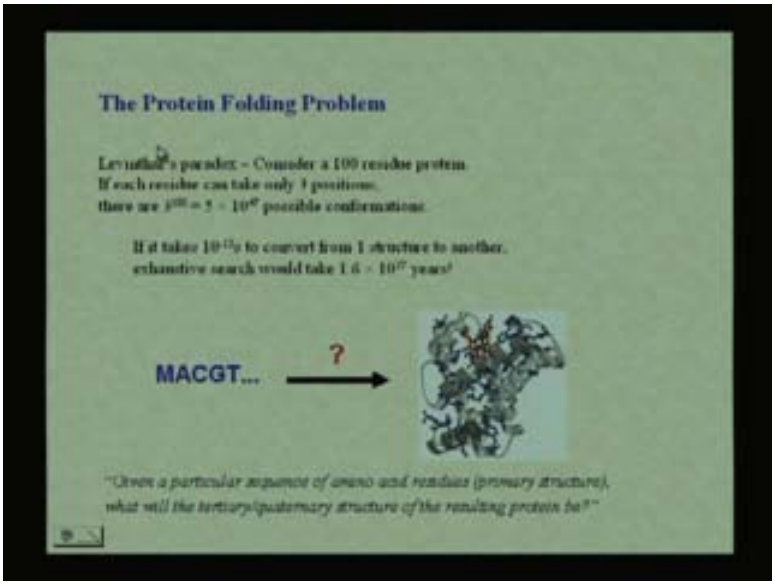


Now we should need to know the structures due to the following reasons. The structure will help us to understand the function, the mechanism evolution etc. it will help us in structure based drug design and it is also will help us basically to solve the protein folding problem because we will have more structures from which we can identify which amino acid sequence folds into which structure.

Here we have a random coil in which we have an α-helix, we have a β-strand and then we have α-helix. Here we have a sequence in one letter code in which we will have to know whether H means helix, C means coil and E means extended sheet that is the

nomenclature here is H C E. Here H is helix, C is coil, E is extended sheet. Now if we have this sequence we will have to know actually which part will be a helix, which part will be a coil and which part will be a sheet that is going to help us in analysis. Here we are considering about the Protein Folding problem.

(Refer Slide Time 5:31 min)

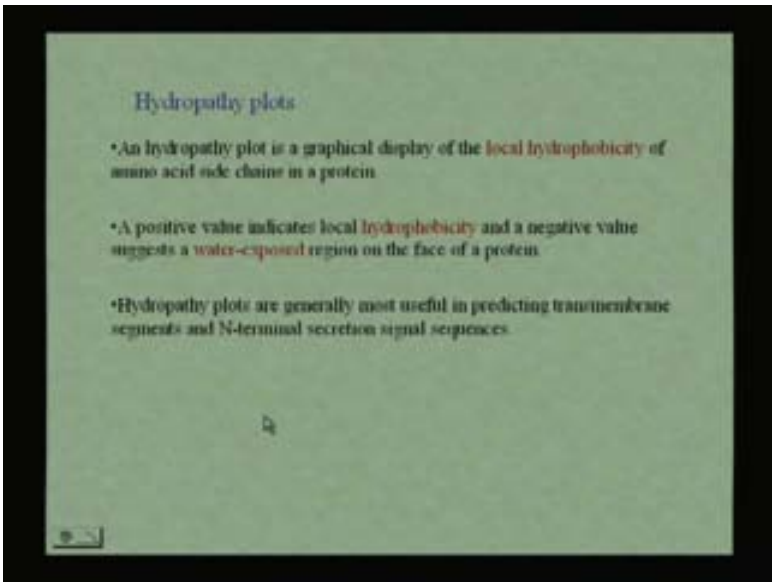


If you have a hundred residues protein you will know what is that mean it means you have hundred amino acids in your polypeptide chain. Now, if we consider that each residue can take only three positions then you know that you can have rotations about the bonds in the amino acids that connect amino acids together.

So, actually it can take on many more positions but if I consider that this hundred residue polypeptide chain can actually take only three positions then there are three to the hundred possible conformations for this polypeptide chain which is about ten to the forty seven possible conformations. Now the protein folds in one single structure and that is the only structure folds into. So in that ten to forty seven possible conformations that are available for a certain protein that is only hundred amino acid residues long if the protein decides to fold into a specific protein and it took less than a 10^{12} second to determine whether it was fold into anyone of these possible conformations. Then it would take 10^{27} years for a single protein to fold which it does in a matter of mille seconds. So it knows exactly how it supports to fold and where is this information? All the information will be avoilable in this sequence. And this is the big question that is still unanswered. We do not know how a particular sequence of amino acids residues that is the primary structure will go to which tertiary structure. Because you understand based on the conformational flexibility there are a very large number of conformations available to it but it will fold to it into a single structure. It is like that the example I give you of the necklace. You have a necklace of beads you pick it up, drop it on the table. It is never going to fall in the same conformation twice you have even 2D it will not do that and forget about the 3D.

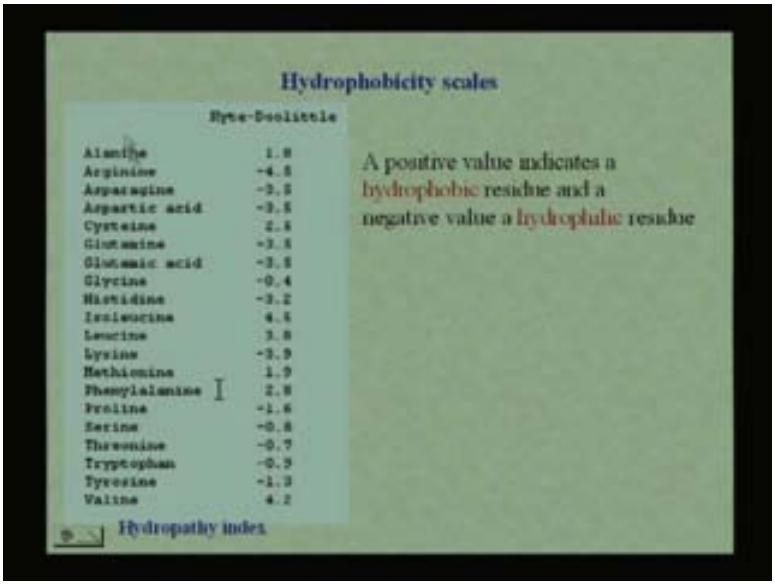
Therefore the whole problem of protein folding is what is a tertiary structure will be for a given particular sequence of amino acid? But what we can do is we can go for just small predictions. We can say from the structures that are already available how this particular sequence might form a helix or we can find out which part is going to be in the central region of the protein by determining which are hydrophobic in nature. So if I find a stretch of amino acids that are going to be hydrophobic in nature I can say that they might be forming the central part of the folded protein. That is the some information I can get which I will be a bit better of then just the primary sequence of the protein.

(Refer Slide Time 9:07 min)



This is what is called a hydropathy plot. The hydropathy plot is a graphical display of the local hydrophobicity of the amino acids side chains in a protein. Why do I want to do that? I know remember I showed you a table which gives you the hydrophobicity values of different amino acid residues. If you have a positive value then you have hydrophobic residues, if you have a negative value then you have a water expose region or a hydrophilic region. These hydropathy plots are actually most useful in predicting trans membrane segments but we will have to know how we can find a hydrophobic region. And what I may do with a hydrophobic region? I will predict that this hydrophobic region forms the center of the protein because I know the protein has a hydrophobic core with a hydrophilic surface to it. Then we learnt in the previous class that how we can explain whether a helix that is on the surface, we can tell which part is going to be inside which part is going to be outside. So now I will be slightly better of in saying about the whole protein sequence as to determining which part will be in the middle of the protein and which part is likely to be on the surface of the protein. So here we have a hydrophobicity scale which I showed earlier. And we have hydrophobic residues for the positive values, hydrophilic residues for the negative values.

(Refer Slide Time 10:56 min)

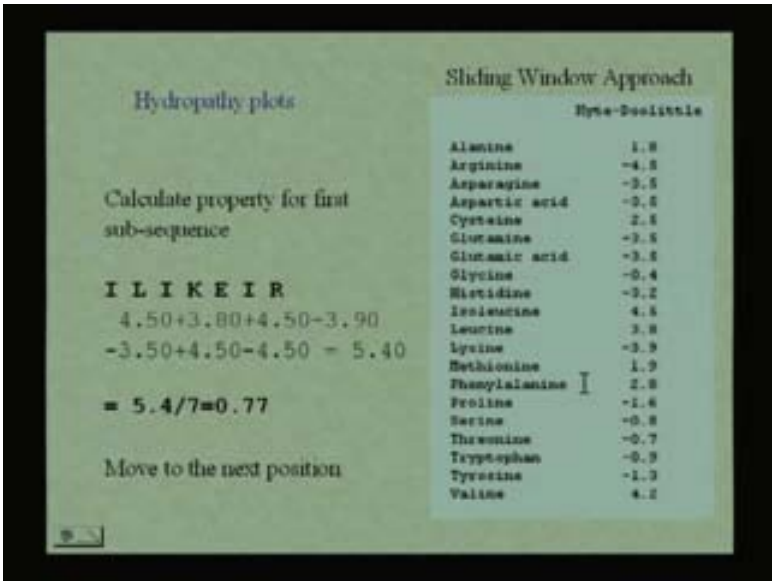


Now what can I do with this? I can go through a hydropathy plot. it is called a Sliding Window Approach. I will go through it very slowly and we will have to plot a hydropathy plot to determine whether a part of the protein is going to be on the surface or whether the part of the protein is going to be in the center of the protein. So what we have to do is we have to calculate the property for a sub sequence. What do mean by a sub sequence? Say I have this as the amino acid sequence.

We have I where I is Isoleucine then we have Leucine and another Isoleucine, Lysine, Glutamic acid, Isoleucine, Arginine. Now what I need to know is from the table how I can determine which part is inside which part is outside.

I have a specific sequence that I have here. Now I take the values for these amino acid residues and take the average of them. So what we will do is let us just put another amino acid for say Gly and Ala. So the first thing that I do here is I add the value for isoleucine. Now the value for Isoleucine is 4.5, the value for Leucine is 3.8, again 4.5, for Lysine is -3.9 why is it minus? Because it is hydrophilic in nature, for Glutamic acid is -3.5, Isoleucine is 4.5, Arginine is -4.5, Glycine is -0.4 and Alanine is 1.8.

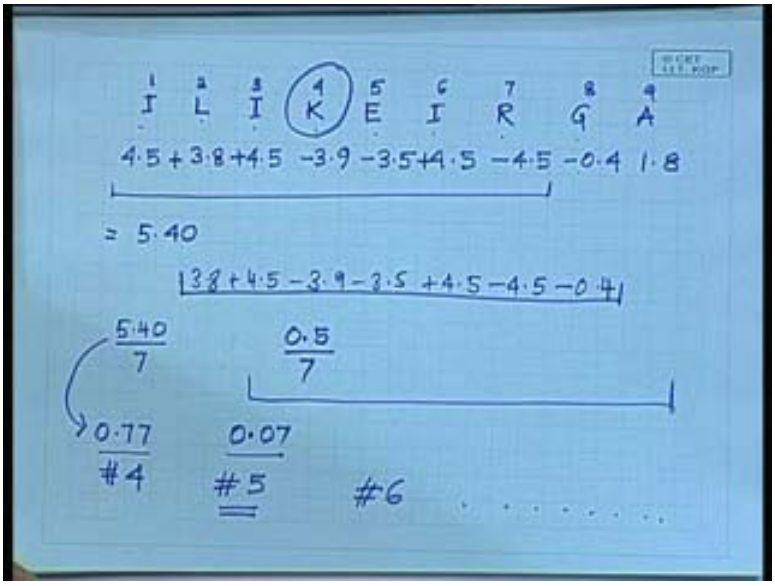
(Refer Slide Time 13:29 min)



Now this is called Sliding Window Approach. So this is residue number 1 then 2,3,4,5,6,7,8 and 9. I take the first seven residues that are called the window. I take a window of seven, I take the average of these so that I have to add all them and divide by 7. We can work that out and we find out that we get a value. We add all these up together from one through seven and then we get a value. We have to add 4.5 + 3.8 + 4.5 – 3.9 - 3.5 +4.5 – 4.5 the total comes to 5.40. So here we have the total as 5.4 we want the average of this. So we divide by 7 then we get 0.77 this is assigned to the central residue. So in this case the residue number 4 will have an average hydropathy index value of 0.77.

Then we move to next window, a sliding window approach. So what do I have to do now? I have to go from two to eight. When I go from two to eight I have to add all these numbers from Leucine, Isoleucine, Lysine, and Glutamic acid, Isoleucine, Arginine and Glycine together and then I have to divide by 7 again. Here 5.40 / 7 gave me 0.77 that is assigned to the central residue. So this is assigned to residue number #4. Then I take the other set then here I am going to loose 4.5 from this and add -0.4 basically. So what I am going to loose from 5.40, what is the value going to be for number #5? I will have a specific value here. So if I add all these values together from two that is 3.8 + 4.5 – 3.9 - 3.5 +4.5 – 4.5 -0.4.

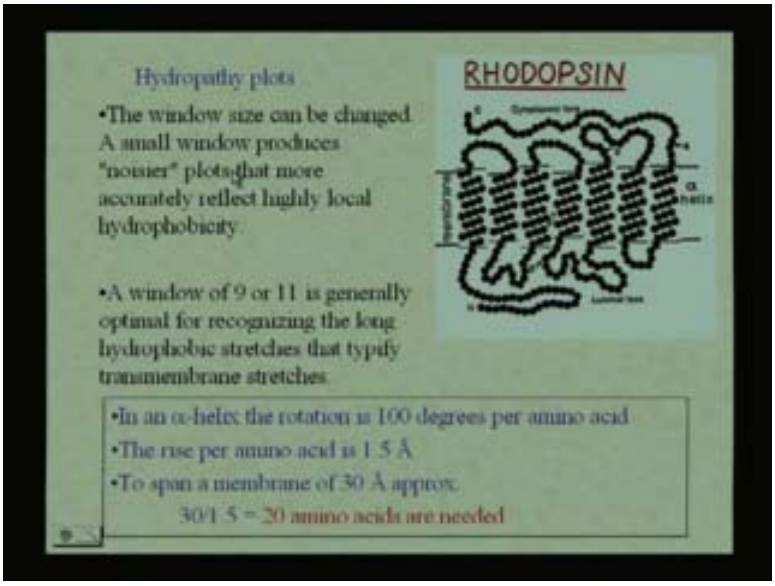
(Refer Slide Time 17:35 min)



I will get 0.5 then what do I have to do is have to divide by 7 that is going to give the value 0.07 so this is asign to residue number #5. Then I have to slide my window once more. Actually you will have to do this for the whole protein but we are not going to do it now. So I have to go from residue number 3 to 9. Then I get another value that I have assigned to residue number #6 and so on. Eventually what I may go to get? I am going to get values from leaving out the first three residues and the last three residues I will get values for the average hydrophobicity for the set of amino acids that formed this particular window. Then what you can do is make a plot.

Basically you have understood that you can change the window size. We can make it in a nine residue window or an eleven residue window but we make it an odd residue window so that we can assign it to the central amino acid.

(Refer Slide Time 19:03 min)



Usually if you have small windows then you have noisy at plots. This is usually nine or eleven is used, here we have used seven but that is fine. Now this is when we have membrane in this ease. We have a lipid bilayer, we have the Cytoplasmic face and we have the inside basically and the outside.

Now, if we look at the types of residues we understand from the helical wheel which will be hydrophilic in nature and which will be hydrophobic in nature. If we have a membrane that is around 30 Aº we know that the rise per amino acid residue is 1.5 Aº. What is that? That is the vertical rise per amino acid residue. The pitch that we saw was for a complete turn was 5.4 Aº and for a single amino acid we have 1.5 Aº. Now if know that my membrane is 30 Aº thick then how many hydrophobic residues should I have there? Twenty, because each 1.5 Aº is for every amino acid residue, I have to span 30 Aºs.

So if I have a stretch of amino acid residues that actually form a helix here that to spam the whole membrane I know if it were a single helix all of them would be hydrophobic in nature because I have my lipid hydrophobic tales that have to interact with the helix. So what can I do? I can say that when I am spamming the membrane with a helix then the nature of the residue in this helix is hydrophobic in nature. So all the ones that are sticking out here, all the side chains that are out here are going to be hydrophobic in nature.

Now what I need in such a specific protein sequence is a stretch of twenty amino acids because I have a 30 Aº stretch and I know that for every amino acid I traverse 1.5 Aº angstroms in height where 30 Aº is the thickness of the membrane. So the rise is 1.5 Aº per amino acid. So I need twenty hydrophobic amino acids to construct a hydropathy plot.

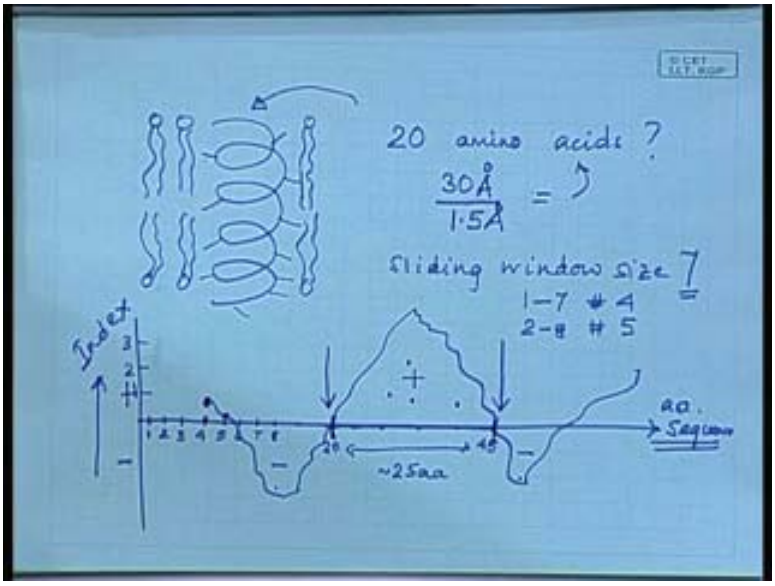
Now for the hydropathy plot here on the Y-axis we will have an index and on the X-axis we will have the sequence of the protein. So we have the amino acid sequence on the X-axis and we have index on the Y-axis. What is this index? It is the average index that we found out earlier. So I have residue 1 here and then 2, 3, 4, 5, 6, 7, 8 and so on. Then I had a sliding window where the window size was seven. So now I assigned the first value to residue number 4 which was 0.77 in this case. This is positive this is negative. So somewhere say if this is 1 this is 2 and this is 3 so 0.77 is some where here. I just make a plot. Then when the windows slide over instead of from residue one through seven which I assigned to residue number 4 and I went from two to eight which I assigned to residue number 5 that came out be 0.07 so that was very low down here. So I can complete a whole plot for protein. What do you need to construct a hydropathy plot? You need the sequence of the protein and you need the hydrophobicity values construct a hydropathy plot. Therefore we have the amino acid sequence and we also have the hydrophobicity in this ease. Then I have to find the average depending upon my window size. Then here I have a possible plot like this. This region is positive, this region is negative.

Now what can I say about the positive regions? They are hydrophobic in nature. Now you understand that when you take the average, a hydrophilic residue counteracts the effect of a hydrophilic residue. But if you have only a stretch of hydrophobic amino acids this value would be a high positive value. If you had a high positive value then you can have a stretch of highly hydrophobic amino acid residues. So, when we look at this I can say this is a highly hydrophobic stretch. When I am talking about a normal protein that is not a membrane protein I can safely say that this part is going to be the central part of the protein or the central core of the protein because it is hydrophobic in nature, it will not be on the surface.

But usually when we do these hydropathy plots they are mainly done for membranes because it tells you that this region is probably spamming the membrane. Why because if I have said the residue is from approximately number 20 to number 45 here. So what is my stretch of amino acids? I have approximately 25 amino acids which are hydrophobic in nature and I know this is a membrane protein and I can that this part forms the helix. I can very safely say that it is this part that is forming the helix of the membrane protein because this part is hydrophobic in nature. And I know if I have a single Transmembrane helix all of the residues have to interact with the lipid bilayer which is hydrophobic in nature.

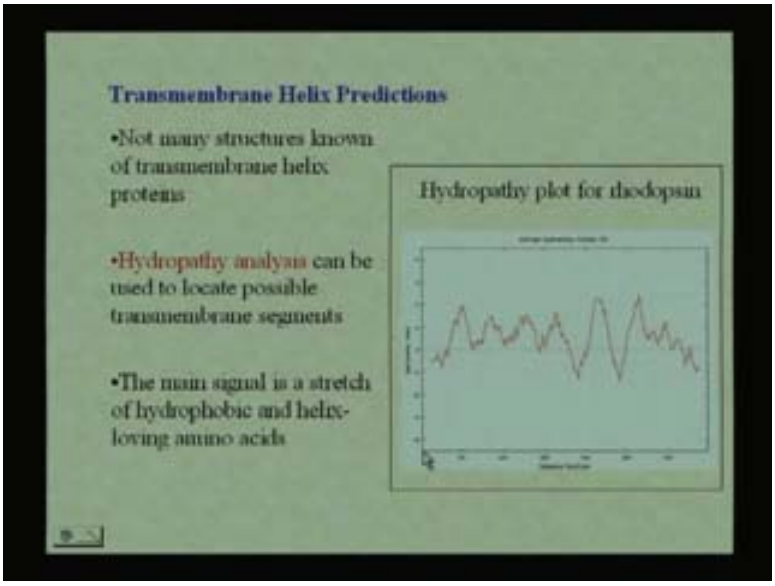
So I can plot a hydropathy plot that tells which region will be hydrophobic in nature and which region will be hydrophilic in nature. So I can say that these regions are going to be on the surface and I can say that these regions are going to be varied in the core of the protein. And I can say that a Transmembrane helix is going to be on the membrane side.

(Refer Slide Time 29:15 min)



Usually the hydropathy analysis is used to locate Transmembrane segments. But you can also do it for a regular protein because the reason being that not many structures of Transmembrane helix proteins are known. And the main signal is a stretch of hydrophobic and helix loving amino acids. What do you mean by helix loving amino acids? Residues those are likely to form α-helix. So that is what a hydropathy plot would look like. This is a hydropathy plot for a rhodopsin. So I can say all the positive parts if see this is the residue number 50 to 100, 150, 200, 250, 300 and so on. So these are stretches that are larger than twenty amino acid residues. Based on the scale, if it goes form 0 to 350 then these are larger than 20.

(Refer Slide Time 30:23 min)

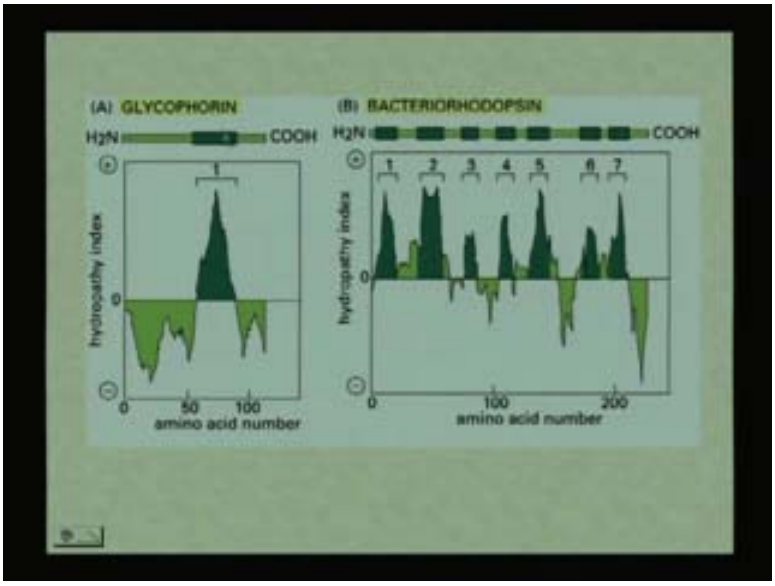


I can say I have 1, 2, 3, 4, 5, 6, 7 probable helices. What are these helices? They are interacting with lipid bilayer of the membrane because rhodopsin is a membrane protein. So this is a typical hydropathy plot. And what is the information you can get from this? Now you understand that if you have a stretch of a hydrophobic amino acid then this is the region that will be the helix part of the Transmembrane protein or rather this will be Transmembrane segment. This will be the helix that is going to interact with the lipid bilayer of the membrane. So again this is a very simple plot.

All the information you need is the sequence and the table. You can also construct a helical wheel. What is the information you need to construct the helical wheel? Just the sequence because for every amino acid you will get rotation is 100º. So you need the amino acid sequence for the construction of the helical wheel. You need additionally the hydrophobicity values of the amino acid residues for the hydropathy plot.

These are two other proteins where this is BACTERIORHODOPSIN and this is GLYCOPHORIN. Now you know which stretch is hydrophobic in nature. We know which stretch is hydrophobic in nature, which are mostly hydrophilic in nature. If this was for a normal protein then you could say the region number 1, 2, 3, 4, 5, 6, 7 would form the inner core of the protein the central part of the protein. So you could safely say these probably were on the outside. So you would be better of just having the sequence of the protein and no idea is about how the protein is folding.

(Refer Slide Time 32:32 min)



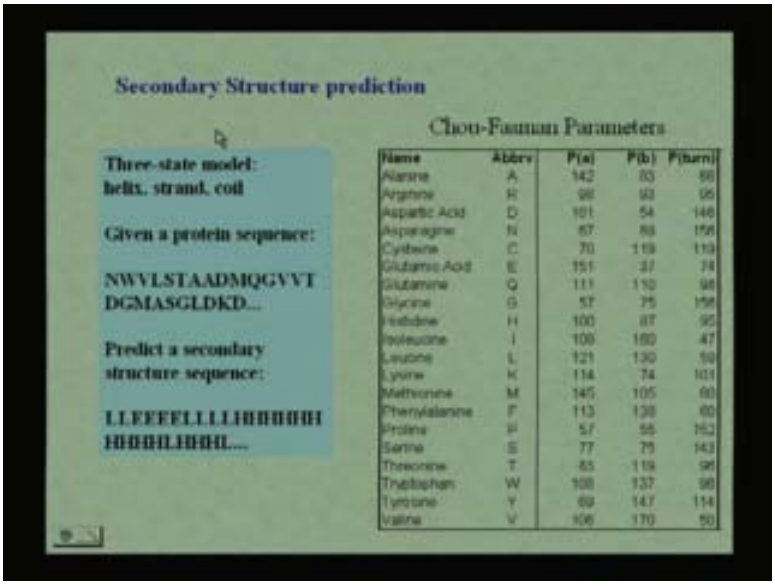
We know the primary sequence for over two hundred thousand proteins and we know the crystal structures for twenty five thousand proteins which is miserable. If you know only the structure of the protein, can you say the function or can you a design drug that is going to act on it? You con not say and it does not lead you anywhere with knowing the sequence for only two hundred thousand proteins and the crystal structures for only five thousand proteins. So we have to know what the structures of the all these proteins will be and we have to go for these prediction methods.

What does this give you an idea of? This gives you an idea that all we learned a helical wheel from just the sequence. If I know the sequence which will form a helix then I can say whether which part of the helix is going to be inside which is going to be outside.

Now, if we are looking at the sequence of this and I know from the hydrophobicity in disease where I have a hydrophobic region I can safely say that this hydrophobic region will form part of the protein core of the protein. But in this case when we are talking about Transmembrane segments these are the regions that traverse the mapping.

Now we want to go for secondary structure prediction. I want to know where a helix will form, so I am bit bolder now I had the sequence and from the sequence I could construct a helical wheel. But what is the idea of constructing a helical wheel if you don't know where the helix is going to be? You can not keep on doing with for the whole protein.

(Refer Slide Time 34:29 min)

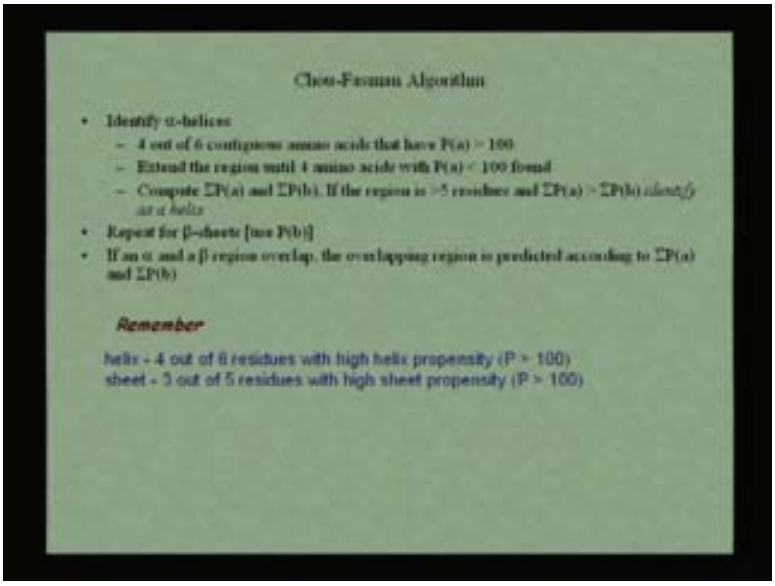


You can do the hydropathy index plot for the whole protein and then figure out which regions are inside or outside. But you have the protein sequence always available to do a secondary structure prediction. You do not have the structure always available.

If we want to construct a three state model we have the helix, the strand and the coil. So we have basically α, a β and a turn. These are just some numbers so we need another table if we want to go for a secondary structure prediction. This is a very famous way or rather very easy to predict whether you have a helix or not. These are called the Chou - Fasman Parameters. It tells you the chance or rather the propensity that you are going to have an Alanine in α-helix. Here this value is called the propensity. The larger the number will have the larger probability that a helix is going to be in that specific secondary structure.

Therefore what this table tells you is that you see all the twenty amino acids here, here the numbers tells you whether these amino acids will form a's that is alpha helices or b's that is beta strands or coils that is turns because you have your protein sequence. This is L that is another notation that is also used apart from C. So you want to know where the coils are or the turns of the loops are, you want to know where the helices are and you to know where the sheets are because that will give you a better idea of how a protein is going to fold. Because you will have some information from hydrophobicity, you will have some information about the secondary structure. So that will lead you into a better idea of how a protein is actually going to form its tertiary structure from its amino acid sequence. So we have the sequence of the protein, from the sequence of the protein we just have to look at the numbers.

(Refer Slide Time 37:20 min)



If six out of six contiguous numbers or rather six contiguous amino acids in which if four of them have p(a) > 100 then a helix will form.

If I had a helix, we know whether helix begins here so MQGVVT. Here M >100 so we have one amino acid greater than hundred. The Q where it is Glutamine which is also greater than hundred so I have two out of two greater than hundred. G is Glycine which is 57 which is less than hundred. So I have two out of three. Then V which is 163 so three out of four, again V is greater than hundred then four out of five, T which is Threonine is 83. So after go bit further I may write here MQGVVT. So we should have four that are greater than hundred. I have MQVV out of the six MQGVVT which are greater than hundred. So here I have a helix HHHHHH.

What about the next one? Then we extend the region until four amino acids with P(a) is less than hundred are found. So all you have to do is you just have a slide, again you are sliding window where you are looking at a window size of six. This six telling you that if you have four out of the six which are P(a) > 100 then you have an α-helix. If the P(a) > 100 then you do not have an α-helix any more.

How do you look for a β-sheet? If the P(b) > 100 for four out of six than you have a β-sheet. The problem arises when the α and the β regions overlapped. Then you have to do some mathematics. you have to some P(a) value of the six residues, some P(b) value of the six residues which ever is higher it is going to be that. (Refer Slide Time 40:25 min)

So what information do you have? You have a lot of information from the amino acid sequence. Now actually you can say whether you have a helix or not. So now it will make sense from an understanding of whether you have a helix or not. Then you can construct a helical wheel and you can say which part is going to be outside and inside. So we gradually getting into know more and more of the structure.

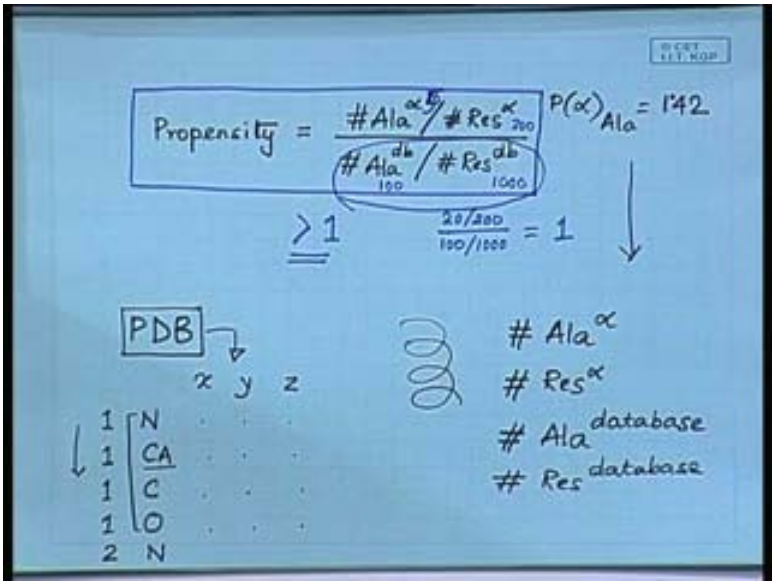
We have a helix which is 4 out of 6 residues with high helix propensity. Now I am talking about propensity, it is sought of a probability but you see the numbers are greater than hundred. In some tables you might see like we have the $P(\alpha)_{\mathbf{Ala}} = 142$ or some times put it as $1\dot{4}2$. The way Chou and Fasman got all these numbers was by a statistical analysis on the structures that are available. What do I mean by a structures are available? The crystal structures are solved for proteins are available in protein data bank. Now it is freely available where you can download protein structures. The protein structures are you have the x, y and z coordinates for all the atoms except the hydrogen atoms because X-ray Crystallography cannot look at hydrogen atoms.

So I am looking at residue number #1 then I will have nitrogen for residue number #1. Residue number #1 will also have a C_α atom associated with it, residue number #1 will also have carbon atom associated to it where it is part of the carboxylic group and residue number #1 will also have oxygen associated with it. What do I need to draw it? I need these values. Only if I have these values I can draw it in three dimensional spaces. So the protein data bank gives you these values for the twenty five thousand structures that are available in it. (Refer Slide Time 43:20 min)

So when I go to residue number two it starts again with nitrogen because I am going from the amino terminus to the carboxylic terminus. Then if you had a side chain you would have apart from a C_α, you would also have a C_β. so the C beta would be return after this. So this would be the back bone. Then you would have a C_β and so on but what we need to know is there are a set of structures available for which you can do an analysis. Here for say you look at all the helices that are there in the proteins and you have to count the number of Alanines that are there in the alpha helices. You have to count all the residues that are there in alpha α-helix only.

Then you have to count the number of Alanine in the database including the ones in the alpha. So you have to calculate all the Alanines that are present. You have to calculate the number of residues that are there in the database which ever database you are using. Propensity is a ratio. It is a ratio of the number of Ala in α divided by the number of residues in α to the number of Ala in database divided by the number of residues in the database which we can write it as Propensity = $[\#Ala^{\alpha} / \#Res^{\alpha}] / [\#Ala^{db} / \#Res^{db}]$. So this is your propensity calculation. This number is greater than one because you have to remember that you looking at a large sequence which is a polypeptide sequence of a large set of proteins.

(Refer Slide Time 45:55 min)

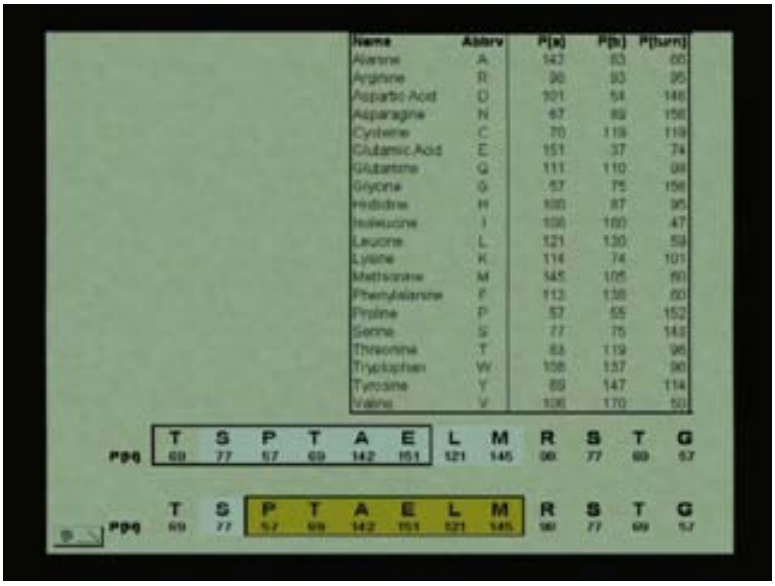


You want to whether alpha is preferred in helices. If I look at this, this will give me some idea about whether alpha is preferred in helices or not because this gives me the number of Alanine in the whole database. So if I have say 8% of Alanine in the whole database then I can calculate these as a percentage. For say I have thousand residues in the database and hundred of them are Alanine, From that thousand residues in the database two hundred are there in the alpha helices in which twenty are Alanine. So what would the value be, (20/200) / (100/1000) so this is equal to 1. This is nothing great that I have in the helices. Just I have ten percent as I have in the rest of the protein. I do not have any information about it. But if I had fifty Alanines then the value would be greater than one. Then I can see more of Alanine in helices which makes it significant than in the normal case of a protein. That is how they came up with these numbers in the slide here. So 1˙42 means that this number is greater than one which means the Alanine would like to be in α-helix.

But let us think about a Proline, all of you know how a Proline looks like. It will break the helix because you cannot have a turn properly and it is an amino acid that bends on to itself. So the propensity of it to forms a in an α-helix should be very low. Here the value of Glycine is 57 which is also a very low value why because it does not like to be an α-helix or it is not seen rather in α-helix for the analysis that has been done for the set of proteins which is true for mostly all the cells. Now if I look at a turn where these have mostly Glycine and Proline, the Glycine is because of its flexibility and Proline is because it basically helps in the turn back of chained at times. So look at these numbers 152 and 156 which are pretty high. And Asparagine is also high which is also 156.

So that is how these propensity values were actually determined. The propensity values have since been determined again for a very larger set of amino acids but this table is still used today for a rough prediction of where α-helix or β-sheet will be.

(Refer Slide Time 49:47 min)



We just need to have the table to figure out where are helix is going to be and where are sheet is going to be and the rest of it will going be coil. Here you have a turn set also, you have a p turn also.

So this is our table and this is our sequence. So I have the T S P T A E here I have put in values S is 77, T is suppose to be 83 where is 69. So we have Threionine, Serine, Proline and so and so forth. Now when I am at this point how many do I have that are greater than hundred just two. So I cannot say that I have a helix formation. I slide my window down to serine. I have an additional one greater than one hundred but it is still three out of six which is not good enough. Then I slide it again so now I have four out of six. So what can I say now? The helix begins.

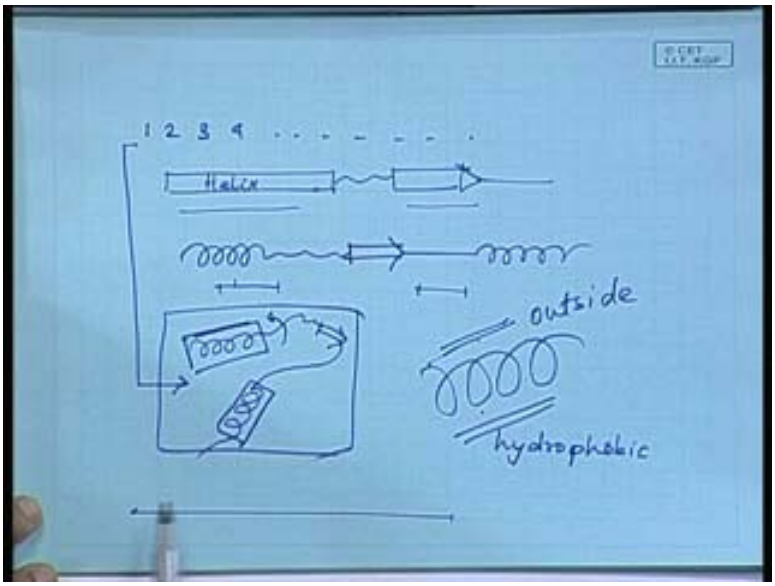
Here the helix begins basically just after the Proline but you need to know is you have a sequence. And you can say from the sequence and from the table that where the helix is going to be, where the sheet is going to be and where the turn is going to be. Then based on that what we can do is from this information we can roughly determine the sequence of the protein. So now I have the sequence 1 2 3 4…. and so on and what can I say is I have a helix here, I have a turn here and then I have a sheet here so we can say what we have. Then what do I do? I can do a hydropathy plot. What is hydropathetic plot going to tell me? It will tell me that which parts of these are hydrophobic in nature. So I can say this part is hydrophobic and again this part is hydrophobic.

So I can say that this part is going to be inside, this part is going to be outside and then again this part is going to be inside. So I have some information of how the protein is going to be fold. Now I can construct a helical wheel for this, I can also construct a helical wheel for this.

This face would be hydrophobic because this turn can rotate basically, I can have a rotation about this which would make either face in or out. Then what would I have to do? I have to construct a helical wheel. If I know that this face is hydrophobic I have to turn it around to make it to come to the core of the protein. So I have a hydrophobic region and a hydrophilic region which is therefore on the outside.

Now I am better of in determining how the protein is going to interact, how it is going to fold into giving the final tertiary structure. So what we learnt is that we can determined where we might have the helix from the Chou -Fasman parameters, we can determine where we can have the hydrophobic regions or where we can have Transmembrane segments form a hydropathy plot and we can determined whether this part is outside or inside from the helical wheel.

(Refer Slide Time 54:40 min)



So we are better of just on the primary amino acid sequence of the protein. Thank you.