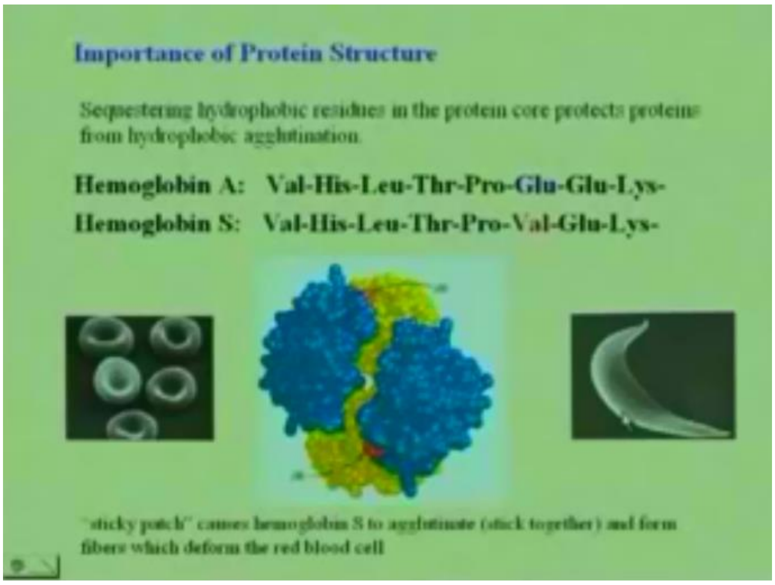
Biochemistry
Prof. S. DasGupta
Department of Chemistry
Indian Institute of Technology – Kharagpur

Lecture - 5
Protein Structure - III

This is lecture number three on protein structure.

(Refer Slide Time: 00:52)



Importance of Protein Structure

Sequestering hydrophobic residues in the protein core protects proteins from hydrophobic agglutination

Hemoglobin A: Val-His-Leu-Thr-Pro-Glu-Glu-Lys-

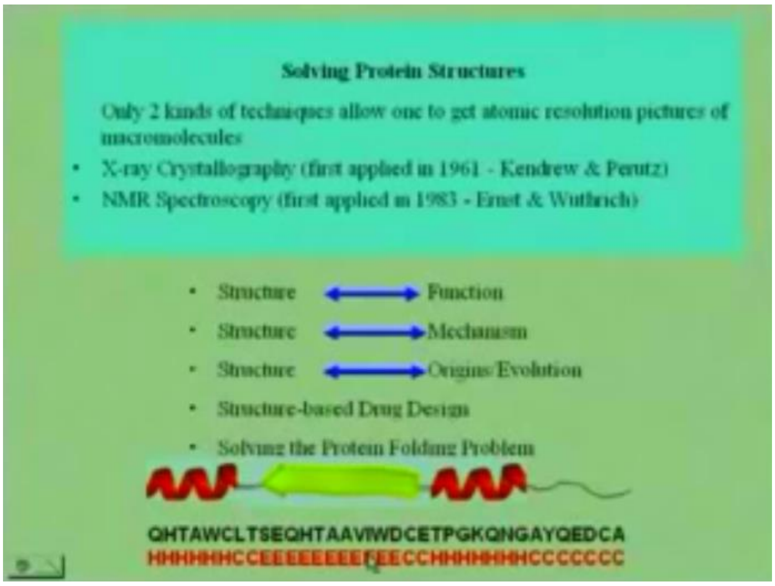
Hemoglobin S: Val-His-Leu-Thr-Pro-Val-Glu-Lys-

"sticky patch" causes hemoglobin S to agglutinate (stick together) and form fibers which deform the red blood cell

Last time we discussed about the importance of protein structure and I just want to go through this slide once more to show how important the property of an amino acid is, in this case actually giving you disease. So, what happens here is, you have a difference in the folding of the protein. The reason why we have this change is because of a specific genetic mutation. This is what is called a mutation, an amino acid mutation, where we have glutamic acid go to valine.

Now, because of this mutation, we have a hydrophobic residue on the surface and because of this property is entirely different from the glutamic acid, we have these sequestering together or sticking together to form a fiber, which is what is going to give a disease. So, the folding of the protein is extremely important and there are certain forces associated with the folding of the protein, which we will be doing it later on.

(Refer Slide Time: 02:01)



Solving Protein Structures

Only 2 kinds of techniques allow one to get atomic resolution pictures of macromolecules

- X-ray Crystallography (first applied in 1961 - Kendrew & Perutz)
- NMR Spectroscopy (first applied in 1983 - Ernst & Wuthrich)

- Structure ⟷ Function
- Structure ⟷ Mechanism
- Structure ⟷ Origins/Evolution
- Structure-based Drug Design
- Solving the Protein Folding Problem

QHTAWCLTSEQHTAAVIWDCETPGKQNGAYQEDCA
HHHHHHCCEEEEEEEEEEEEECCHHHHHHHHCCCCCC

But for now, we are going to understand, more of how we can actually get to solving protein structures, because now, we 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 how to get the protein structure, there are actually only two kinds of techniques that are available to get atomic resolution pictures because you want to know exactly where the atoms are, which are going to tell you what your protein looks like.

And you understand that it is an extremely large system, so it is a macro molecule, so it is very difficult to pin point, where the atoms are if you are to take a snapshot of it. Now, there are two techniques available. One is x-ray crystallography and the other is NMR spectroscopy. Now, x-ray crystallography is by far the only method that is still taken to date as the method to solve a protein structure, but NMR spectroscopy is very fast catching up.

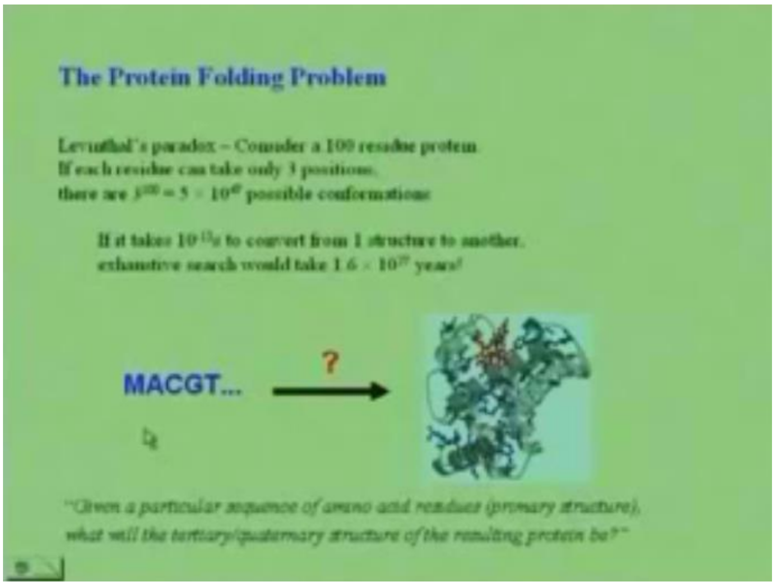
Now, the difficulty of x-ray crystallography is getting a crystal. You understand if you wanted to crystallography you have to have a crystal of the protein. Now, getting a protein crystal is very difficult, because of its large size it actually either forms a powder or it just sticks together and does not form a crystal at all. So, if you do not have a crystal you cannot do a crystallography and that is why it is very difficult to get protein crystals, which is why people are going to what is called protein structure prediction.

Because, it is easy to get the sequence of the protein, you can find out the amino acid sequence of the protein very easily, a method that we will be studying also. And to get from the sequence to the structure is what the problem is. Another reason why we need to know the structure is because of the following reasons. The structure is going to help us understand the function, the mechanism, evolution, it is going to help us with what is called structure based drug design and it is also going to help us basically solve the protein folding problem because we have more structures from which we can identify which amino acid sequence folds into which structure.

So, say, we have something like this, what is this? We have a random coil, we have Alpha-helix, and we have a beta strand and we have alpha-helix here. What we want to know is, this is say the sequence in one letter code. What do I want to know? I want to know whether H means helix, C means coil and E means extended sheet. That is what the nomenclature here is, H, C, E. H is helix, C is coil and E is extended sheet.

Now, what I want to know they for is, if I have this sequence how I can actually say which part is going to be a helix, which part is going to be a coil and which part is going to be a sheet. That is going to help me in my analysis.

(Refer Slide Time: 05:17)



The Protein Folding Problem

Levinthal's paradox – Consider a 100 residue protein.
If each residue can take only 3 positions,
there are $3^{100} = 5 \times 10^{47}$ possible conformations

If it takes 10^{-13}s to convert from 1 structure to another,
exhaustive search would take 1.6×10^{27} years!

"Given a particular sequence of amino acid residues (primary structure), what will the tertiary/quaternary structure of the resulting protein be?"

Now, this, I am going to mention something about this, because this is a very current topic of investigation, a very current topic of research, which is called the protein folding problem. Now just consider this. If you have a 100 residue protein, you know now what does that mean. It means you have 100 amino acids in your polypeptide chain. Now, if we consider that each residue can take only 3 positions, what do I mean by that?

You know that you can have rotations about the bonds in the amino acids that connect the amino acids together. So if, actually it can take on many more positions, but if I consider that this 100 residue polypeptide chain can actually take on only 3 positions, then there are 3 power 100 possible conformations for this polypeptide chain. Which are about 10 to the 47 possible conformations?

Now, the protein folds in one single structure as I showed you in the first day itself when we looked at ribonuclease A, that is the only structure it folds into. So, off the 10 to the 47 possible conformations that are available for a certain protein that is only 100 amino acid residues long, if the protein decides it wants to fold into a specific protein and it took less than a picosecond to determine whether it was going to fold into any one of these possible conformations, then it would take 10 to the 27 years for a single protein to fold which it does in a matter of mille seconds.

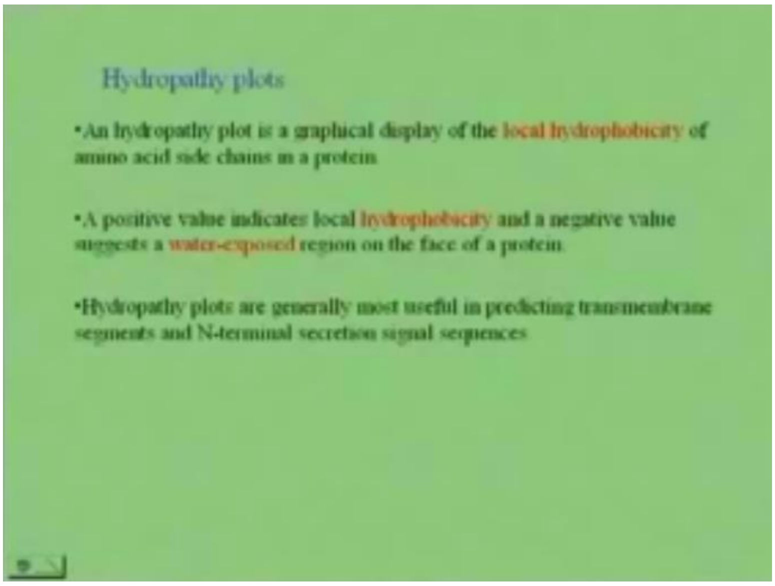
So, it knows exactly how it is supposed to fold and where is this information? It is in this sequence. All the information is in this sequence and this is the big question that is still unanswered. We do not know how a particular sequence of amino acid residues, that is the primary structure, is going to 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 into a single structure. It is like that example I give you of the necklace. You have a necklace of beads. You pick it up and drop it on the table. It is never going to fall in the same conformation twice. Even 2D it will not do that and for 3D forget about it. So, the whole problem of protein folding is, given a particular sequence of amino acid, what is the tertiary structure going to be? But, what we can do is, we can go for a small prediction.

We can say from the structures that are already available, how, say this particular sequence, might form a helix or might form what is called, we can find out which part is going to be in the central region of the protein by determining what? Determining which are hydrophobic in nature? okay? So, if I find a stretch of amino acids that are going to be hydrophobic in nature, what can I say about them? I can say that they might be forming the central part of the folded protein.

That is some information, I can get which I will be a bit better off than just the primary sequence of the protein. So, how do we do that?

(Refer Slide Time: 09:06)



Hydropathy plots

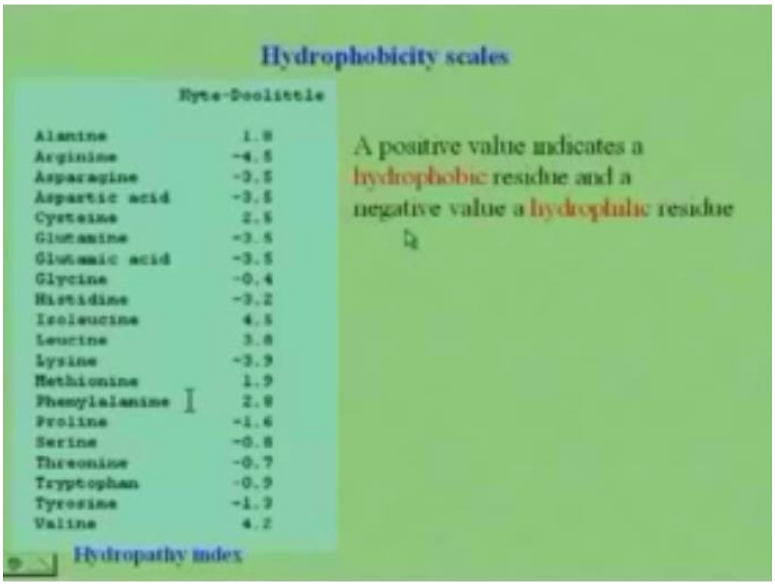
- An hydropathy plot is a graphical display of the local hydrophobicity of amino acid side chains in a protein
- A positive value indicates local hydrophobicity and a negative value suggests a water-exposed region on the face of a protein
- Hydropathy plots are generally most useful in predicting transmembrane segments and N-terminal secretion signal sequences

This is what is called a hydropathy plot. What is a hydropathy plot? It is a graphical display of the local hydrophobicity of the amino acid side chains in a protein. Why do I want to do that? I know, remember I showed you a table, which I will show you again, which gives you the hydrophobicity values of different amino acid residues. If you have a positive value, then you have a hydrophobic residue.

If you have a negative value, you have a water-exposed region or a hydrophilic region. These hydropathy plots are actually most useful in predicting transmembrane segments but, for now, we will know how we can find a hydrophobic region. What am I going to do with a hydrophobic region? I will then predict that this hydrophobic region forms the center of the protein. Because, I know that the protein is a hydrophobic core with a hydrophilic surface to it.

And then we learned in the previous class, how we can say whether a helix that is on the surface, we can tell, which part is going to be inside and which part is going to be outside. So now, I will be slightly better off in saying about the whole protein sequence as to determining, which part is going to be in the middle of the protein and which part is likely to be on the surface of the protein.

(Refer Slide Time: 10:41)



Hydrophobicity scales

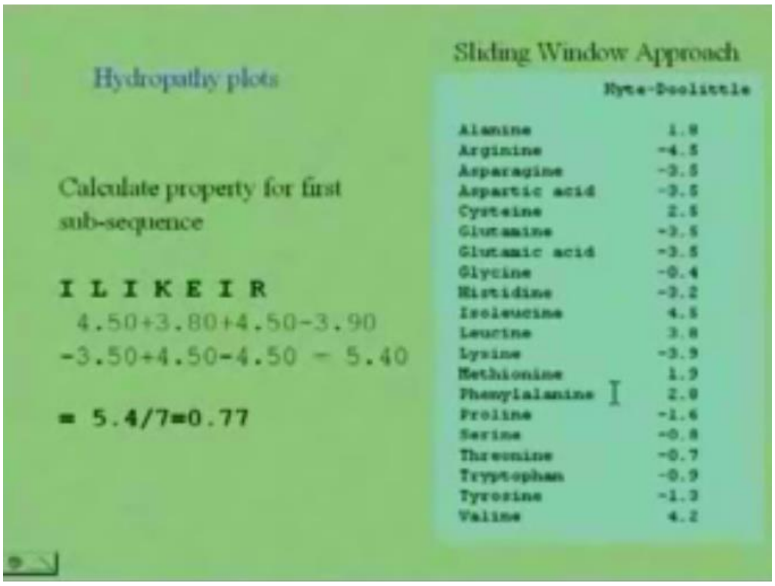
	Kyte-Doolittle
Alanine	1.8
Arginine	-4.5
Asparagine	-3.5
Aspartic acid	-3.5
Cysteine	2.5
Glutamine	-3.5
Glutamic acid	-3.5
Glycine	-0.4
Histidine	-3.2
Isoleucine	4.5
Leucine	3.8
Lysine	-3.9
Methionine	1.9
Phenylalanine	2.8
Proline	-1.6
Serine	-0.8
Threonine	-0.7
Tryptophan	-0.9
Tyrosine	-1.3
Valine	4.2

Hydropathy index

A positive value indicates a hydrophobic residue and a negative value a hydrophilic residue

So, what do we have here? This is a hydrophobicity scale. This I showed you in one of the previous classes and what do we have here? For the positive values, we have hydrophobic residues. The negative values are hydrophilic residues. Now, what can I do with this?

(Refer Slide Time: 10:58)



Hydropathy plots

Calculate property for first sub-sequence

I L I K E I R
4.50+3.80+4.50-3.90
-3.50+4.50-4.50 = 5.40

= 5.4/7=0.77

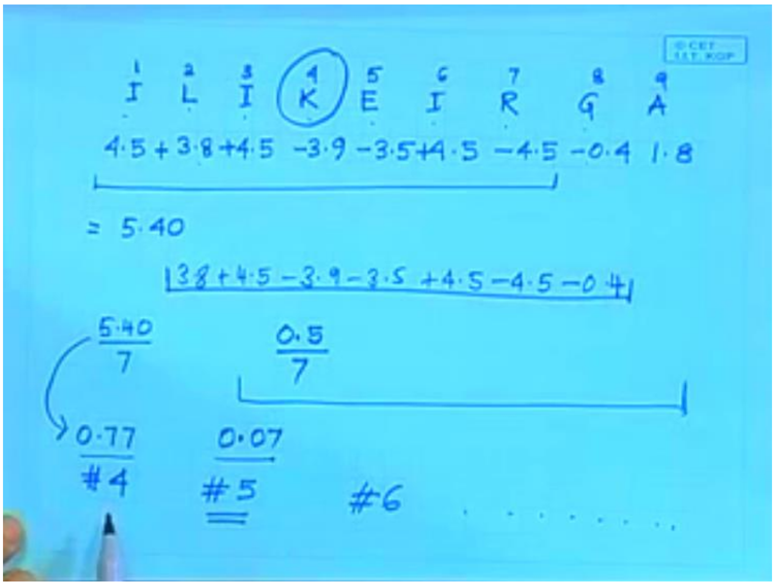
Sliding Window Approach

	Kyte-Doolittle
Alanine	1.8
Arginine	-4.5
Asparagine	-3.5
Aspartic acid	-3.5
Cysteine	2.5
Glutamine	-3.5
Glutamic acid	-3.5
Glycine	-0.4
Histidine	-3.2
Isoleucine	4.5
Leucine	3.8
Lysine	-3.9
Methionine	1.9
Phenylalanine	2.8
Proline	-1.6
Serine	-0.8
Threonine	-0.7
Tryptophan	-0.9
Tyrosine	-1.3
Valine	4.2

I can go to what is called a hydropathy plot. It is called a sliding window approach. I am going to go through it very slowly and we are going to plot a hydropathy plot to determine whether a part of a protein is going to be on the surface or whether a part of a protein is going to be in the center of the protein. So, what do we do? We have to calculate the property for a sub-sequence. What do I mean by a sub-sequence? Say, I have this as my amino acid sequence.

So, let’s just write this down. We have, what is I, I is isoleucine, then we have leucine, another isoleucine, lysine, glutamic acid, isoleucine and arginine. What I need to know now is, from the table, how I can determine, which part is inside, which part is outside. So, what do I do? I have a specific sequence that I have here. I now take the values for these amino acid residues and take the average of them. So, what do I do?

(Refer Slide Time: 12:33)



Let’s just put another amino acid here, 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. 4.5. Lysine is -3.9. Why is it minus? Because it is hydrophilic in nature. Glutamic acid is -3.5. Isoleucine is 4.5. Arginine is -4.5. Glycine is -0.4. And alanine is 1.8. Now, this is called as sliding window approach.

So, this is my residue number 1, 2, 3, 4, 5, 6, 7, 8, 9. I take the first 7 residues. That is called my window. I take a window of 7. I take the average of these. So, what do I have to do? I have to add them all up and divide by 7. You can work that out and we find out that we get a value of, what is the value that we get? We add all these up together from 1 through 7 and then we get a value, which I will show on the slide here.

We have to add 4.5+3.8+4.5-3.9-3.5+4.5-4.5. What does that total come to? Did you work it out? The total comes to 5.40. What do we want? We go back to the slides here. We have the total as 5.4. We want the average of this. So, we divide by 7. 0.77, this is a sign to the central residue.

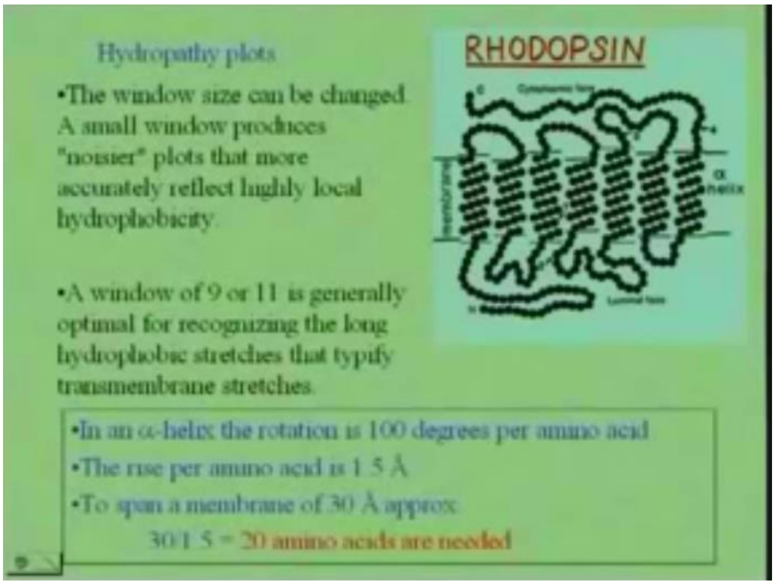
So, residue number 4 in this case, is going to have a hydropathy index value, average hydropathy index value of 0.77. Then what do we do? We move to the next window. We have sliding window approach. So, what do I have to do now? I have to go from 2 to 8. When I go from 2 to 8, I have to add all these numbers from leucine, isoleucine, lysine, glutamic acid, isoleucine, arginine and glycine together.

And, then I have to divide by 7 again. So, 5.40 divided by 7 gave me this that is a sign to the central residue. So, this is a sign to residue number 4. If I take the other set, then what am I going to lose? I am going to lose 4.5 from this and add -0.4 basically. So, what am I going to lose from 5.40? So, what is my value going to be for number 5? I will have a specific value here, so, if I add all these values together from 2 that is 3.8+4.5-3.9-3.5+4.5-4.5-0.4, I am going to get what? What do I have to do? Divide by 7 that is going to give me, so how much am I going to get here? 0.07.

So, then what do I do again? This is a sign to residue number 5. Then, I have to slide my window once more. Well actually you have to do this through the whole protein, but we are not going to do it now. So, I have to go from now residue 3 to 9. Then, I get another value that I assign to residue number 6 and so on. Eventually, what am I going to get? I am going to get values from leaving out the first 3 residues and the last 3 residues.

I am going to 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.

(Refer Slide Time: 18:41)



Hydropathy plots

- The window size can be changed. A small window produces "noisier" plots that more accurately reflect highly local hydrophobicity.

- A window of 9 or 11 is generally optimal for recognizing the long hydrophobic stretches that typify transmembrane stretches.

- In an α-helix the rotation is 100 degrees per amino acid
- The rise per amino acid is 1.5 Å
- To span a membrane of 30 Å approx

  30/1.5 = 20 amino acids are needed

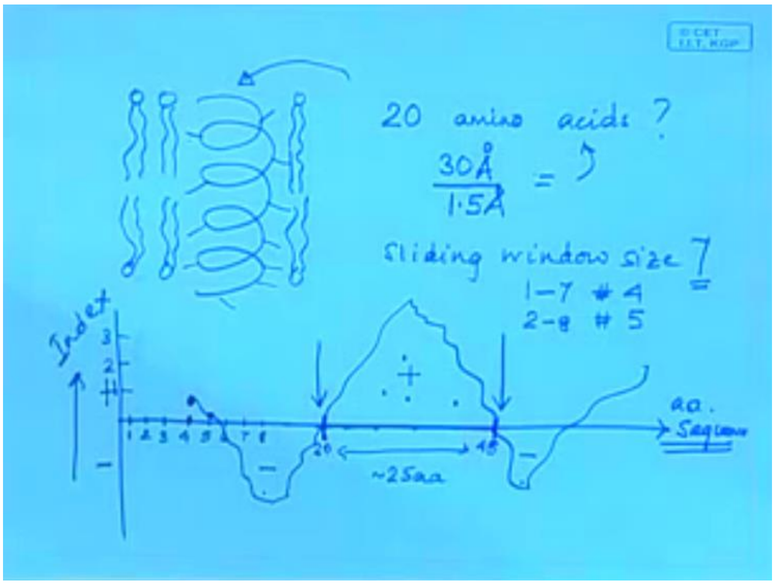
Now basically you understand that you can change the window size. We can make it a 9-residue window or 11 residue window, we make it an odd residue window, so that we can assign it to the central amino acid. Usually what happens is, if you have a small window you have noisier plots. We will look at a plot and see what it looks like. This is usually a 9 or 11 is used. We have used 7, but that’s fine.

Now, this is when we have membrane helices. This is something I was mentioning in the previous class. We have our lipid bi-layer. We have the cytoplasmic phase and we have the inside basically and the outside. Now, if we look at the types of residues that we have here, we understand now from the helical wheel which are going to be hydrophilic in nature and which are going to be hydrophobic in nature.

Now, if we have a membrane that is around 30 angstroms, we know that the rise per amino acid residue is 1.5 angstroms. What is that? That is the vertical rise per amino acid residue. The pitch that we saw was for a complete turn. That was 5.4 angstroms for a complete turn. And this is for a single amino acid we have 1.5 angstroms. Now, if I know that my membrane is 30 angstroms thick, then how many hydrophobic residues should I have there? 20. Why? Because each 1.5 angstroms is for every amino acid residue. I have to span 30 angstroms.

So, if I had a helix or if I had a stretch, let us not call it a helix, right now. If I have a stretch of amino acid residues that actually form the helix here, that were to span the whole membrane, I know if it were a single helix, all of them would be hydrophobic in nature. Why? Because I have my lipid hydrophobic tails that have to interact with the helix. So what can I do?

(Refer Slide Time: 21:28)



I can say that when I am spanning the membrane. This is my membrane. I am spanning the membrane with the helix. What is the nature of the residues in this helix? They are hydrophobic in nature. Now, so all the ones that are sticking out here, all the residues that the side chain that are out here are going to be hydrophobic in nature. Now, what I need in such a specific protein sequence is I need a stretch of 20 amino acids. Why? Because I have 30 angstroms.

What is 30 angstroms? It is the thickness of the membrane. So, I have 30 angstroms stretch and I know that for every amino acid I traverse 1.5 angstroms in height. So, the rise is 1.5 angstroms per amino acid. So, I need 20 amino acids to, what type of amino acids? Hydrophobic. How do I determine that? I construct a hydropathy plot. Ok now, what we did find for the hydropathy plot? This is going to be the index.

Here, on the y axis, we are going to have index. And on the x axis, is going to be the sequence of the protein. So, we have the amino acid sequence on the x axis. And we have the index. What is this index? It is the average index that we found out in the previous side. So, if I have the residue 1 here 2, 3, 4, 5, 6, 7, 8 and so on. Then, where was my, I had a sliding window, what was my sliding window size? 7. Size was 7.

What do I have now? I assign the first value to residue number 4, which was in this case .77, so this is positive, this is negative. So, somewhere here. Let us mark 1, 2, 3. So, if this is 1, 2, this is 3. So .77 is somewhere here. I just make a plot. Then, when the window slid over from residue, from 1 through 7 which I assigned residue number 4, I went from 2 to 8, which I assigned residue number 5. That came out to be .07 and that was very low down here.

So, I can complete a whole plot for the protein. What do you need for this? What you need to construct a hydropathy plot? You need the sequence of the protein and you need the hydrophobicity values. That is all the information you need. So, we have our amino acid sequence and we also have the hydrophobicity indices. Then, I have to find the average depending upon my window size. Then see, I have a plot like this. That is possible? This region is positive. This region is negative. Now, if we go back to the data in our slides here.

What can I say about the positive regions? They are hydrophobic in nature. So if I have, now you understand when you take the average, a hydrophilic residue counteracts the effect of a hydrophilic residue. But, if you had the stretch of hydrophobic amino acids only, then this value would be a high positive value. If you had a high positive value, what can you say? You can have a stretch therefore of highly hydrophobic amino acid residues.

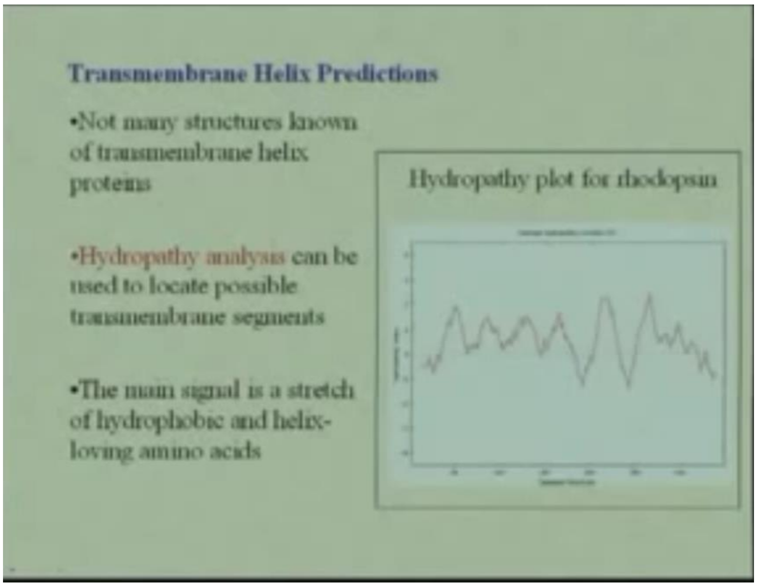
So, when we look at this and I say that this stretch is a highly hydrophobic stretch, when I am talking about a normal protein, that is not a membrane protein, I can safely say that this part is going to be of form part of the central part of the protein. Central core of the protein. Why? Because, it 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 spanning the membrane. Why? Because, if I say my residue is approximately from number 20 to 45 here, so what is my stretch of amino acids? I have approximately 25 amino acids that are hydrophobic in nature and I know that this is a membrane protein, what can I say about it? I can say that this part forms the helix. I can very safely say that it is this part that is forming the helix of my membrane protein.

Why? Because, this part is hydrophobic in nature and I know that if I have a single trans membrane helix, all of the residues have to interact with the lipid bilayer, which is hydrophobic in nature. So, I can plot the hydropathy plot that is going to tell me, which region is going to be hydrophobic in nature and which region is going to 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 buried in the core of the protein. And for a transmembrane helix, I can say, it is going to be on the membrane side. Now, usually as I mentioned the hydropathy analysis is used to locate transmembrane segments.

(Refer Slide Time: 29:22)



Transmembrane Helix Predictions

- Not many structures known of transmembrane helix proteins
- Hydropathy analysis can be used to locate possible transmembrane segments
- The main signal is a stretch of hydrophobic and helix-loving amino acids

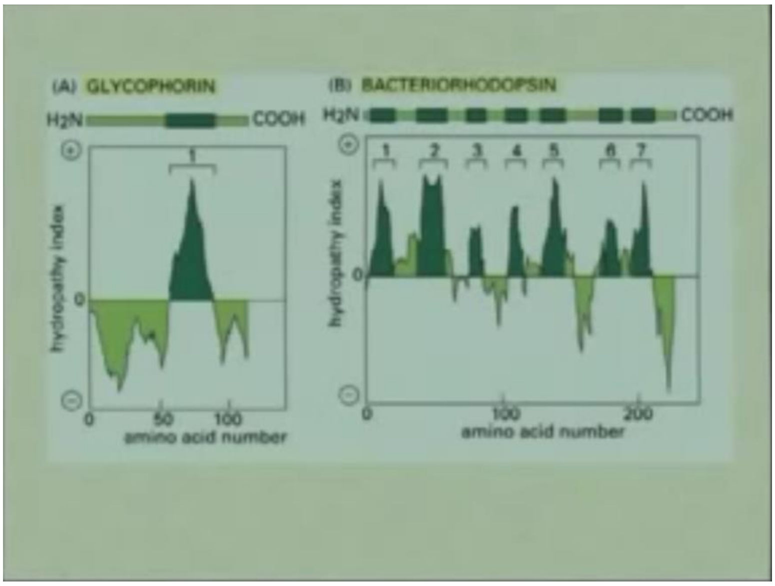
Usually, but you can also do it for a regular protein. Because the reason being that not many structures of transmembrane helix proteins are known and what is the main signal? A stretch of hydrophobic and helix loving amino acids. What do you mean by helix loving amino acids? Residues that are likely to form an alpha helix. So, that’s what a hydropathy plot would look like.

This is the hydropathy plot for a rhodopsin, the one that I showed you on top, the structure. So, what can I say? All these positive parts, this is residue number 50 to 100, 150, 200, 250, 300. These are stretches that are larger than 20 amino acid residues, based on the scale if it goes from 0 to 350, then these are larger than 20. So what can I say? I have 1, 2, 3, 4, 5, 6 and 7 probably helices. What are these helices? They are interacting with the lipid by layer of the membrane because rhodopsin is a membrane protein.

So, this is the typical hydropathy plot that you can see and what is the information that you get from this? You understand now that if you have a stretch of hydrophobic amino acids then this is the region that is going to be the helix part of the transmembrane protein or this is going to be the transmembrane segment rather. This is going to be the helix that is going to interact with the lipid by-layer of the membrane.

So this is a very simple plot. What is the information you need? All the information you need is the sequence and the table. You can also construct a helical wheel. What is the information that you need for the helical wheel? Just the sequence and nothing else, because you know that for every amino acid the rotation is 100 degrees. So, all you need for the construction of the helical wheel is the amino acid sequence, what you need for the hydropathy plot is additionally the hydrophobicity values of the amino acid residues.

(Refer Slide Time: 31:45)



So, these are other 2 proteins, where you can figure out. This is bacteriorhodopsin and this is glycophorin. What do we have here? So, you know which stretch is hydrophobic now. Right We know which stretch is hydrophobic and which almost is hydrophilic in nature. Now, if this were for the normal protein say, then what could you say about these regions? You could say that the region number 1, 2, 3, 4, 5, 6 and 7 would form the inner core of the protein, the central part of the protein.

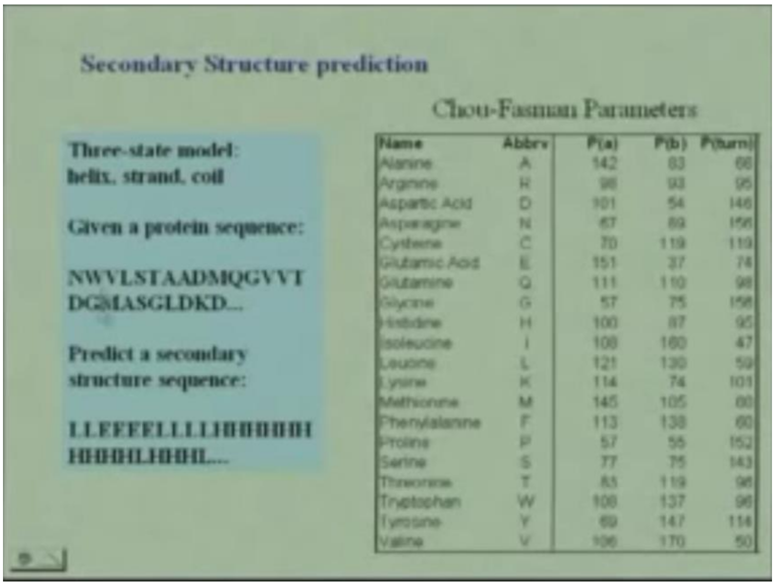
So, you could safely say that these probably are on the outside. So, you would be better off than just having the sequence of the protein and no ideas to how the protein is folding. People are not as good as in predicting the protein structure but research is going up to 77 percent today. But, still that is not good enough. The reason being that the primary sequence we know for over 200000 proteins and the crystal structure we know for 25000 which is miserable.

Considering that only if you know the structure of the protein can you say the function or you can design a drug that is going to act on it, knowing the sequence for 200000 proteins and the crystal structure only for 25000 proteins doesn’t lead us anywhere. So, we have to know what the structures of all these proteins are going to be and we have to go for these prediction methods.

So, what does this give you an idea of? First of all, we learnt from a helical wheel, from the sequence, if I know that the sequence is going to form a helix, then what can I say? I can say, which part of the helix is going to be inside and 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 is going to form a part of the protein, core of the protein.

But, in this case, when we are talking about the trans membrane segments, these are the regions that traverse the membrane.

(Refer Slide Time: 34:07)



Secondary Structure prediction

Three-state model:
helix, strand, coil

Given a protein sequence:

NWVLSTAADMQGVVT
DGMASGLDKD...

Predict a secondary
structure sequence:

LLEEEELLLLHHHHHHH
HHHHLHHHL...

Chou-Fasman Parameters

Name	Abbrv	P(a)	P(b)	P(turn)
Alanine	A	142	83	66
Arginine	R	98	93	95
Aspartic Acid	D	101	54	146
Asparagine	N	67	89	156
Cysteine	C	70	119	119
Glutamic Acid	E	151	37	74
Glutamine	Q	111	110	98
Glycine	G	57	75	156
Histidine	H	100	87	95
Isoleucine	I	108	160	47
Leucine	L	121	130	59
Lysine	K	114	74	101
Methionine	M	145	105	60
Phenylalanine	F	113	138	60
Proline	P	57	55	152
Serine	S	77	75	143
Threonine	T	83	119	96
Tryptophan	W	108	137	96
Tyrosine	Y	69	147	114
Valine	V	106	170	50

So, now we want to go for secondary structure prediction. I want to know, where a helix is going to form. So, I am a bit bolder now. I had my sequence. From my sequence, I could construct a helical wheel. But, what is the idea of constructing a helical wheel if you do not know whether helix is going to be. You can’t keep doing it for the whole protein. You can do the hydropathy index plot for the whole protein and then figure out, which regions are going to be inside or outside, but if I want to do a secondary structure prediction, what does it mean?

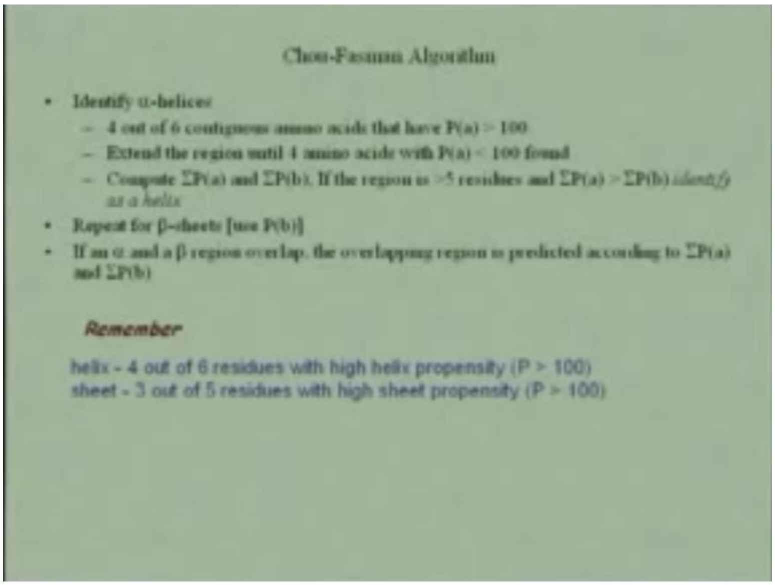
You have to remember that you have the protein sequence, always available to you. That is always available to you. You do not have the structure always available. So, if we want to construct a three-state model, we have the helix, the strand and the coil. So, we have basically an alpha, a beta and a turn. These are just some numbers. So, we need another table, if we need to go for secondary structure prediction. It was done in 1977.

It is a very famous way or very easy way rather to predict whether you have a helix or not. These are called the Chou-Fasman parameters. It tells you that the chance or rather the propensity, this value is called the propensity that you are going to have an alanine in alpha helix. The larger the number, the larger the probability that the helix is going to be in that specific secondary structure.

So, let us go through it once more. What this table tells you is that you see all the 20 amino acids here. It tells you that the numbers here tell you whether these amino acids are going to form as that is alpha helices, Bs that is beta strands or coils that is turns. Because, what do you want to do? You have your protein sequence. You want to know where your loops. This is L that is the another notation that is used apart from C.

So, you want to know where the coils or turns or loops are. You want to know where the helices are and you want to know where the sheets are, ok because that will give you a better idea of how your protein is going to fold, because you'll have some information from your hydrophobicity. You'll have some information about the secondary structure. So, that will lead you into a better idea of how your protein is actually going to form its tertiary structure from its amino acid sequence.

(Refer Slide Time: 37:09)



Chou-Fasman Algorithm

- Identify α-helices
  - 4 out of 6 contiguous amino acids that have P(a) > 100
  - Extend the region until 4 amino acids with P(a) < 100 found
  - Compute ΣP(a) and ΣP(b). If the region is >5 residues and ΣP(a) > ΣP(b) *identify as a helix*
- Repeat for β–sheets [use P(b)]
- If an α and a β region overlap, the overlapping region is predicted according to ΣP(a) and ΣP(b)

Remember

helix - 4 out of 6 residues with high helix propensity (P > 100)
sheet - 3 out of 5 residues with high sheet propensity (P > 100)

So, how do we do this? Again, we have the sequence of the protein. Form the sequence of the protein; you just have to look at the numbers. You don’t have to calculate anything now. You just look at the numbers. Out of 6 contiguous numbers or rather 6 contiguous amino acids if 4 of them have P(a) value greater than 100 then a helix is going to form. So, you are not doing any mathematics. You are just putting the numbers down.

So, let’s just go back to the previous slide. So, if I had a helix, you can just make up the helix. Let us look at this? We know whether helix begins here. So, MQGVVT. M is greater than 100, so, we have one greater than 100. What is Q? Q is glutamine and greater than 100. So, I have 2 out of 2 greater than 100. G is glycine, 57, less than 100. So, I am 2 out of 3. V is 106. 3 out of 4. Again V, 4 out of 5. T is threonine which is 83. So, I have to go a bit further.

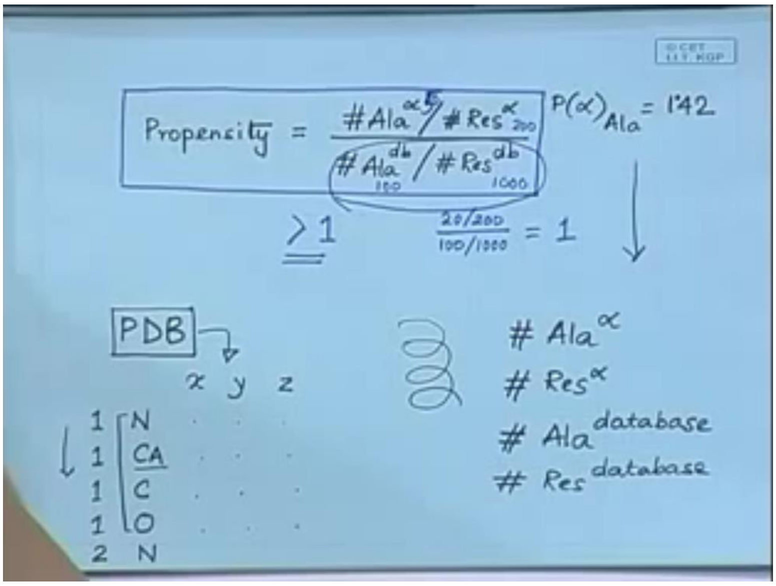
So., we have to have four that are greater than 100. Do I have four? I have MQVV out of the six MQGVVT. So, I can say that I have a helix here. What do I have here? It is HHHHHH. It is very simple. You just look at it. What about the next one? Then, we extend the region until 4 amino acids until P(a) less than 100 are found. So, you keep on doing that. So, all you have to do is you just have a slide, basically a sliding window again where you are looking at the window size of 6.

What is this 6 telling you? It is that if you have 4 out of 6 that have a P(a) value that is greater than 100, then you have an alpha helix. If your P(a) value is less than 100, you do not have alpha helix anymore. How do you look for a beta sheet? If your P(b) is greater than 100. 4 out of your 6 P(b) is greater than 100, then you have a beta sheet. So it's very simple. The problem arises when the alpha and the beta region overlap.

Then, what do you have to do? Then, you have to do some mathematics. Then, you have to sum the P(a) value of the six residues and sum the P(b) value of the 6 residues, whichever is higher it is going to be that. It is very simple and what information do you have? A lot of information. You have from the amino acid sequence; you can say whether you have helix or not.

So, now it'll make sense from the understanding whether you have a helix or not then you can construct a helical wheel and then say which part is going to be inside and outside. So, we are gradually getting to know more and more of the structure. So, we have a helix that has 4 out of 6 residues with high helix propensity. Now, I think I should explain to you what a propensity means before we go any further.

(Refer Slide Time: 41:12)



$\text{Propensity} = \dfrac{\#Ala^{\alpha}_{5} / \#Res^{\alpha}_{200}}{\#Ala^{db}_{100} / \#Res^{db}_{1000}}$
$P(\alpha)_{Ala} = 1{\cdot}42$
≥ 1
$\dfrac{20/200}{100/1000} = 1$
PDB
x y z
1 N
1 CA
1 C
1 O
2 N
$\# Ala^{\alpha}$
$\# Res^{\alpha}$
$\# Ala^{database}$
$\# Res^{database}$

Now, I am talking about propensity. It is sort of probability but you see that the numbers are greater than 100. In some tables, you might see like we have P(alpha) for alanine is 142. In some papers or books, it is sometimes put as 1.42. But in that case, when you do or when you try to predict whether you have a helix or not you'll be just looking for the values instead of 100 you look for 1.

That's as simple as that. Now, we are talking about the propensity value here. Now the way, Chou-Fasman got all these numbers was by what is called the statistical analysis on the structures that are available. What do I mean by the structures are available? The crystal structures that are salt for proteins are available in what is called the protein data bank. It is now freely available where you can download protein structures.

What do we mean by protein structures? If you have x, y and z coordinates for all the atoms except the hydrogen because x-ray crystallography cannot look at the hydrogen. So, if I am looking at residue 1, I am going to have a nitrogen for residue number 1. Residue number 1 will also have a C alpha atom associated with it. Residue number 1 will also have a carbon atom associated with it that is part of the carboxylic group.

Residue number 1 will also have an 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 the 3 dimensional spaces. So, the protein data bank gives these values for the 25000 structures that are available in it. So, now when I go to residue number 2 it starts again with the nitrogen, because I am going from the amino terminus to the carboxyl terminus.

So, then if you had a side chain, you would have apart from having C alpha you would also have C beta. So, the C beta would be written after this. So this would be the backbone, then you would have the C beta and so on. But, what we need to know is that there are a set of structures available for which you can do an analysis. What is that analysis? What you do for the analysis that was done here, say that you look at all the helices that are in the proteins and you count the number of alanine that are there in the alpha helix.

You count the number of residues in the alpha helix, all the residues that are there in the alpha helix only. Then, you count the number of alanine in the database including the ones in the alpha. So, you calculate all the alanines that are there. You calculate the number of residues that are there in the database, whichever database you are using. Propensity is a ratio. It is a ratio of the number of Ala in alpha divided by the number of residues in alpha to the number of Ala in the database divided by the number of residues in the database.

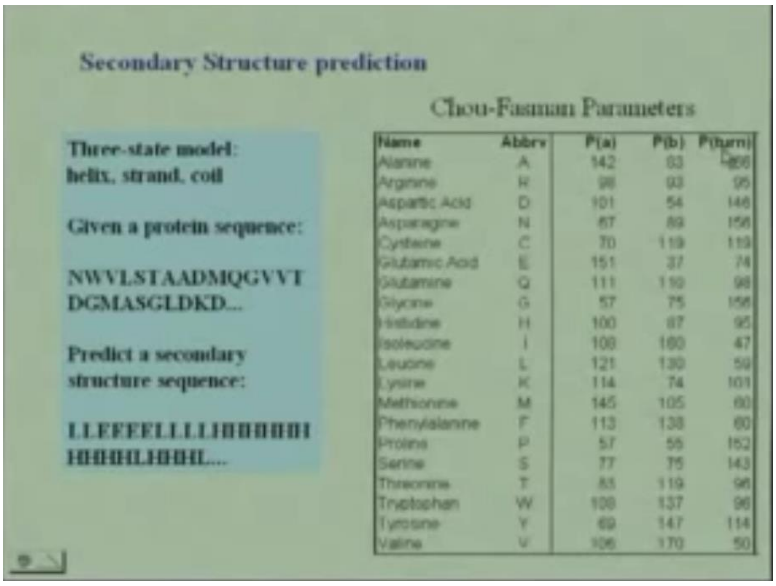
So, what does it tell me? This is your propensity calculation. If this number is greater than 1, because you have to remember that you are looking at a large sequence, a polypeptide sequence of a large set of proteins. You want to know whether alpha is preferred in helices.

What do I mean by that? If I look at this, this will give me some idea to whether alpha is preferred in the helices or not because this gives the number of alanine in the whole database.

If I have, say, 8 percent of alanine in the whole database, I can calculate these as a percentage. Say, I have 1000 residues in the database and 100 of them are alanine, that is possible. Of the 1000 residues in the database 200 are there in alpha helices of which 20 are alanine. So, what would my value be? It is 20 by 200 to 100 by 1000. What is it? It is 1. What does it tell me? There is nothing great that I have in helices.

I just have it, 10 percent as I have in the rest of the protein. I don't have any information about it, but if I had 50 of these of that were alanine, my value would be greater than 1. Then, in the normal case of a protein that I see here, I see more of alanine in helices, which makes it significant. So, that's how they came up with these numbers in our slide here.

(Refer Slide Time: 48:01)



Secondary Structure prediction

Three-state model:
helix, strand, coil

Given a protein sequence:

NWVLSTAADMQGVVT
DGMASGLDKD...

Predict a secondary
structure sequence:

LLEEEELLLLHHHHHHH
HHHHHLHHHL...

Chou-Fasman Parameters

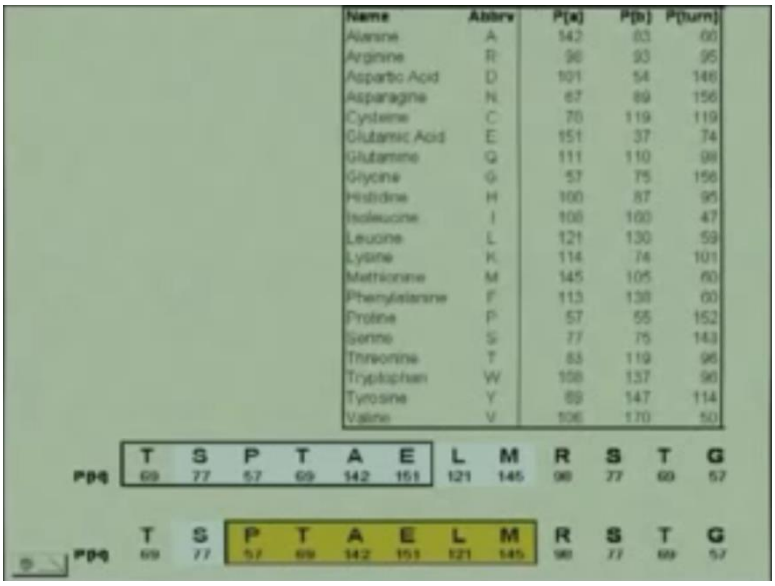
Name	Abbrv	P(a)	P(b)	P(turn)
Alanine	A	142	83	66
Arginine	R	98	93	95
Aspartic Acid	D	101	54	146
Asparagine	N	67	89	156
Cysteine	C	70	119	119
Glutamic Acid	E	151	37	74
Glutamine	Q	111	110	98
Glycine	G	57	75	156
Histidine	H	100	87	95
Isoleucine	I	108	160	47
Leucine	L	121	130	59
Lysine	K	114	74	101
Methionine	M	145	105	60
Phenylalanine	F	113	138	60
Proline	P	57	55	152
Serine	S	77	75	143
Threonine	T	83	119	96
Tryptophan	W	108	137	96
Tyrosine	Y	69	147	114
Valine	V	106	170	50

So, 1.42 is the way which you have in lot of tables. So 1.42 means that this number is greater than 1. It means that alanine would like to be in alpha helix, but let's think of a proline. All of know what a proline looks like. It'll break the helix. Why? That's because you cannot have a turn properly. It's a imino acid that bends on to itself. So the propensity of it to form an alpha helix should be very low. Look at the value. It's 57.

So, it is glycine which is very low. Why? That's because it doesn't like to be in an alpha helix or it is rather seen in alpha helix for the analysis that has been done, for this set of proteins, which is true for mostly all the sets. Now, if I look at a turn, turns have mostly glycine and proline. Glycine because of its flexibility and proline because it’s imino acid and it basically helps in the turn back of chain at times.

So, look at these numbers, 152 and 156 which are pretty high. Asparagine is also high. So that is how these propensity values were actually determined. The propensity values are determined again since for a very larger set of amino acids but this table is still used today for a rough prediction of where your alpha helix or beta sheet is going to be. So, what do we need? We just need to have that table to figure out where our helix is going to be and where our sheet is going to be and the rest of it is going to be coiled. There are turns also. You have a turn set also here as a P(turn) also.

(Refer Slide Time: 50:08)



Name	Abbrv	P(a)	P(b)	P(turn)
Alanine	A	142	83	66
Arginine	R	98	93	95
Aspartic Acid	D	101	54	146
Asparagine	N	67	89	156
Cysteine	C	70	119	119
Glutamic Acid	E	151	37	74
Glutamine	Q	111	110	98
Glycine	G	57	75	156
Histidine	H	100	87	95
Isoleucine	I	108	160	47
Leucine	L	121	130	59
Lysine	K	114	74	101
Methionine	M	145	105	60
Phenylalanine	F	113	138	60
Proline	P	57	55	152
Serine	S	77	75	143
Threonine	T	83	119	96
Tryptophan	W	108	137	96
Tyrosine	Y	69	147	114
Valine	V	106	170	50

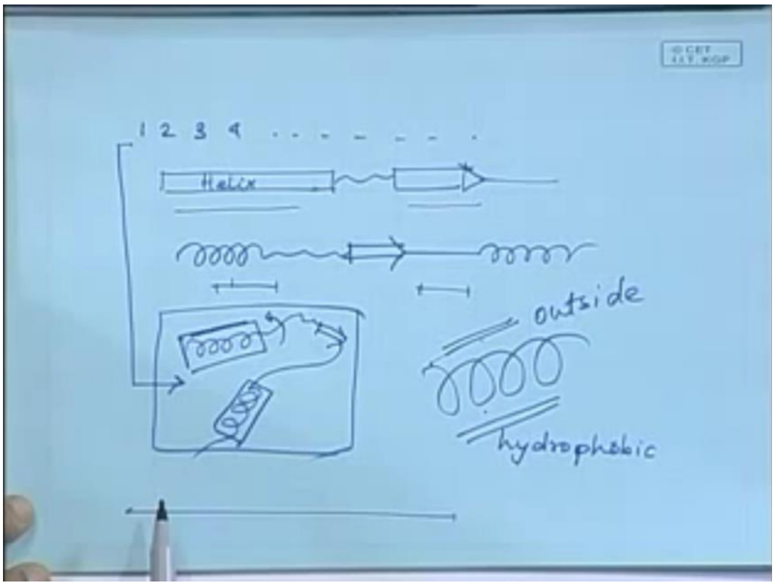
	T	S	P	T	A	E	L	M	R	S	T	G
P(a)	69	77	57	69	142	151	121	145	98	77	69	57

	T	S	P	T	A	E	L	M	R	S	T	G
P(a)	69	77	57	69	142	151	121	145	98	77	69	57

So, this is our table and this is our sequence. What do I have? I have my TSPTAE. What have I put in here? I have put in values. S is 77 and T is supposed to be 83 but it is 69 here. I do not know why anyway. So, what we have is threonine, serine, proline and so on and so forth. Now, when I am at this point, how many do I have that are greater than 100? It is just 2. So I cannot say that I have a helix formation. I slide my window down to serine.

I have an additional one greater than 100, but it is still 3 out of 6. It’s not good enough. So I slide it again. What do I have now? It is 4 out of 6. So, what can I say now? My helix begins. Your helix basically begins just after the proline here, but what you need to know is you have a sequence and what can you say from the sequence? You can say from the sequence and from the table where your helix is going to be, where your sheet is going to be and where your turn is going to be.

(Refer Slide Time: 51:57)



Then, based on that, what we can do, we can from this information roughly determine from the sequence of our proteins. I have my sequence 1, 2, 3, 4 and so on. What can I say? I can say that I have a helix here, I have a turn here and then I have a sheet here or something like that. So, I can say, basically it should be more like this. So, we can say what we have. Then what do I do? I can do a hydropathy plot.

What is a hydropathy plot going to tell me? It is going to tell me which parts of these are hydrophobic in nature. Say this part is hydrophobic and again this part is hydrophobic, so what can I say? I can say that this part is going to be inside and that is going to be outside and this part again is going to be inside. So, what do I have? Some information of how the protein is going to fold. So, I can go from my primary sequence to some idea of what it’s going to look like.

Then, I can also determine whether I am right if I construct a helical wheel for this now. I can also construct a helical wheel for this. What would I expect in this one? This face would be hydrophobic because you know that this turn can rotate basically and I can have rotation about this which would make either face in or out. Then, what would I have to do? Construct a helical wheel.

If this face is hydrophobic, I have to turn it around to make it come to the core of the protein. So, I have a hydrophobic region and a hydrophilic region. This is therefore on the outside. So, I am better off in determining how my protein is going to interact, how it is going to fold into giving me my final tertiary structure. What we learnt is we can from the Chou-Fasman parameter determine where we might have a helix.

We can determine from the hydropathy plot where we can have the hydrophobic regions or where we can also have transmembrane segments. We can from the helical wheel determine whether this part is outside or inside. So, we are better off than we are just the primary amino acid sequence of the protein. We will stop here today. Thank you.