

Bioinformatics
Prof. M. Michael Gromiha
Department of Biotechnology
Indian Institute of Technology, Madras

Lecture – 15a
Secondary Structure Prediction II

In this lecture we will discuss about the methods developed for predicting protein secondary structures. In the last class we discussed about regular secondary structures. What are the regular secondary structures?

Student: periodic.

Alpha helices, beta strands, turns right and irregular structure you can call as coil right.
How the alpha helix are formed?

Student: by i and $i+4$.

Right, by the hydrogen bonds between NH and CO groups right in the main chain right in $i+4$, then this alpha helix then you have 2 different types of helices. What are the 2 types of helices?

Student: 3/10 helix

3/10 helix and another pi helix right then the beta strand right, you can also see the hydrogen bonding pattern between the NH and CO groups. Now, 2 types of beta strands, one is parallel and the other is antiparallel right and you can see the difference in the hydrogen bonding patterns in the parallel and the anti parallel beta strands. Then we discussed about turns, 3 different types of turns and type 1 turn is quite frequently occurring right in protein structures.

Then we discussed about Ramachandran plot right. What is Ramachandran plot?

(Refer Slide Time: 01:22)

Refresh

Protein secondary structures
Ramachandran plot
Dictionary of secondary structures of proteins
Statistical analysis (Propensity)
Chou-Fasman method

$\frac{N(\alpha)/\beta}{N}$

Student: It is plot between phi and psi.

Plot connecting phi and psi right and he used the 2 dihedral angles phi angle, psi angle right. How many atoms are required to get the dihedral angle?

Student: 4.

4 atoms right, because it is angle between 2 planes right atoms 1 2 3 and 1 plane 2 3 and another plane. So, angle between these 2 planes right that is called dihedral angles right.

So, here in the protein main chain you can see the 2 types of dihedral angles one is phi and one is psi, phi is a rotation along.

student: CN.

N and $C\alpha$ right, N- $C\alpha$ right and psi is $C\alpha$ -C.

Student: $C\alpha$

Say another one that is a peptide one. So, Ramacahndran he tried to use these atoms as the hard spheres and see the rotations which are allowed right which avoid the steric hindrance. So, based on that he identified the regions where the alpha helices are populated right as well as where allowed regions for the beta strands. So, he develop the plot right and that plot is called the Ramachandran plot. Then if you know the structures

of proteins right based on the hydrogen bonding pattern, Kabsch and Sander right they developed algorithm. Which algorithm?

Student: DSSP.

DSSP right, dictionary of secondary structures proteins, to assign the secondary structures for each residues. Then if you have the experimental data then we can utilize the information to develop different models for predicting the secondary structures right there are several methods we discussed right. So, one of the earliest one that is based on the statistical analysis, Chou and Fasman first he tried to analyze the preference of amino acid residues in each secondary structure. Which is the values he developed?

Student: Propensity.

Propensity right how to get the propensity values.

Student: So, the division of

Right, first for any residues right. So, in any specific confirmation right alpha helix divided by n_i , number of residues in the protein right then you can see the N_α that is the number of residues right this is to N . So, this is for any residue i , here this is for the whole protein to the any residue i it prefers to alpha helix then you do a higher values right depending upon the total number of residues in alpha helix.

So, correct. So, likewise you can get the value for the any specific residues this is only this is for i stands for 20 residues right, you can get the values for 20 residues. From that information you can see the residues right whether the propensity is more than one or less than 1, if it is more than 1 then it is preferred in alpha helix, if less than 1 it is not preferred in alpha helix.

So, today we will discuss about the other methods I discussed about the 5 different methods I mentioned. What are other methods?

(Refer Slide Time: 04:24)

Information Theory: GOR Method (Garnier-Osguthorpe-Robson)

• Information (i) for each residue

Central residue, 8 neighbors on each side (window length of 17 residues); 4 states (helix, strand, turn and coil)

$\begin{matrix} & \text{A} & \text{E} & \text{I} & \text{V} & \text{K} \\ \text{H} & & & & & \\ & \text{H} & \text{H} & \text{H} & \text{H} & \\ \text{E} & & & & & \text{E} \end{matrix}$

Student: Information.

This information theory, this is called the GOR method and hydrophobicity profiles.

Student: MSA.

Multiple sequence alignment neural networks right and the consensus method right the joint methods you can say that. This structure we will discuss the different other methods to predict the secondary structures and the performance of these methods.

With the Chou and Fasman method, is a statistical analysis you can identify the sequence, but the accuracy is about 60 percent 55 to 60 percent. So, here they considered the propensity of each residue to be in any secondary structure either alpha helix or a beta strand. Now, with that information Garnier, Osguthorpe and Robson right they try to use information for the central residue as well as the neighboring residues. Why it is required because there are some segments say any pentapeptides or hexapeptides or even tetra peptides, the same peptides they can adopt different secondary structures right. For example, if you see a segment AEIVK, they can be in helix, they can also be in the strand. In this case if you use this segment a high propensity in helix is in strand this can be difficult to distinguish.

So, Garnier, they try to get the information for the central residues as well as the residues which are occurring nearby the central residue. So, they collect the information for

example, if this is the central residues, they consider the propensity of that particular residue isoleucine in helix for example, and see the neighboring residues about the propensity of the neighboring residues on both sides of the central residues.

So, they derive some sort of frequency matrices and use the information for prediction right because they use several information they call this theory as the information theory because they develop the method based on the information of that particular residue. They consider central residues and they use 8 neighbors from left side and then right side right this may number of driving residues depends we can use any window length and see finally, which window length provides the optimal highest performance right then you can fix it right, you can use it 3 5 7 several window lengths right and can optimize to get the window lengths with the highest performance. So, they use 8 neighbors with 8 neighbors. What will be the window length?

Student: 17.

17 right because 8 plus 1 plus 8 right, this is a central one, this is on the left side this is turned the right side right N terminal C terminal side. So, 17 residues and they consider 4 different states right they consider helix, strand, turn and coil. So, 4 different states they consider, how they derived information. So, they develop that formula to obtain the information content for I is information content.

(Refer Slide Time: 07:31)

**Information Theory:
GOR Method (Garnier-Osguthorpe-Robson)**

Information content

$$I(SS_i=X; \sim X; aa) = \ln\left(\frac{P(SS_i=X|aa)}{P(SS_i=\sim X|aa)}\right) - \ln\left(\frac{P(S_i=X)}{P(S_i=\sim X)}\right),$$

↑
Helix
not in helix
wt+
not helix

$\underline{SS_i} \rightarrow$ secondary structure at position i in the sequence

$\underline{X} \rightarrow$ any secondary structure: helix (H), strand (E), turns (T) and coil (C)

$\underline{aa} \rightarrow$ any amino acid residue

For any secondary structure right and any particular residue right, here you can see any residue i , this can be any secondary structure for example, this X this can be helix right. So, preference of that particular residue in helix over the same residues which is not in helix right this is the negation there is, which one is the residues which are not in the particular secondary structure.

So, how to get this information they use this, this particular formula there is a logarithmic of this the probability of any particular residue to be in helix right this is for example, if you take the helix and the probability of the residue which is not in helix, minus because this is considered only for that particular residue. Now, we have, like Chou and Fasman, we have to normalize with the whole protein. So, in this case the whole protein they take the residues in helix and the residues which are not in helix.

So, for any residue with any particular secondary structure they try to use the information and whether that residues right prefer to in particular secondary structure over the other secondary structures as well as that secondary structure over now other than that particular secondary structure. So, here you can see this a SS_i is a secondary structure at a position i in the sequence and X is any secondary structure you can do it for 4 different secondary structures right helix, strand, a coil and the turns, and aa stands for any amino acid residues that you can use this equation to get the preference of say any residues. So, how they derive the probability right which information required to derive the probability?

Student: Occurrence.

We need experimental data right because if you have only the sequences right if you know only the UniProt sequences then we do not know whether they residue alanine prefers even helix or strand we do not know in this case we need experimental data. So, they collected the experimental data with known secondary structure information. For example, the structures are known then you can use the DSSP it can take the hydrogen bonding information and the assign secondary structures. DSSP uses 8 classifications, but we can simplify into 4 classes based on by combining all the helices as one helix, like alpha helices, pi helices and the 3/10 helices makes one group called a helix.

So, they consider the proteins with known secondary structures right with the second is assignment right then we know that assignment like you know sequence and for each

sequence we know the secondary structure assignment right particular residues belongs to helix or strand or coil or turn then we calculate the frequency of each amino acid in each of the window portions.

(Refer Slide Time: 10:25)

GOR Method

Derivation of probability:

- Proteins of known secondary structure were collected
- Calculated the frequencies of each amino acid in each of the 17 (-8 to 8) window positions in helix, strand, turn and coil
$$\begin{matrix} & \text{(17 X 20 matrix)} \\ \downarrow & \swarrow \\ \text{positions} & \text{amino acids} \end{matrix}$$

For example, if you have the window portion 17 what is the preference of residue at x, x-1, x-2, x-3 and so on, likewise x+1, x+2, and x+3. So, they have 17 residue positions and calculate the frequency of each amino acid for each secondary structures right. So, they can develop some matrixes. So, you can see the 17 by 20 matrix, here 17 stands for?

Student: Window.

These are positions right 17 positions and 20 stands for.

Student: 20 amino acids

20 amino acid right, they can calculate.

So, now they use this matrix right to calculate the probability as well as information content. Because this matrix contains information about the preference of any particular residue at different positions, 17 positions right. So, this matrix can be used to obtain the probability of any new residue to be in particular position depending upon the secondary structure.

(Refer Slide Time: 11:14)

GOR Method

- The matrix was used to calculate the conditional probabilities and information content

- Total Information content for each secondary state is calculated from $\sum_{aa} I(SS_i=X:\sim X; aa)$ and secondary structure is assigned based on the maximum score.

$$\begin{matrix} H & x_1 \\ E & x_2 \\ T & x_3 \\ C & x_4 \end{matrix} \rightarrow \max(x_1, x_2, x_3, x_4)$$

So, now, they calculate total information content for each state, for example take helix they use is information. What is a preference of that particular residues to be in helix compared to be not in helix? They do the same for strand and you can do the same for the turn and the same for the coil then when you will how many values you will get.

Student: 4 values.

4 values right, if you have the 4 second secondary structure helix strand turn or coil. So, we get 4 different values. Here will get x_1, x_2, x_3, x_4 and they take this 4 and from this, which one is the preferred secondary structure.

Student: Maximum.

The maximum, maximum of these x_1, x_2, x_3, x_4 right this will give you the preferred secondary structure right this is the principle used in Garnier method. I will explain with one example.

(Refer Slide Time: 12:31)

GOR Method						
	Helix	~Helix	Total	P(SS=H aa=A)	$\frac{210}{300} = 0.70$	$H \nearrow A \nwarrow$
Alanine (aa=A)	210 19	90 19	300	P(SS=H aa=A) $\frac{210}{300} = 0.70$	$\frac{210}{300} = 0.70$	
All residues	810 900	990 900	1800	P(SS=H) $\frac{810}{1800} = 0.45$ P(SS=~H) $\frac{990}{1800} = 0.55$	$\frac{810}{1800} = 0.45$ $\frac{990}{1800} = 0.55$	
$I(SS=H:\sim H; aa=A) = \ln(0.70/0.30) - \ln(0.45/0.55)$ $= 0.847 - (-0.20) = 1.047 \rightarrow \text{Preferred}$						

For example if you have a protein with 1800 residues or a set of protein with 1800 residues, 810 are in helix and 990 are not in helix; that means, 990 are spanning to different secondary structures either strand or coil or turn. Then you take in a particular see in alanine. So, that 300 alanine right this for example, right this is not exact number this is I put for example. So, 210 are in helix and 90 not in helix. Using this information we can derive the information content right for example, the probability of particular residue alanine in helix right, this is alanine in helix right. What is the probability, totally how many residues?

Student: 300.

300 alanine right, they we are talking about alanine, so 300 alanine. How many alanine in helix.

Student: 210.

210. So, this you can calculate the ratio this will 0.7 then you can see the same residue alanine which are not in helix maybe it is not in helix it is negation. So, you can use it how many helix how many alanine or not in helix.

Student: 90.

90 right, they for 90 divided by total 300 residues. So, here you can calculate the probability of alanine which are not in helix just 0.3 right. Now, this is for a particular residue. Then go with the full residue, full proteins right all the residues right. So, here if it say any secondary state helix. How many residues are in helix?

Student: 810.

810 right if you take 810. So, normalize with the total residues 1800 this will be 0.45. Now, that the, how many residues which are not in helix. So, you can see 990 divided by 1800 this 0.55. We use this information right to get the probability of alanine to be in alpha helix this is called information content. So, you can see the probability of alanine in helix which is given as logarithmic of 0.7 by 0.3 this is for alanine and here this is for the whole protein.

So, now if you calculate this, this equal to 0.847 minus this logarithmic is -0.2. So, if you subtract then finally, we get the value of 1.047. From this number we can say whether this alanine prefers to be in helix or not, say prefers to be helix or not.

Student: Yes.

Yes right. If the value is more than 0 then it prefers to be helix if it less than 0 is not to prefer in helix. For example, if you have 300 residues 300 alanines, 150 helix and 150 is not in helix, likewise 1800 you can put 900 and helix and 900 not in helix right. For example, if you put this as 900, this as 900 here also if you put 150, 150. So, then what is the value we get.

Student: 0.5.

Logarithmic of this will be.

Student: 0.5.

Right 1, right minus ln of 1 right, this will be 0.

So, if there is equally preferred that is not just 150, 150 if the ratio is same right is not the exactly same number right. For example, here 100 and 200, likewise here also if you see the ratio of 2 is 3 right, a fine. So, if you do this then we get the value of 0 this will tell you that for if it is randomly distributed; that means, the distribution is same the full

protein level as well as any residue level then will get value of 0. So, if it is more than 0 right then we will see is of preferred ones, if less than 0 right you can see this not preferred in particular secondary structure.

(Refer Slide Time: 16:29)

Amino acid residue	Residue position (conformation)								α -helix								
	$j - 8$	$j - 6$	$j - 4$	$j - 2$	j	$j + 2$	$j + 4$	$j + 6$									
Gly	-8	-10	-15	-20	-30	-40	-50	-60	-15	-10	-5						
Val	0	10	15	20	30	40	50	60	0	10	5						
Leu	0	5	10	15	20	25	28	30	14*	10	5	0					
Ile	5	10	15	20	25	20	15	10	0	-10	-5	0					
Ser	0	-5	-10	-15	-20	-25	-30	-35	-39	-30	-25	-11	-10	-5	0		
Thr	0	0	0	-5	-10	-15	-20	-25	-26	-25	-20	-15	-10	-4	0	0	
Ala	0	-5	-10	-15	-20	-15	-10	0	5	10	20	25	20	15	10	5	0
Gln	0	0	0	0	10	20	30	40	76	73	78	73	70	60	40	20	0
Glu	0	0	0	0	5	10	20	30	40	30	20	10	0	0	0	0	0
Lys	0	40	50	55	60	60	56	50	42	18	5	0	0	4	0	0	0
His	10	20	30	40	50	50	50	50	12	-20	-10	0	0	0	0	0	0
Arg	0	0	0	0	0	0	0	0	-9	-15	-20	-30	-40	-50	-60	-30	-10
Phe	0	0	0	0	0	5	10	15	15	10	5	0	1	0	0	0	0
Tyr	-5	-10	-15	-20	-25	-30	-35	-40	-45	-40	-35	-30	-25	-20	-15	-10	-5
Trp	10	20	40	50	55	50	40	30	12	10	0	10	20	30	40	50	10
Cys	0	0	0	0	0	0	0	-5	-10	-12	-10	-5	0	0	0	0	0
Met	10	20	25	30	35	40	45	50	53	50	45	40	35	30	25	20	10
Pro	-10	-20	-40	-60	-80	-100	-120	-140	-77	-60	-30	-20	-10	0	0	0	0

So, here I show the matrix one of the examples. So, this is one matrix for the alpha helix that you can see this is alpha helical conformation. So, this is a central position j. So, how many window lengths they used.

Student: 17.

17 right. So, you can have the values 1 2 3 4 5 6 7 8 likewise here also 8. So, 8 plus 8 16 plus central 1 this equal to 17, this varies from j-8 to j+8 for example, if this is a sequence AIKTSV right. So, if you take this is a central residue j, this is j+1, this is j-1 and so on right. You can see the 8 7 residues this side and 7 residues, 8 residues here and 8 residues there, is a matrix.

If you see the matrix we can see the preference of some specific residues for assemble to the jth position right; which residues are preferred in j. So, if you see the plus that's preferred one you can see the plus values here right. So, it is plus value corresponds to alanine this is 65. So, here is 14 this corresponds to valine and here 32 this corresponds to leucine. So, you can see the glutamic acid here this is glutamic acid. So, this is 78 glu. So, from this one which residues are preferred at the central position? Alanine, valine.

Student: Valine.

Glycine, glutamic acid, this is similar to the propensity obtain by.

Student: Chou-Fasman

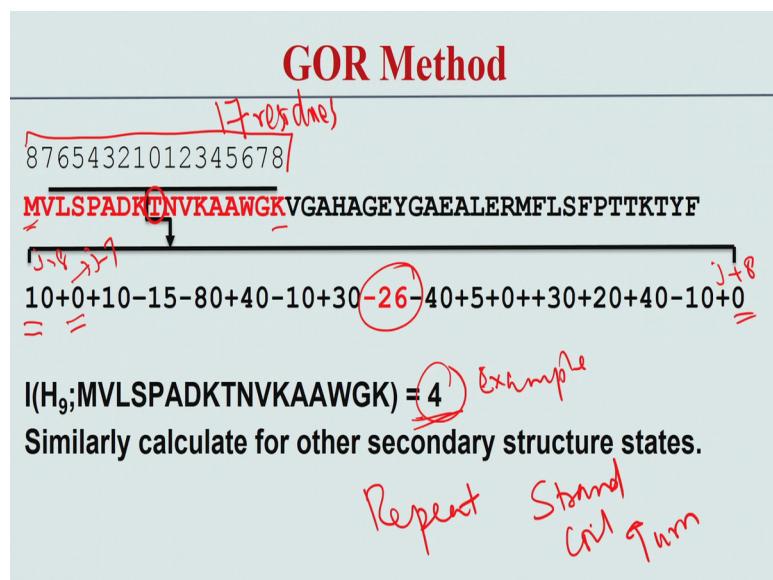
Chou and Fasman, yesterday we discussed the previous class we discussed about the propensity of residues right. You could see the preference of these residues say glutamic acid, alanine, valine, leucine they prefer to be non alpha helix right. In this case the preference agrees with the Chou Fasman method. So, the data seem to be fine.

Now, here they added more information regarding neighboring residues for example, $j-1$, $j-2$ and see some cases is they do not have any preferences and some cases you can see the preferences. Say glycine is not preferred at any position right totally random, we see alanine is preference at the j as well as other positions and some cases you can see the preference at some positions are not in the other positions. So, you can use this matrix to get the information. So, how many matrixes they derived; this is for helix.

Student: 4.

4 matrixes right one for helix, one for strand, and turn and coil.

(Refer Slide Time: 19:09)



So, now you have the sequence. So, this is amino acid sequence and how we obtain the information content using Garnier, and how they transfer the data to predict the

secondary structure. For example, if you take the central residue threonine. So, these are the different windows you can see 17 residue windows, from here to here right. So, the 17 residues and you substitute the values from the table. For example, if alpha helix for T, T central one. What is the value for threonine? T is here this is a central value this is this equal to -26 right, you got -26 here. Then take this j-8 right j-8 which residue methionine. So, go with a methionine methionine here this j-8 this is 10, put the value 10, this is j-7, j-7 is valine if you see here the valine j-7 that is equal to 0, that is equal to 0. Then go with the j+8 j+8 is lysine. So, if we see the lysine what is the 3 letter code for lysine?

Student: Lys.

Lys right this is here. So, lysine is here. So, these value is 0 right, it 0. So, we have 17 values for the 17 positions right that is the difference between Garnier, and Chou and Fasman. Chou and Fasman use only the 3 on in right. So, I get the values and sum of all the numbers then we will get the value of the in a window length right. So, this is not the correct one. So, this for example, you put +4 right this for example.

So, if you out of these numbers right finally, for example, if you say the value is +4. Now, it prefers to be for helix, but we do not know whether the prefers alpha helix compared with others second restructures. For example, if we get the beta sheet 15 or 20 right then; that means, the preference for beta strand is more than alpha helix. So, in this case what we have to do.

Student: We have to repeat.

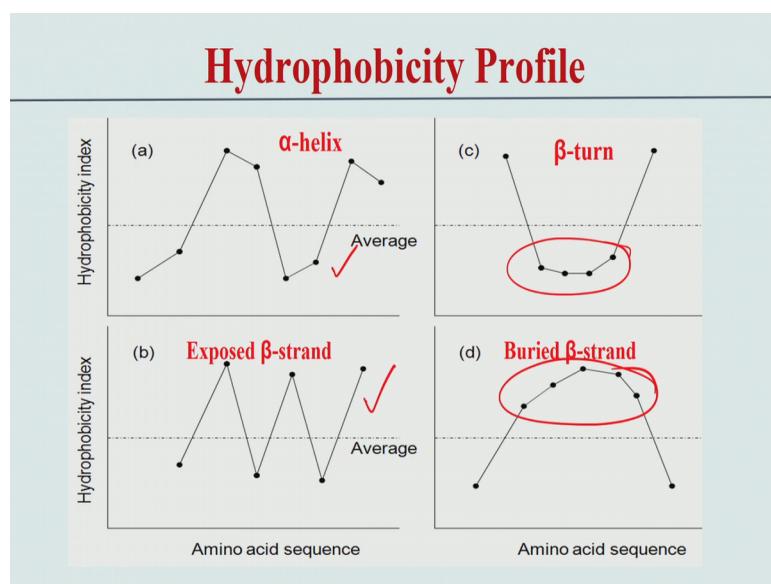
We have to repeat, for different secondary structures this for the helix, let do for the strand, coil and turn. So, we get 4 numbers. There for among the 4 numbers choose the best and assign these residues is an alpha helix, likewise you can assign the values for all the sequence and we get based on the preference you cannot say in different secondary structures. And in the end we can do some end corrections right for example, there are 6 or 7 residues which is continuously predicted helix in between there is a strand, then you can take this whole segment has an helix that we can make some end corrections compared with the experimental data right to improve the performance of the method this weather also I can predict with an accuracy of for about 65 percent.

So, then the other methods, we can talk about hydrophobicity profiles. So, what is hydrophobicity profile? We discussed earlier classes. What is the hydrophobicity profile?

Student: Plot connecting amino acids to hydrophobicity values.

Is a plot connecting amino acid residues with respect to hydrophobicity values, you can use any hydrophobicity scales I will discuss about a particular scale right. Here you can see the hydrophobic residues have more values and the polar residues have less values. So, new plot these values right we can see a pattern right they identified some specific patterns for helix or stand or coil or turn as we know the helices 3.6 residues per turn and if this residues are in amphipathic, right spanning in 2 different phases right then 2 in one phase and 2 another phase based on that they derived a specific type of patterns.

(Refer Slide Time: 22:58)



This can work only if the helices are amphipathic in nature. The helices are not amphipathic then we cannot see this pattern in this case you can the methods will fail right in this case we need to add more information right. But if the helices are amphipathic in nature then we can easily identify the helical segments based on the patterns. Generally in all protein structures about 70-75 percent of residues are amphipathic right, the helices. So, they could be able to identify the segments.

Then we go for the beta strand. Beta strand, there are two different types of beta strand some beta strands it is completely buried right because the residues which are buried are

highly hydrophobic in nature. So, the mainly if you compare to the helices and strands, strands are more hydrophobic than helix. So, these residues which are strand they are highly hydrophobic. And second one the beta strands are also in amphipathic nature in this case you can say 2 phases and one phase you can see the hydrophobic residues and the other phase you can see the hydrophilic residues right. If it is amphipathic nature or the exposed beta strand you can see pattern which are shown in this figure right this is the pattern you can see.

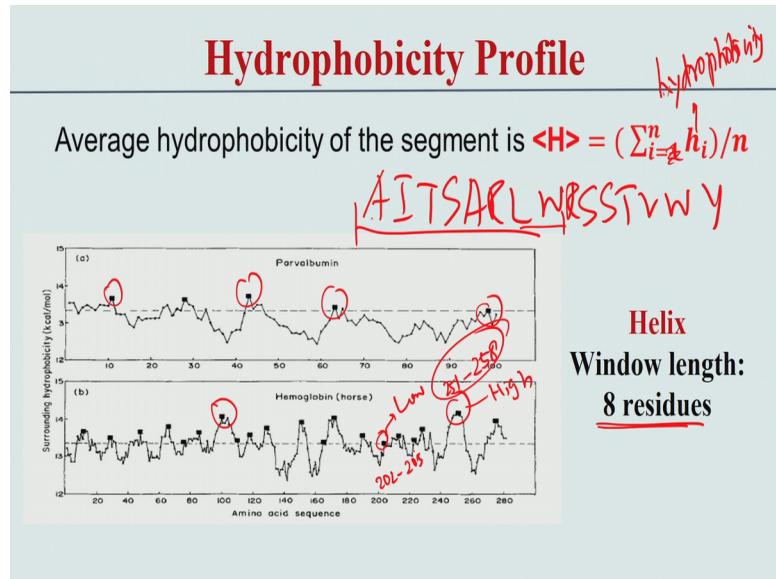
On the other hand if the residues are in the buried right and adopt the beta strand conformation you can see a pattern which are shown in the figure d right, we can see most of the residues which are highly hydrophobic. So, now, if you have the patterns for helix there are two different patterns for the strand based on the location of the strands whether it is in helix whether is not helix is whether in the exposed beta strand or at the buried beta strand.

Now, there is another case just for the turns, turns are generally hydrophilic like then, like loops because they accommodate with the polar residues right in this case you can see a stretch of residues which are less hydrophobic. So, if they made these type of graphs based on the known information and the physical basis of how these proteins fold and what are the preference of these different residues in secondary structures this is not a statistical analysis, it is based on the characteristics of the residues in protein structures.

So, if you have any sequence right for example, if we have the sequence, now we can make a profile and if you have the profile then from that profile you can identify the regions which have any of these patterns. Now, if you identify any specific patterns then you can say this could be helix and this could be strand and this could be the coil or the buried beta strand or exposed beta strand and so on. This is one way there also different methods which can be used for predicting the second the structure based on the average hydrophobicity right because this is a single residue hydrophobicity and you can see some stretch of residues as we discussed in earlier classes, there is one residue which are having different value. So, then you do not get the patterns the pattern will those.

In this case you can get the average values and try to use the average value for predicting the second the structures right. So, we will explain how to do that. For any sequence for example, if you have a sequence you can make different window lengths.

(Refer Slide Time: 26:14)



For example if it takes window length of 8 residues take 1 to 8, 1 2 3 4 5 6 7 8 and the value is equal to i equal to 1, to n , h_i by n , h_i by n . What is h_i ?

Student: Hydrophobicity.

It is a hydrophobicity value. So, you get the for each residue we have the value it is here assign the values and add up the values and divide by n . So, n equal to this segment this 8 residues segment n equal to 8 right you can do that. So, we get the different values.

And you can use these values to put in secondary structures how to do this and how to assign these segments. In this case we see a protein and if you get the amino acid sequence its possible to see whether this residues can contains more alpha helices or more beta sheets are the mixture of helix and strands. How to do this? If we have a sequence and we have the pattern right this is a pattern for the helix and see how many pattern the segments which has this pattern right take any sequence and make a plot and see how many segments we have this type of patterns. Then already we discussed about the preference of residues, Chou and Fasman method. So, we know that some residues are preferred in alpha helix. If you take a minimum of 4 residues that is corresponds to one turn and say all the 4 residues are accommodated with the preferred residues like alanine, valine, glutamic acid.

So, how many segments you can get the preference of residues and how many overlaps with the particular profile and the preferred residues. They could get many patterns then you can see this protein is mainly having lot of helices. In this case it is fine to identify the helical segments instead of doing all the secondary structures, this first one and the second case they will also do the same pattern for the strands and see how many segments contained exposed beta strands and buried beta strands and uses Chou and Fasman parameters. So, we know the preferred residues in beta strand take only preferred residues and see a segment of 4 residues and then see how many segments you can identify using preferred residues then we get the number of segments, comparing these two if it is highly dominated by helical segments we see the protein contains mainly helices.

If this contains mainly more number of beta strands right and you can see this protein contains mainly beta strands. And some cases we will have both 10 strands segments and then 15 beta strand, a helical segment. In this case you can see these protein contain the mixture of both right this is how we will see. Then if we can see a protein which dominated by helices then we see that there will be many several helical segments.

So, to identify the helical segments what to do first we get the window length of 8 residues and construct a plot. You can see these are 2 different examples for 2 different proteins right. So, here X axis is the sequence 1 means that means, 1 to 8 average of 1 to 8; 2 means average of 2 to 9 and the y axis is the hydrophobicity value average value. And if you make the plot you can see there are several peaks right, several peaks and if you see these peaks some of them are very high from the average. For example, this is high peak and some peaks which are here just the low just about the average, low peaks.

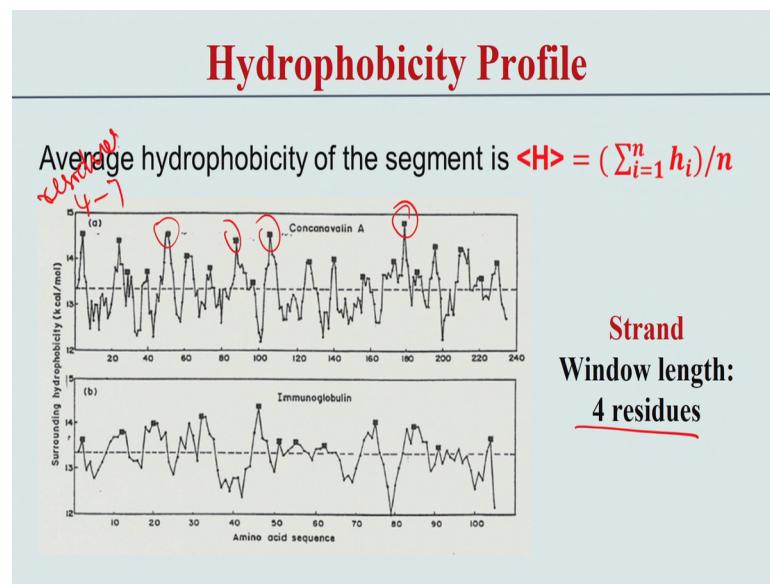
So, if you see these twp peaks and compare the known structures you see there will be a longer segment which corresponds the high peaks and shorter segments of helices in the lower peaks. So, then we see for example, you this is a region with 251 for example, then you can see that this is a means for to 251 to 258 this belongs to helix. The lower peak for example, this is 202 it corresponds a shorter helix say 4 residues. So, is a 205 likewise we assign the secondary structure. And then we can make the modifications to refine the n segments based on the preferred residues in n, at the N position and the C positions as well as the patterns where they have right this type of patterns, from that you can identify the regions which you belong to helical segments.

So, what first what we have to do? First you have to see whether the proteins are dominated with the helices or dominated the beta strands or contains both right. How to do that?

Student: take 4 residues

Take the 4 residues segment and see how many segments you can identify using preferred residues as well as the patterns in helices. Check for helix check for strand. For example, helix has 25, strand has 2 then this protein belongs to the alpha helix protein you can predominate the alpha helix. If beta strand is 20 and helix is 3 then we can say these proteins mainly dominant in beta strand if helix equal to 10 strand equal to 15 right or vice versa right then you can say it contains both mixture of that. If it is helix then we deal with the helical segments we do not bother about the strand. So, get the average profile make into two groups one is a high peak and low peak, high peak strands for longer segments low peak strands for the shorter segments and we can find n corrections make the helical position, regions. So, this is your example for the beta strand.

(Refer Slide Time: 32:07)

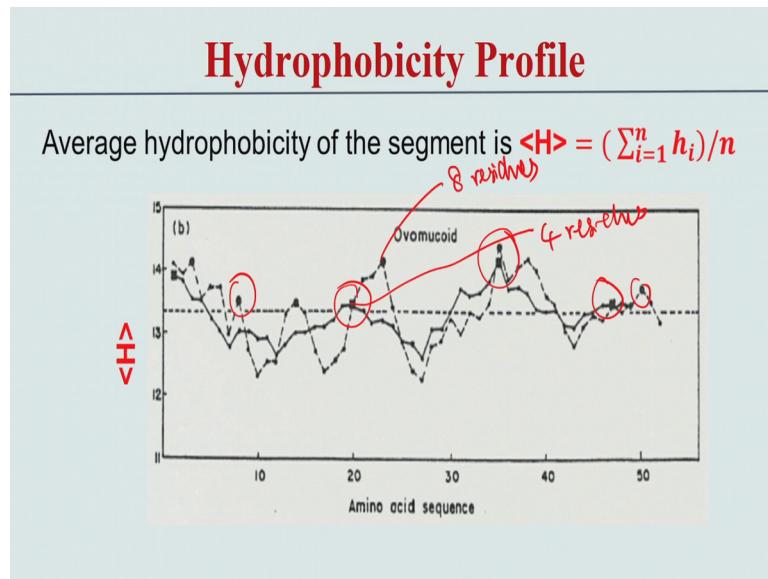


So, here we use window length the 4 residues, because short beta strands are frequently occurring in protein structures they this way we take 4. Again to make this plot and see there are several peaks. So, each peak represents a beta strand. For example, in this case it is 1 2 3 4, if it is 4 then 4 to 7 that residues 4 to 7 because 4 residues segment these corresponds to a beta strand and then we can make the end correction based on the

preferred residues in the N terminal or the C terminal as well as the patterns right if you have any patterns then you can extend the pattern till the pattern ends.

So, likewise we can make the end corrections to identify the beta strands. So, if with contains both helix and strands right, we have to made the figure one for 8 residues which one is for 8, this for 8 residues and you can see the small one right this figure, this figure that is for the 4 residues; one is for the helix one is for the strand.

(Refer Slide Time: 33:05)



So, then we see the patterns many cases it is distinct one is here, one is helix is here and this and the strand somewhere, helix and strand, different distinct places. So, if do so, we can discriminate the occurrence of helices and strands, we can see some overlap between 4 residues 8 residues segment. Then you can get the values and compare the numbers and say whether this can be a helix or the strand right. You can get this segment and see the average value, based on the average value we can decide whether this can be a helix or this can be a strand. So, now, there are 3 different ways based on the classes. So, that you can predict the secondary structure using hydrophobicity profiles.