

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

**Lecture - 14b**  
**Secondary structure prediction I**

So in 1983, Kabsch and Sander right, they developed a program called the DSSP right this is dictionary of secondary structure of proteins, this why they called as DSSP. To analyze the known 3D structures they are not doing any prediction, they are taking the real structures from the protein data bank and they estimate, they just check the hydrogen bonding pattern and then they have the phi psi angles.

(Refer Slide Time: 00:40)

# DSSP:

## Dictionary of Secondary Structures in Proteins

Sequence and secondary structure for 4MBN chain A

*Sequence*

1  
51  
101  
151

```
VLSGEWQLV LHYNAKVEAD VAGHGQDILI RLFFKSPETI EKFDYFKHLK
HHHHHH HHHHHHHGGG HHHHHHHHHH HHHH HHHH HT GGGTI

TEAMKASED LKKGHTVLIT ALGAILKKGK HHEAEALPLA QSHATKKHIP
SHHHHH HH HHHHHHHHHH HHHHHHHHHH HHHHHHH HHHHHS

IKYLETISEA IIVLHPSHP GPFGADAQGA NNKALELFRK DIAAKYKELG
HHHHHHHHH HHHHHHHH G GGS HHHHH HHHHHHHHHH HHHHHHHHHH

YQG
```

*Sequence*

*Secondary Structure*

*DSSP*

*4*

5	A	G	H	>	S	+	0	0	35	2,-0.2	4,-1.6	1,-0.2	-1,-0.2	0.823	107.4	48.1	-63.8	-34.5
6	A	E	H	>	S	+	0	0	51	2,-0.2	4,-1.8	1,-0.2	-1,-0.2	0.883	109.7	52.9	-77.1	-34.6
7	A	W	H	X	S	+	0	0	15	-4,-2.7	4,-2.6	2,-0.2	5,-0.3	0.894	105.3	56.2	-63.6	-34.4
8	A	Q	H	X	S	+	0	0	133	-4,-2.1	4,-2.5	1,-0.2	5,-0.2	0.938	107.3	48.2	-56.7	-47.2
9	A	L	H	X	S	+	0	0	55	-4,-1.6	4,-1.5	1,-0.2	-1,-0.2	0.855	112.8	50.0	-60.7	-40.5
10	A	V	H	X	S	+	0	0	0	-4,-1.8	4,-2.0	2,-0.2	-1,-0.2	0.917	114.6	40.0	-65.7	-50.1
11	A	L	H	X	S	+	0	0	44	-4,-2.6	4,-2.2	2,-0.2	-2,-0.2	0.842	107.7	61.6	-78.6	-27.0
12	A	H	H	X	S	+	0	0	120	-4,-2.5	4,-0.6	-5,-0.3	-1,-0.2	0.965	109.4	43.1	-64.9	-40.0

M. Michael Gromiha, NPTEL Bioinformatics Lecture 14

From these patterns they assigned each residue to several 8 different secondary structures. There are 3 different types of helices right what are 3 different helices we discussed?

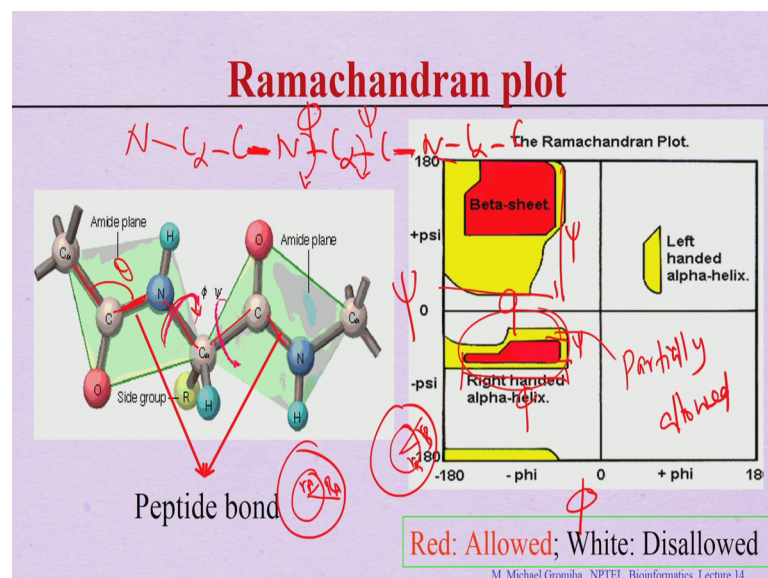
Student: 3/10 helix

Alpha helix and 3/10 helix and pi helix, and 2 types of strands right and the bridge and 1 bent, 1 turn right and 1 coil, 1 is irregular. So, they make 8 different categories of secondary structures. So, here this is the output of the DSSP, here they give the chain information this is a chain information and here this is the amino acid sequence, and this

is the secondary structure assignment, and here they have this hydrogen bonding patterns. If it is the residues number they put 2 4 1 -1; that means, plus or minus they forward the direction and the -1 in a reverse direction, they put which residues are hydrogen bonded with respect to this particular residue right and they get some energy values also this what they get the energy values.

Then finally, here they get the phi psi values this is phi, this is psi right you can see the almost residues are within that range right if you see this range right and here if you see the map ok.

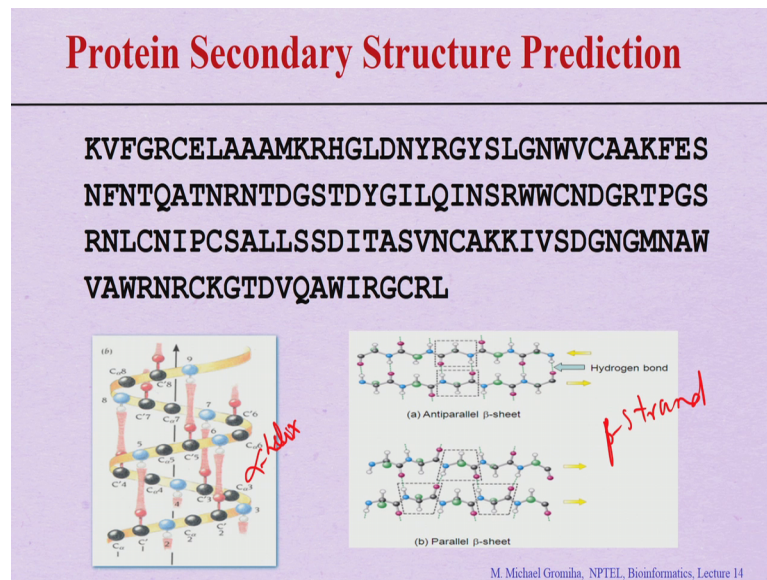
(Refer Slide Time: 01:58)



Here is the range right. So, this is the angle right. So, you can see the values are within that particular range. So, you can see this is from the alpha helix.

So, now the question is if you have this primary sequence, right DSSP can assign the values if the 3D structures are known right.

(Refer Slide Time: 02:36)



This is one example for the myoglobin, first line it shows the sequence and second line shows the secondary structure, these structures are not known for example.

This is for example, you give the sequence is it possible to predict the secondary structure from this sequence from this sequence where there are alpha helices right this is alpha helix, where you form the beta strand. So, there are various methods because once we have the data there was sequence, and the structures then they can compare their results right sequences and the structures and based on the information we can derive some methods to predict the secondary structures, and you can evaluate because there is some experimental known data are available. So, you can also evaluate. So, there are various methods which have been proposed for predicting the secondary structures of proteins. Now first one is the based on statistical analysis.

(Refer Slide Time: 03:16).

## Methods

1. **Statistical analysis**  
(Preference of residues, Chou and Fasman, 1974)
2. **Information theory (GOR)**
3. **Hydrophobicity profiles**
4. **Multiple sequence Alignment**
5. **Machine learning techniques**  
(Neural networks, Support vector machines etc.)
6. **Consensus (Joint)**

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 14

This is a simplest one for example, alanine proposed to be an alpha helix, then if you say few alanines, then we see that this segment prefers to be an alpha helix right. For example, if you see the students they will come at 9 o'clock. They go somewhere 10 o'clock come to a lab right and 1 o'clock go to lunch, then if you see his behaviour he will be for 10 o'clock he will be in the lab its 1 o'clock he will be in the lunch right likewise if some residues they prefer to be accommodating any specific secondary structures, then you can see that these residues are highly populated. So, these segment forms an alpha helix right this is the statistical analysis.

Chou and Fasman; they analysed data of the known proteins, all the known proteins available at that time 1974 they developed the propensity values for all the 20 residues based on that they classified the residues into few groups, which residues prefer to form alpha helix, which residues they do not prefer to be in alpha helix right some of them which are indifference may or may not form right likewise for the beta sheets. Then after this information this you can get achieve an accuracy of about 55 to 60 percent, sometimes same residues they can adopt alpha helices and beta sheets right.

The prediction method can predict either helix or strand with the preferred same segment in the 2 different secondary structures then one will be correct one will be wrong. In this case Garnier's group right they developed a method called GOR based on information theory. Here they did not consider only a central residue, they made a window; window

of this is 70 residues right 8 residues at the left side and 8 residues at the right side right and get the information for all the residues ; that means, any segments these are different secondary structures based on the neighbouring residues. So, you can discriminate the same segment belongs to helix and belongs to beta sheet depending upon the neighbours. So, they used that in that information to predict the secondary structures.

Later on you try to use various properties; one of the most prominent properties is the hydrophobicity, because it can easily distinguish the residues based on polar, nonpolar charged and all. So, they try to see the periodicity of these residues, they take the sequence and find are there any periodicity of residues in the sequence right if you look into the sequence it is difficult to find. So, if we have a kind of figures, kind of profiles easily we can see right what is the hydrophobicity profile?

Student: Sequence vs hydrophobicity

It's a plot connecting sequence versus values right. So, we discussed about several patterns right for alpha helix for the strand, for the transmembrane regions. So, they have some specific patterns and if you scan these patterns in these profiles right and then you can easily identify the secondary structures.

Then after that they try to include the multiple sequence alignment. So, right the time grows, they think the single sequence difficult to predict the secondary structure. If you have more number of sequences of known information, then we gather the information from that sequence alignment and you can try to use for prediction. So, while I have discussed about multiple sequence alignment, how many residues are required for the multiple sequence alignment?

Student: 3 or more than 3.

3 or more sequences, if you want to see any residues are conserved how many sequences will can give the reliable results?

Student: 100.

Yeah if you see more 100 right at least you need 50, more than 100 then you can give reliable results. So, in this case they try to gather the information for different sequences and make the alignment, from this alignment they try to predict the secondary structures

right for example, position number 5, accommodate they highly conserved accommodate by alanine and all the 100 sequences this residue is in helix, then your own sequence you can also tell that this residue number 5 alanine that belongs to alpha helix, in the with the high confidence because if it is highly conserved.

So, this program depends on the sequence alignment, if you get good alignment then your accuracy is high. If the alignment is not very good then you have the dilemma in your assignment. So, then we see the multiple sequence alignment, then we discussed about the profiles placed in the matrices right.

And then they included all the aspects to predict secondary structure. Then they started to use machine learning because all these things depending upon the data sets right if we need to manipulate the data right that the codes, and they try to use different machine learning techniques for example, support vector machines, and then see we put train all the known information with these machine learning techniques, and to predict this secondary structures either helix or strand or coil. So, more details we will discuss in the next subsequent classes.

Then during this course right they started to use join methods or a consensus method right this is easy to do, but also it relies on different servers, different methods. In this case either they use for voting for example, if you get 10 methods available already we can based on statistics or information or whatever right . So, different methods are available, then they run for in the 10 servers right and see the best. For example, 7 methods predicts a residue as helix, you can say this residue belongs to alpha helix right.

Just by voting procedure, you can predict the secondary structures and you can evaluate. Second one is you can get the output from different servers and you can make a metaserver. Again run this output right you train this data to get this prediction results for you have to sequence.

So, current scenario, they try to mix everything they try to include as much information as possible either from single sequence or the statistical preferences or the multiple sequence alignment the matrices right and put together in machine learning right to get the secondary structure prediction. Currently they are using the dip learning right because we can handle large amount of data.

So, in that case you can enhance the prediction performance.

(Refer Slide Time: 09:41)

### Statistical analysis: Propensity

The propensity of an amino acid residue i in any conformation (helix or strand or turn or coil) has been defined as the percentage of residue i in that conformation to the percentage of **all** residues in the same conformation.

$$\text{propensity}_\alpha(i) = \frac{\% \text{ of residue } i \text{ in } \alpha\text{-helix}}{\% \text{ of all residues in } \alpha\text{-helix}}$$

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 14

So, then we will discuss about the statistical methods, the statistical methods, first we need to have the preference of residues in any secondary structures either in helix or in strand. How to calculate the propensity, what is the meaning of propensity?

Propensity of any amino acid for example, alanine right that depends the preference of that that particular residue in a particular conformation for example, if you take alanine or any residue *i* to be in any conformation for example, helix or strand right how to do it? First see the percentage of the residue *i* in that conformation compared with all the residues in a same conformation. So, can you see a propensity of *i* using the equation percentage of residue *i* in alpha helix, now divided by percentage of all residues in alpha helix how to calculate the percentage of residues *i* in alpha helix?

Student: Number of *i*th residue in alpha helix divided by all residues.

All the all the residues of type *i* for example, alanine, 10 alanines right and 7 in alpha helix right what will be their percentage of residues in alpha helix.

Student: 70 percent.

70 percent right. So, then we can see that then normally you see the total number of all the residues in alpha helix the whole protein the 100 residues and 80 residues on alpha helix then how many residues in alpha helix.

Student: 80 percent.

80 percent now you compare. So, if your whole protein is 80 percent in alpha helix and the preference is high only if it exceeds, the number right. Because if we randomly distributed you can get 80 percent right, then you can see whether it is underestimated or overestimated then that will introduce a bias, and you can see some preference of these residues in alpha helix.

(Refer Slide Time: 11:21)

**Statistical analysis: Propensity**

% of residue  $i$  in  $\alpha$ -helix =  $n_{\alpha}(i)/N(i)$

$n_{\alpha}(i)$  = number of residues of type  $i$  in  $\alpha$ -helix

$N(i)$  = number of residues of type  $i$  in the whole dataset

% of all residues in  $\alpha$ -helix =  $n_{\alpha}/N$

$n_{\alpha}$  = total number of residues in  $\alpha$ -helix

$N$  = total number of residues in the whole dataset

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 14

You can say percentage of residues in alpha helix using this equation  $n_{\alpha}$  by  $N$  of  $i$  where this is the  $i$ th residue say alanine in alpha helix, divided by  $i$ th residue in the whole protein.

The likewise again the whole protein is  $n_{\alpha}$  the total number of residues  $n_{\alpha}$  is number of residues in alpha confirmation.

(Refer Slide Time: 11:40)

## Propensity

```
VLSEGEWQLV LHVWAKVEAD VAGHGQDILI RLFKSHPETL EKFD RFKHLK  
HHHHHHH HHHHHHHGGG HHHHHHHHHH HHHHH HHHH HT GGTT  
  
TEAEMKASED LKKHGVTVLT ALGAILKKKG HHEAELKPLA QSHATKHKIP  
SHHHHHH HH HHHHHHHHHH HHHHHHTTT HHHHHHHH HHHHHTS  
IKYLEFISEA IIVLHSRHP GDFGADAQGA MNKALELFRK DIAAKYKELG  
HHHHHHHHHH HHHHHHHH G GGS HHHHHH HHHHHHHHHH HHHHHHHHT  
  
YQG
```

→ Segment e  
→ Secondary structure

DSSP

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 14

You can discuss this with one example here first one is the sequence and a second one is the secondary structure. So, you can see H is a helix and you can see the G, that is 3/10 helix right and you can see it is turn and you can see some gaps. Gaps means you can say it is coil because there is not known this is irregular shape right. So, this is the S is the bend right and you can see the different assignments using captions, generally they use the DSSP. So, this is the result of the DSSP.

Now, if you want to see the propensity whereas, say if you take this protein or say what is the propensity of alanine to be in alpha helix, how to calculate.

(Refer Slide Time: 12:27)

## Propensity

E.g. **Ala**: % of Ala in  $\alpha$ -helix =  $N_a(\text{Ala})/N(\text{Ala})$   
 $= 15/16 = 0.94$

% of all residues in  $\alpha$ -helix =  $N_a/N = 115/153 = 0.75$

Propensity of Ala =  $0.94/0.75 = 1.25$

Propensity of Gly:  $0.5/0.75 = 0.66$

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 14

So, you can see the alanine alpha helix right, this is number of alanine in alpha helix divided by number of alanine you see total number how many of alanines here? You can see all the alanines 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16 there are 16 alanine out of 16 how many of them in alpha helix, if you see the red ones they are 15 of here right.

So, this is not alpha helix right. So, there are 15 alpha helix. So, you can get the numbers that is 0.94, then take the all the residues in alpha helix in this case you can see 100 and 53 residues right if you see here 50 had a 150 153 out of 153 you can see 150 in alpha helix. So, that is 0.75. 75 percentage of the residues in alpha helix, now if you see any residue to be preferred in alpha helix that should be more than.

Student: 1.

More than 1, that's more than 75 percent because the whole protein 75 percent of alpha helix right. So, then you can get the propensity of alanine. So, 0.94 divided by 0.75; this is equal to 1.25.

So, the value is more than 1, then you can say that this residue prefers to be in the particular secondary structure. In this case for example, alpha helix then I take the glycine. We see the glycine, how many glycines here? 1 2 3 4 5 6 7 8 9, 9 or 10 glycines right. So, finally, we can say this is the 0.5 the percentage of glycine this is the percentage of glycine, and this is the percentage of helices right this is 0.075. So, divide

this one, you will get 0.66 see in this case it is less than one. So, from this what can we infer?

Student: not preferred

Glycine is not preferred in alpha helix; this we did based on just one protein. So, currently more than 137000 proteins are available in the protein data bank, you can reduce redundancy right and from the non-redundant set you can calculate the propensity, then you can get a value about 20 different amino acid residues. Then how to derive, how to get these numbers right if you want to derive with this propensity values from the whole database right, how to get the numbers? How to write an algorithm to get this propensity values? What is the necessary information you need?

Student: Sequence data

We need the sequence; all the 20 different amino acid residues right and you need the secondary structure assignment.

### Algorithm

1. Compute the occurrence of 20 residues in helix
2. Compute the occurrence of 20 residues in whole protein
3. Compute the ratio  $\rightarrow \frac{\% \text{ residues in helix}}{\% \text{ residues in whole protein}}$  Sequence  
Sec-Str.
4. Compute total number of residues in helix
5. Divide with total number of residues in a protein  $\frac{N_h}{N}$
6. Divide 3 by 5, to get the propensity of all the 20 amino acid residues in helix.

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 14

You need the sequence and you need a secondary structure assignment right and for 20 different amino acids. First we to get the occurrence of 20 residues in helix, and we need the same occurrence in the full protein. So, full protein how many residues of a particular type for each type we can get, and how about in helix, then you calculate the ratio this will give you.

Student: percentage

Yeah percentage of residues type i.

Student: In helix.

In helix and then the second step you take the total number of residues in helix, and total number of residues in a protein that is  $n_\alpha$  by  $N$  right, this will give you the percentage of residues in alpha helix. Then you divide this by this then you can get the propensity of all the 20 residues in helix you can repeat for all the proteins.

(Refer Slide Time: 16:08)

TABLE 5.2 Chou-Fasman parameters					
Residue	$P_\alpha$	Residue	$P_\beta$	Residue	$P_\tau$
Glu ✓	1.53	Hβ Met	1.67	Asn	1.68
Ala ✓	1.45	Val	1.65	Gly	1.68
Leu ✓	1.34	Ile	1.60	Ser	1.56
His	1.24	hβ Cys	1.30	Pro	1.54
Met	1.20	Tyr	1.29	Asp	1.26
Gln	1.17	Phe	1.28	Tyr	1.25
Trp	1.14	Gln	1.23	Cys	1.17
Val	1.14	Leu	1.22	Trp	1.11
Phe	1.12	Thr	1.20	Lys	1.01
Lys	1.07	Trp	1.19	Arg	1.00
Ile	1.00	Iβ Ala	0.97	Thr	1.00
Asp	0.98	iβ Arg	0.90	Phe	0.71
Thr	0.82	Gly	0.81	His	0.69
Ser	0.79	Asp	0.80	Met	0.67
Arg	0.79	bβ Lys	0.74	Ile	0.58
Cys	0.77	Ser	0.72	Ala	0.57
Asn	0.73	His	0.71	Gln	0.56
Tyr	0.61	Asn	0.65	Leu	0.53
Pro ✓	0.59	Pro	0.62	Glu	0.44
Gly ✓	0.53	Bβ Glu	0.26	Val	0.30

**Hα: Strong helix former**  
**hα: Helix former**  
**lα: Weak helix former**  
**iα: Weak helix breaker**  
**bα: Helix breaker**  
**Bα: Strong helix breaker**

And finally, you can get their propensity values for example, if you take the helix, this is the propensity values. These are data obtained from Chou and Fasman. If you see these numbers, some of them are more than 1, some of them are less than 1, based on these numbers which residue is preferred to be in alpha helix?

Student: Glu,

Glu

Student: Ala and Leu

Alanine and leucine. Which residues is not preferred to be in alpha helix? You can see Pro, Gly and so on. This, we have high preference right. So, these residues are low preference, right less preference.

Then, if you see their values, some of them are very high values such as 1.5, 1.45 and something that very low 0.5 or 0.6, and several residues they are oscillating around 1 right that also depends upon the dataset. If you use 100 proteins you can derive the values if you use 1000 proteins or 10000 proteins right although there is not major difference, but you can see a similar trend right maybe some minor changes hence they divide this group into 6 groups; first one is  $H\alpha$ .

This is strong helix former. So, because they have high tendency to form helix because the value is very high and they have this  $h\alpha$  right that is in this group. So, they have a tendency to form  $h\alpha$  right to form the helix, there is helix formers then here just around one little bit around one, here this is weak helix former.

Then come down there is just a little bit less than 0. So, they may get weak helix former or helix breaker, because they does not prefer to be in the helix and you can see a  $b\alpha$  these are helix breakers and  $B\alpha$ , they a strong helix breaker. So, if you have the propensity based on the numbers, they grouped into 6 groups.

Strong helix former, helix former, weak helix former, weak helix breaker, helix breaker and strong helix breaker and you can use this information to predict the secondary structures. So, if you have any segment, which contains mainly these residues, then there is a high probability to be in helix.

If you have more residues which are in this category, then it has less probability to be in alpha helix this is for helix then we can do this for beta strand or coil or turn, which is the beta strand. Now here are they classified 6 different groups, now you can see the high values, this is the strong beta formers and these are the beta breakers. And if you compare this  $P\alpha$  and  $P\beta$  some residues are distinct and some residues are same for example, Proline right, here it is not preferred in alpha helix also is not preferred in beta strand. Glutamic acid: it is a breaker in the case of beta, but it is a strong helix former in alpha. So, if you have glutamic acid in a sequence, then you can see there are tendency to be in alpha helix right likewise some difference of preference you can see the preference in beta, but not in alpha, but some cases you can see similar preferences in both alpha helix and beta sheet this will give you a conflicting situation, right the same segment can be predicted with the alpha as well as for the beta.

In this case you can compare the values which one has high values, then you can assign accordingly. So, how to predict the helical segments or beta strands right, from this amino acid sequence?

(Refer Slide Time: 19:51)

### Rules for Identifying Helix

- The propensity values and the residues belonging to these 12 classes are shown in Table 5.2.
- For a protein sequence, assign the appropriate parameters from Table 5.2.

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 14

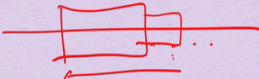
First you see the sequence and assign the propensity values, because if you take this table, there are 12 different groups for the helix 1, 2, 3, 4, 5, 6 right, for the beta also you have 6. Now you can take these preferences and assign the values, then you can see the weightage.

(Refer Slide Time: 20:12)

### Rules for Identifying Helix

**Helix:**

- Values of the six parameters are  $H_\alpha = h_\alpha = 1$ ;  $I_\alpha = 0.5$ ;  $i_\alpha = 0$ ;  $B_\alpha = b_\alpha = -1$ ;
- Scan for window of 6 residues, where score  $\geq 4$  i.e. at least four helix formers and not more than one helix breaker;
- Extend the length in both directions until the score is less than 4;



M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 14

For example if it is helix. So, we take the different parameters for example,  $H\alpha$ .

This is the helix former, you take it as one and this is the weak helix former, taken as 0.5 for example, this is the weak helix breaker, take it as 0, and this as helix breakers they put as -1 because we have to give preference for the residues, which has high preference to be in alpha helix.

If it has low preference it has less preference taken as -1. Then you take the window residues of 6 residues, because alpha helix is longer than the beta strands right. So, take the 6, minimum 6 residues and you can extend it. And the idea is if you take any segment, this segment have high preferred residues to be in alpha helix and less number of low preference residues.

So, at least 4 for helix formers and there should be not more than 1 helix breaker. So, with this assumption they put a score of more than 4. In this case they will have high tendency to have at least 4 helix formers, because we take the value of 1 and not more than 1 helix breaker right. In this case you can get a score of 4, then you can extend the directions for example, if we find, we will reach the sequence and you find this is the helical segment, and then you can extend and see whether the score is down or not. The score is still more for example, they preferred residues are present together, then you can add up, you can add up when there finally, add up to here to make the complete helix.

(Refer Slide Time: 21:46)

## Rules for Identifying Helix

- Continue the search and locate all helical regions in the sequence.
- Refinement: Pro, Asp, Glu: N-terminal; His, Lys, Arg: C-terminal; Pro: Not in inner helix or C-terminal

TABLE 5.2 Chou-Fasman parameters	
Residue	$P_{\alpha}$
Glu	$H\alpha$ 1.53
Ala	1.45
Leu	1.34
His	$h\alpha$ 1.24
Met	1.20
Gln	1.17
Trp	1.14
Val	1.14
Phe	1.12
Lys	$l\alpha$ 1.07
Ile	1.00
Asp	$i\alpha$ 0.98
Thr	0.82
Ser	0.79
Arg	0.79
Cys	0.77
Asn	$b\alpha$ 0.73
Tyr	0.61
Pro	$B\alpha$ 0.59
Gly	0.53

So, now how to do this continued search, and there are some conditions we can assign the terminals for example, you can see some residues which are mainly in the N-terminal, and some of them on the C terminal and Proline, as we discussed it is not presented in the inner helix whether it will form turns.

These are there some common observations, this you can use for refinement. If you can more residues are predicted right if you want to refine it, then you can use this information.

(Refer Slide Time: 22:12)

The slide displays an amino acid sequence: **KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFNTQATNRNTDGSTDYGILQINSRWWCNDGRTPGSRNLCNIPC SALLSSDITASVNC AKKIVSDGNGMNAWVAWRNRCKGTDVQAWIRGCRL**. A handwritten red box labeled "α-helix" encloses the residues **RCELAAMKR**. Below the main sequence, a list of overlapping 6-residue segments is provided: **KVFGRC**, **VFGRCE**, **FGRCEL**, **GRCELA**, and **RCELAA**. A red arrow points to the first segment, **KVFGRC**.

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 14

So, now I give a sequence, this is the amino acid sequence. So, now, the question is which residues from alpha helix, which residues belong to beta strand right.

First take the sequence and if we take the overlapping residues of 6 residues. So, 1 2 3 4 5 6. So, KVFGRC, this is segment 1, and the second one start from here to here right VFGRCE and the third one is start from F 4 5. So, likewise you can make different segments.

Now, you take this one and see the score for instead of just writing the number just you just we check whether this is 1 or 0.

(Refer Slide Time: 22:51)

KVFGRC: $0.5+1+1-1+0+0 = 2.5$	
VFGRC: $1+1-1+0+0+1 = 2$	
FGRCEL: $1-1+0+0+1+1 = 2$	
GRCELA: $-1+0+0+1+1+1 = 2$	
RCELAA: $0+0+1+1+1+1 = 4$	
<b>Score</b>	
MKRH: $1.20+1.07+0.79+1.24 = 4.3$	
KRHG: $1.07+0.79+1.24+0.53 = 3.63$	

Residue	$P_{\alpha}$
Glu	H $\alpha$ 1.53
Ala	1.45
Leu	1.34
His	h $\alpha$ 1.24
Met	1.20
Gln	1.17
Trp	1.14
Val	1.14
Phe	1.12
Lys	l $\alpha$ 1.07
Ile	1.00
Asp	i $\alpha$ 0.98
Thr	0.82
Ser	0.79
Arg	0.79
Cys	0.77
Asn	b $\alpha$ 0.73
Tyr	0.61
Pro	B $\alpha$ 0.59
Gly	0.53

So this is the table, you can see this 1, this you put 0.5, this is 0, this is -1, take the first one you take k this is equal to lysine this equal to 0.5, lysine is here 0.5 right then V, V is here this is equal to 1 then your F is here again, phenylalanine. So, this is equal to 1 right and this G is in this group this is -1. So, you put the -1 and R and C that is here R is here and C is here this is 0. Present the values, just approximately assign, this is 2.5. So, this is less than or equal to 4 right this is this is not the part of helix go the next segment right here these numbers are the same and the last one is E, that is equal to 1 right in this case the score is equal to 2, 2 or 3, 1 2.

Student: 2.

This should be 1.5 right 1, 2 - 1, 1 this is will be 1.5 right this is 1.5 right, this will be 2 now the third 1 if you see. So, these are the same right, these are the same here and the last one is L, you seen this equal to 1, one is equal to 2 again go on adding these numbers when they go with RCELLA, R equal to 0, C equal to 0 all these are 1 because E is 1, L is 1 and R and L is 1. So, final score is 4, now you can see this is a segment you can identify as an alpha helix. Now whether this only these residues form an helix or you can extend it then, if you take the last one MKRH, it is here and then if you do it here.

So, if you do this, this is 4.3 you can add right this is RCLLA, RCLLA now this is next A, if you add the A, A is preferred one. So, it will be more than 4 right, the next one is M. So, M is also preferred 1, this this is belongs to H $\alpha$  right m belongs to H $\alpha$ . So, that also

we can increase it. So, the next one is K. So, lysine. So, it is all 0.5 and if you add the values right this is then this will give you the value more than 4. You can extend up to this MKRH. So, the here you can add this M is equal 1.2 right from here and the K is 1.07. 1.07, lysine is in here and arginine is 0.79.

This should be 1.2 or 4, this where is 4.3 then add next one if I add 1 more KRHG right you can see the last 4 one KRHG right if I add this then the number is down less than four. So, I cut here. So, now, segment starts from this lysine right you can start from here this one right you can write, where it is 4 here. I start from R right you can start from here and up to here this. This is the segment which can form alpha helix. So, how to do this? First what to do?

Student: by 6 residues

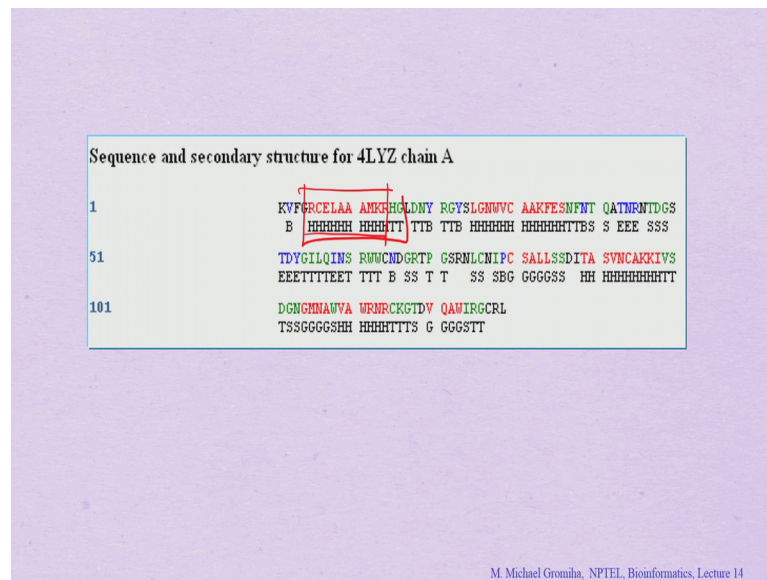
Six residues window, you get the overlapping segments. So, these are the 6 residues overlapping segments and then.

Student: Then see the score.

Then it gets assigned the values from this table, window 6 different categories based on the tendency to form helix or they are not preferred to be in helix they have -1 to +1, just put the numbers assign the numbers and see whether this number is more than or equal to 4. If it is very less go to the next segment and you go to the next segment until the where we get the value of 4 right you get the value of 4 here. So, you can start from here, then this is 4 the question is if you add 1 that will have a high preference or low preference. So, add 1 and see the last 4 residues and whether the number is right you can see the segment the 6 window segment is the value is more than 4.

So, you can see the numbers right this is 4.3 because you can get the numbers from this table here we just assign the numbers 1 or minus 1 0.5, just you see how many helix formers or how many helix breakers. Once we identify the segment then extend it based on the real values right. So, then we extend it until we get the value less than 4, here you can see the get the value less than 4 then we stop there. This is the complete segment you can get ok.

(Refer Slide Time: 27:18)



So, this is the assignment and actually they compare with the experimental data it also ok it is up to here right, it is completely matched with the experimental data.

This is a segment here we couldn't get any helix values are less and here we get the high preference, and the segment is predicted is alpha helix right it matches with the experimental data.

(Refer Slide Time: 27:37)

### Rules for Identifying $\beta$ -strand

- The values of the six parameters are  $H_\beta = h_\beta = 1$ ;  $I_\beta = 0.5$ ;  $i_\beta = 0$ ;  $B_\beta = b_\beta = -1$ ;
- Scan for window of 5 residues, where score  $> 3$ , i.e. at least three strand formers and not more than one strand breaker;
- Extend the length in both directions until the segment has the average propensity  $< 1$ ;
- Continue the search and locate all strand regions in the sequence.

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 14

Now, how to get beta starts give the same procedures, take the  $H_\beta$  small beta is equal to 1,  $I_\beta$  equal to 0.5,  $i_\beta$  equal to 0, and  $B_\beta$  or  $B_\beta$  equal to -1. So, here we use this scan of 5

residues, and see it is score of more than 3 because all beta strands are smaller than alpha helices.

So, I take 5 residue segments right how they define these segments? They tried with vary segments and compare the results with the experiments and finally, which will matches they try to fix with that one. This is how they tried to assign 6 and 5 for the helixes and strands. Then you can extend the length in both the directions unless the average property is less than 1 right you can continue the region and find the strand if there are the same procedure you can adopt. So, I do not have to explain all these details right.

(Refer Slide Time: 28:24)

Rules for Identifying $\beta$ -strand	
<b>Conflict situation:</b> A region containing overlapping helical and strand assignments is considered as a helix (or strand) if <u>average propensity of <math>\alpha</math>-helix (<math>\beta</math>-strand)</u> is <u>greater than that of <math>\beta</math>-strand (<math>\alpha</math>-helix)</u> .	
Residue	$P_{\beta}$
H $\beta$ Met	1.67
Val	1.65
Ile	1.60
h $\beta$ Cys	1.30
Tyr	1.29
Phe	1.28
Gln	1.23
Leu	1.22
Thr	1.20
Trp	1.19
l $\beta$ Ala	0.97
i $\beta$ Arg	0.90
Gly	0.81
Asp	0.80
b $\beta$ Lys	0.74
Ser	0.72
His	0.71
Asn	0.65
Pro	0.62
B $\beta$ Glu	0.26

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 14

So, here we have the values for B $\beta$  and you can assign the values and find out. For example, if you find a conflict situation the same segment can be predicted as helix and can also be predicted as beta sheet in this case what to do? In this case, get the segments and assign the real values and see which one is dominant. If average propensity of alpha helix is greater, then this is alpha helix, if beta strand is greater, then it is beta strand. So, in the conflict situation you can also get the values, and you can compare the numbers and then assign the secondary structure.

(Refer Slide Time: 29:00)

The screenshot shows the Chofas webserver interface. On the left, there is a 'Choose:' section with three options: (A) Program, (B) Protein (sequence/accession), and (C) Analyze protein. Option (A) is selected, and a dropdown menu shows 'Chou-Fasman Secondary Structure prediction'. Below this, there is a 'Protein sequence:' field with a 'FASTA format' dropdown. A text area contains a protein sequence in FASTA format:   
>4LYZ: A|POBID|CHAIN|SEQUENCE  
KVFGRCELAAANKRGLNTRGYSLGWVCAAKFESNFTQATNRNTDOSTDVGILQIN  
SRWVNDGRTPGSRNLCTPC  
SALLSSDITASVWCARKIPSDQNGMNAVAVNRCKGTQVQAVIROGRL  
At the bottom of this section are '(C) Do Analysis' and 'Submit Sequence' buttons. On the right, the Chofas header states: 'CHOFAS predicts protein secondary structure version 2.0.ou66 September 1998. Please cite: Chou and Fasman (1974) Biochem., 13:222-245'. Below this is a 'Chou-Fasman plot of 8, 129 aa: 4LYZ: A|POBID|CHAIN|SEQUENCE'. The plot shows predicted secondary structure elements: helix (indicated by '<----->'), sheet (indicated by 'EEEEEE'), and turns (indicated by 'T'). The sequence is aligned with these predictions. At the bottom right, residue totals are given: H: 47, E: 33, T: 16; percent: H: 36.4, E: 25.6, T: 12.4. A red arrow points to the URL <http://fasta.bioch.virginia.edu/fasta/chofas.htm> at the bottom of the slide.

So, here is one webserver, there are several servers available to predict the second structure based on Chou and Fasman parameters right here they use various programs and here we take the Chou and Fasman parameters, and here we give the sequence in fasta format it uses this method, the same 1974 Chou Fasman, right and finally, you can see this region is the helix, this is what we discussed now the predicted one. And then we can see this is the strand region and it is a turn region. So, finally, can use this server to predict the secondary structures based on Chou and Fasman methods. So, summarizing what did we discuss?

Student: Protein secondary structures.

Protein secondary structures. So, what are the various secondary structures?

Student: helix.

Alpha helix beta strands right and turns right and then coil right how the secondary structures are formed?

Student: Hydrogen bonding.

Where are the hydrogen bonding between?

Student: Back bone

Back bond NH and?

Student: CO

CO groups. So, it is they have the periodic arrangement of residues right. So, alpha helices are, how alpha helices are found?

Student:  $i$  and  $i+4$ .

$i$  and  $i+4$ ; helices right and this is closely packed. If you compare to the alpha helices and beta sheets right this is closely packed and this is loosely packed, whereas, it is loosely packed right then how we obtained this Ramachandran plot, what is the Ramachandran plot?

Student: Allowed and disallowed regions.

Kind of plot connecting phi and psi which tells you which rotations of phi and psi are allowed.

Student: Allowed.

In different secondary structures alpha helices and beta sheets right now which are allowed regions and the partially allowed regions right and the complete disallowed region right fine. So, now, how to assign the second structures, if you know the 3D structures Kabsch and Sander developed a program called?

Student: DSSP.

DSSP; what is DSSP?

Student: Dictionary of Secondary Structure of Proteins.

Dictionary of Secondary Structure of Proteins right, they assigned the secondary structures based on hydrogen bonding pattern right, they design the values right. Now the question is whether we can predict a secondary structure from the sequence. Several methods we discussed right, I will give the details in the next class right similarly we discussed in detail about.

Student: Chou Fasman method

Statistical analysis that is Chou and Fasman method, right. So, how Chou and Fasman method works.

Student: preference

Preference of residues. How to get the preference?

Student: Propensity values.

Propensity values right. for any residue, how far that specific residue prefers to be in in a secondary structure, either helix or strand based on that residues in helix, total number of residues, totally how many of type i, total number of the whole residues is a full protein and how many residues in helix in the complete protein. So, based on this information you can get the propensity. Creatively own example on which residues preferred to be an alpha helix.

Student: Alanine.

Alanine, glutamic acid are preferred to be in alpha helix. Some bases, which are not preferred to be alpha helix for example?

Student: Lysine.

Lysine, proline right there are preferred right right. So, based on this information. How to predict the helical segments?

Student: Chou and Fasman

Here you take a window length all of 6 residues right and then.

Classify this values into 6 groups and assign the numbers and see whether the numbers more than 4 right. We continue till we get the value 4, once you get the 4 there could be a nucleic acid right then extend the residues and see the last 4 residues have the actual value of more than 4. With this 4 add up where it decreases right less than 4 then you stop then you can identify this residue is an alpha helix. Likewise for beta strand right they use how many residues?

Student: 5 residue window

Five residue windows and then repeat the same procedure right to get the beta sheet. If there is a conflict situation for example, same segment it is predicted as both alpha helix and beta sheet right how to do this?

Student: To one with a greater propensity

The one with a greater one that you can assign where it is alpha helix having more propensity right then you can assign this alpha helix or beta strand vice versa. So, now, we can predict this using different servers right, one example is you can see in this server right. So, in you can just do the sequence and give you the secondary structure. We discussed about the statistical methods and next class we will focus on the different other methods for example, the information theory, multiple sequence alignment, high domestic profiles, the neural networks.

Right like machine learning techniques as well as the consensus right this we will discuss in the next class.

Thank you very much.