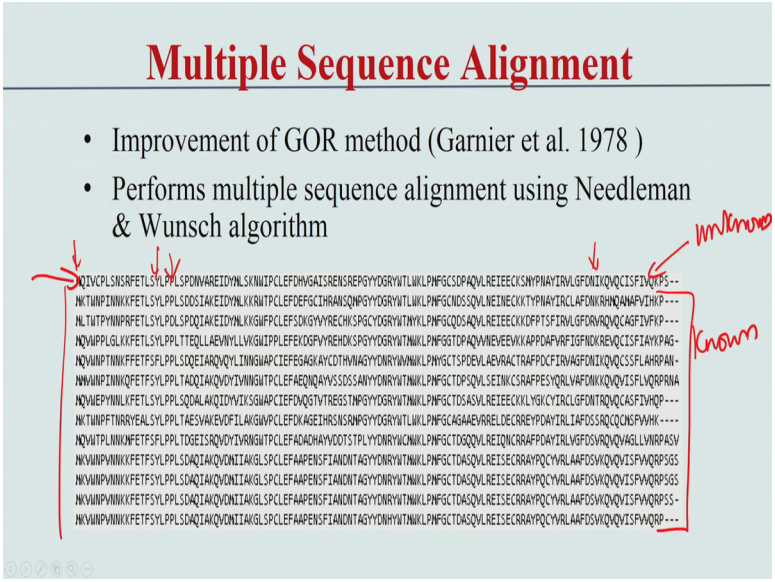
Bioinformatics
Prof. M. Michael Gromiha
Department of Biotechnology
Indian Institute of Technology, Madras

Lecture – 15b
Secondary Structure Prediction III

Next one is the multiple sequence alignment. So, we discussed three different methods; one is a statistical method, Chou Fasman and we discussed about information theory GOR method and the hydrophobicity profiles; based on the patterns in hydrophobicity. These are all based on single sequence, if you have one sequence we do not need anything else; just see the values the preference of residues or the profiles we can predict.

Later on, we realized that instead of using a single sequence; if you obtain the information; similar to that particular sequence; that may increase the prediction performance. So, they started to use the multiple sequence alignment; how do you get the multiple sequence alignment? Take your sequence, get the homologous sequences and you can align; what program we use to get the multiple sequence alignment?

(Refer Slide Time: 01:13)



Student: ClustalW.

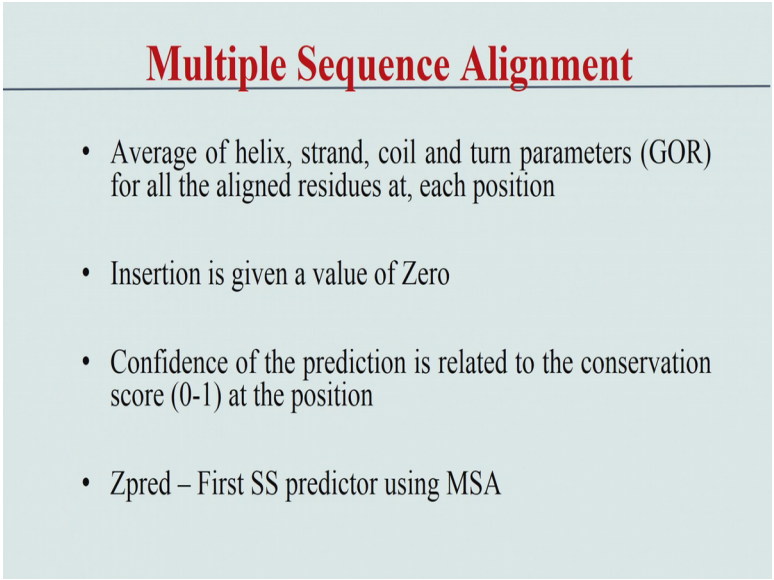
ClustalW; ClustalOmega we can use for the multiple sequence alignment. So, for example, this is your sequence you obtained the homologous sequences and then you can get this multiple sequence alignment. Now we get several sequences and finally, I can see the multiple sequence alignment. And if you look into the sequence; some of them are conserved, what is the conservation?

Student: almost all residues

Conserve same residue at same position; for example, take the first residue it is highly conserved and some cases it is variable; this is P is conserved here and O is conserved here and some positions you can see the variability; changes with several residues; for example, we take here I, K, V and yes there are many residues; they change the variability. So, if it is highly conserved then the possibility of the prediction performance is high.

Because we take the same from different sequences, so what to do; when you make these sequence alignment, we have the multiple sequence alignment there are different ways to do.

(Refer Slide Time: 02:09)



First case is we can get the GOR values; that means, we have the frequency of occurrence matrices. So, get the frequency of occurrence for the helix, strand and coil; for the different residues in the sequence alignment. And then you can see which one is

the highest one; you can predict; here is the confidence increases, if it is highly conserved why?

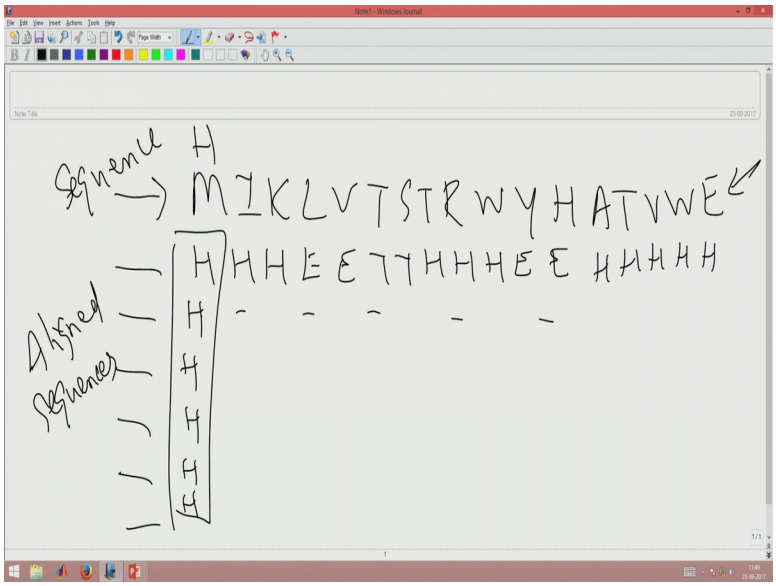
Student: More information.

Because we get the same information for the same sequence; that case you can if you add up all these things, we will get a high value. So if highly conserved; we will get the high confidence values. So, if it is insertion deletion where you say there is a gap; then we say value is 0 and the others they get the table for the helix, strand, and coil and turn and get the values and finally, look at the 4 different states; then you can predict; if the confidence is very high; is based on the conservation score between 0 to 1, if it is 1 then you can get the high conference than the value of 0.

So, the first predict is Zpred; this is the first secondary structural predictor using multiple sequence alignment and using this one. Then there is another way to use a multiple sequence alignment for predicting the second secondary structures. For example, right if you have this multiple sequence alignment; this is unknown and for example, if you get these numbers from known; that means, you search your own sequence with BLAST; with the data set of PDB sequences, with known structure sequences and if you could do this alignment; the first one is unknown and the all other sequences we know the secondary structures.

For example, we take second sequence; we change these sequences into secondary structure; that means the sequence will be; for example, if you have the sequence.

(Refer Slide Time: 04:02)



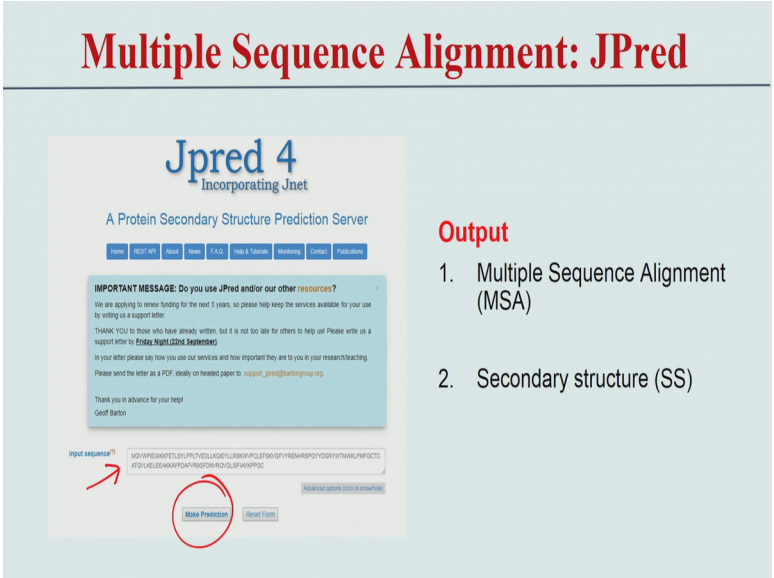
This is your unknown sequence; so, you align with the known values; known structures. Then if you have the different aligned one; the aligned sequences these are the aligned sequences and for each aligned sequence, you have the second structure assignment.

For example this is helix; so, for each case we have the secondary structure assignment. And see the values here if it is helix, helix, helix, helix, helix then you can assign this value for this; so for each positions, if have known secondary structures you can assign the values and then you can get the secondary structure for your query sequence.

So, if the accuracy is high; you can see the conservation is very high. Other way around, if the conservation is high for example, if this highly same residues and this case you can see the prediction values with the high accuracy. So, the alignment score increases or the multiple sequence alignment is highly confident and then the accuracy increases.

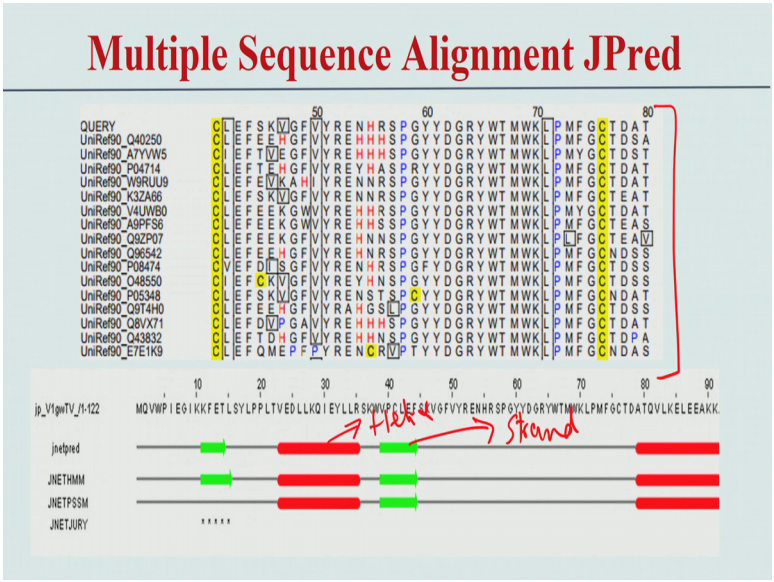
So, they did for several proteins with the homologous sequences and tried to evaluate the performance and see the performance which are the sequences, which are highly aligned they have high performance than the residues which are in the lower performance. This is currently in the research problem, we have the issue in the disorder region; because it is very difficult to get the proper alignment. So, it is very difficult to predict this secondary structures; when they form with the any of the complexes.

(Refer Slide Time: 05:58)



So, now there is the method called Jpred; this is the one of the earliest methods; they used the multiple sequence alignment. It takes any of these amino acid sequence; this input sequence; you know if you make prediction, this will give you the sequence alignments for any given sequence as well as secondary structure.

(Refer Slide Time: 06:17)



For example, if you run your own sequence; so, this is your multiple sequence alignment and here you can see the predicted method; here this stands for?

Student: Helix.

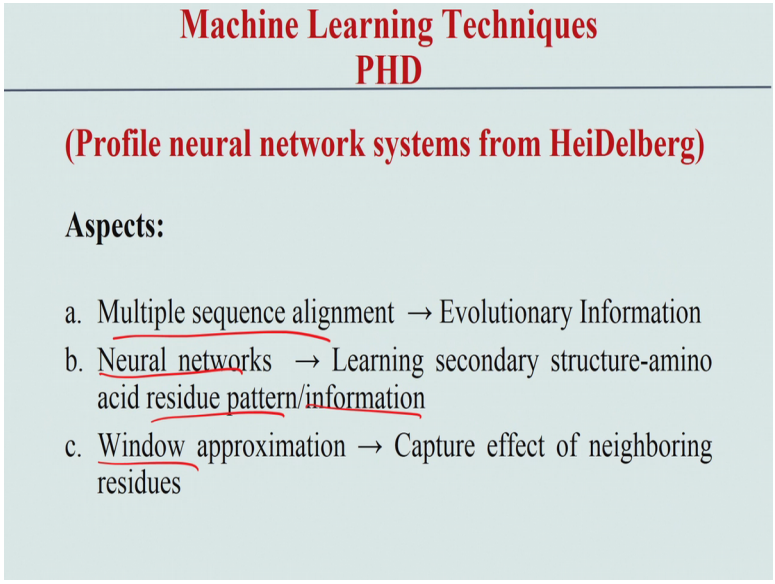
Helix; this arrow stands for? Strand; so, you can see the multiple sequence alignment or this is a highly conserved and if the conservation score is high then you can see the prediction accuracy also very high; they can get this information. So, we discussed about the propensity and the information theory, hydrophobicity profiles and the multiple sequence alignment.

If you see the performance; each method you can see improvement of performance. First we started with a 60 percent, 65 percent, 68 percent and the multiple sequence alignment; if we get good alignments, you can get up to 65 to 70 percent.

So, now they tried to put all the information in a machine learning because here we try to understand which feature is important? How for each feature performs in the secondary structure prediction? Say add all the information and put the machine learning techniques; machine learning techniques is just the mapping of the input features with the output. For example, you have 100 proteins; all the 100 proteins, we know the assignment; segment 5 to 10 is helix; 8 to 12 is strand and so on.

And here we have information central residue we know, neighboring residues we know, patents we know, preferred residues we know; we give all the information. Then the machine learning it is a kind of black box; it maps this input sequence with the output and assign weights to maximize the performance; in the later classes we will discuss more details.

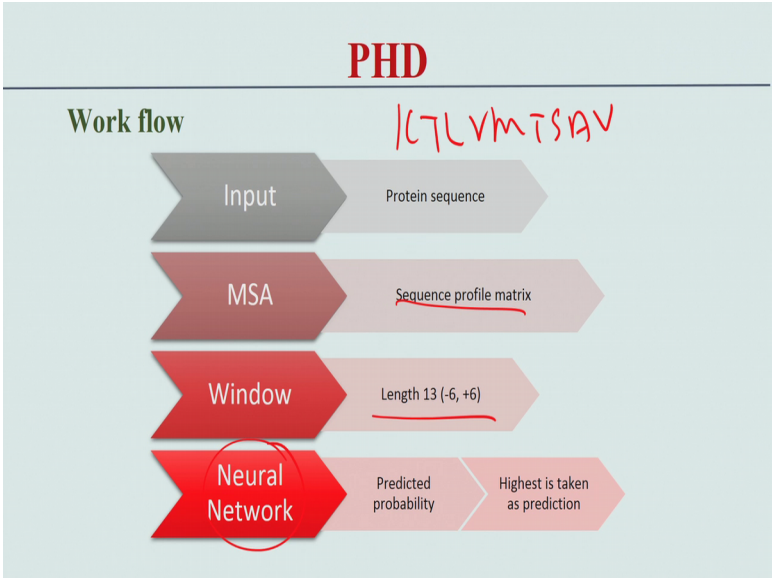
(Refer Slide Time: 07:59)



So, they give the information regarding the sequence alignment for the evolutionary information. And they used any of the machine learning techniques, there are several techniques available neural network, support automations; so, many methods available.

So, for example, if you take the neural networks; it learns the known information that is what we give as input. For example, amino acid pattern or any neighboring residue information and they try to learn this information, they allows you use the different windows. They can use 3, 5, 7, 9 and then optimize the windows and optimize the parameters or any properties to predict the secondary structures; how it does?

(Refer Slide Time: 08:44)



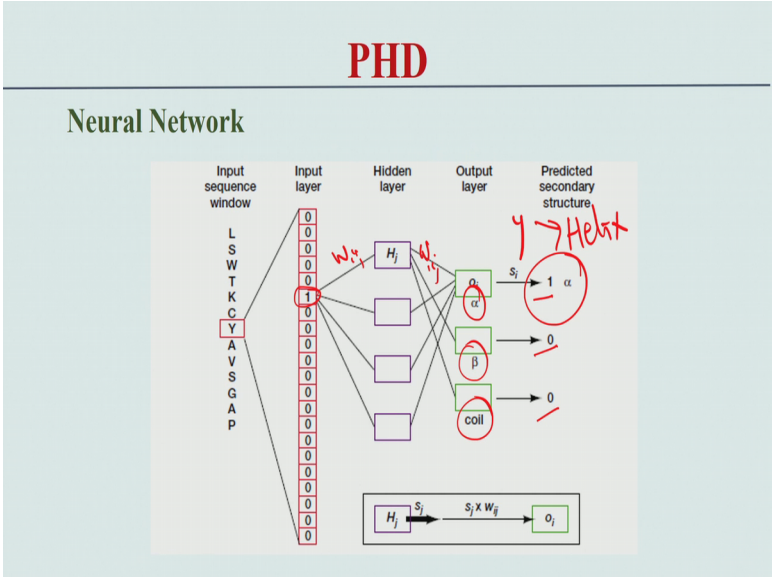
So take the input sequence; this is your protein sequence for it is only if you have this KTLVMTSAV then we get the profile matrix, how to get the profile matrix?

Student: PSSM.

PSSM gets the multiple sequence alignment; from the multiple sequence alignment we have the preferred residues at any particular position; so we get the values. Then they take the window length, they can use different window lengths right. So, for example, if there is 13, there is -6 to +6; all the information they take and put everything in the machine learning. For example, neural networks which will give you the probability of each residue to be in helix or in strand or in coil so different, different.

For example if you take 3 states; it will give you the probability; if this is the case, central residues 3; 1 in; these are the neighboring residues; this is the preferred a residues and so on. It can map, so that this probability of helix is 0.5 strand is 0.2 and coil is 0.1. Then we can see from; what you can see the best one this could be in the helix or strand or coil and so on.

(Refer Slide Time: 09:49)

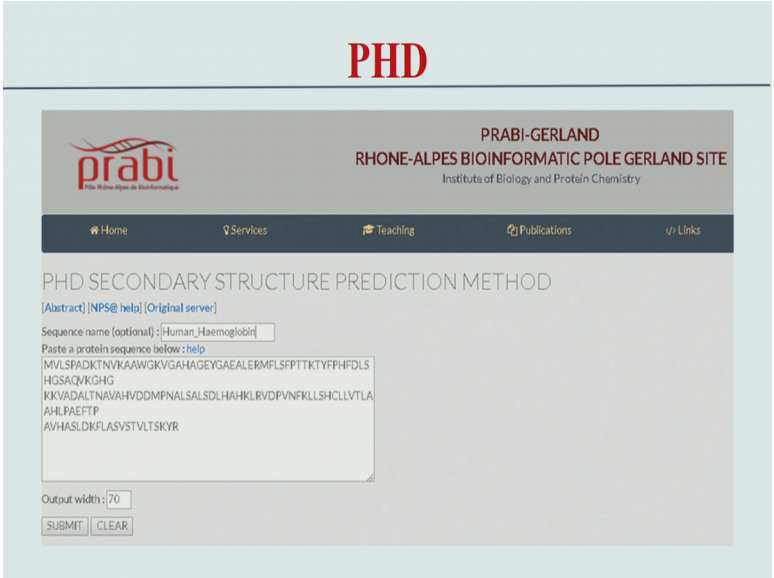


So, here this is one example; so here you have the input sequence. So, this is the central residue; this is Y, the input layer they have the 21 values; 20 values for the 20 different amino acids; 1 for the padding sequence that they align the next one or they can also use it for insertion and deletions.

So, Y is here right; so, they put one here and all others; that is 0. Likewise then they have some hidden layers; here is the place they train your data. So, they assigned different weights; so, put either here or here; different ways they do; between this i and j and based on its weight; they give the values for the output. What is the preference or some preferred probability of that particular residue Y in helix or in the strand or in coil.

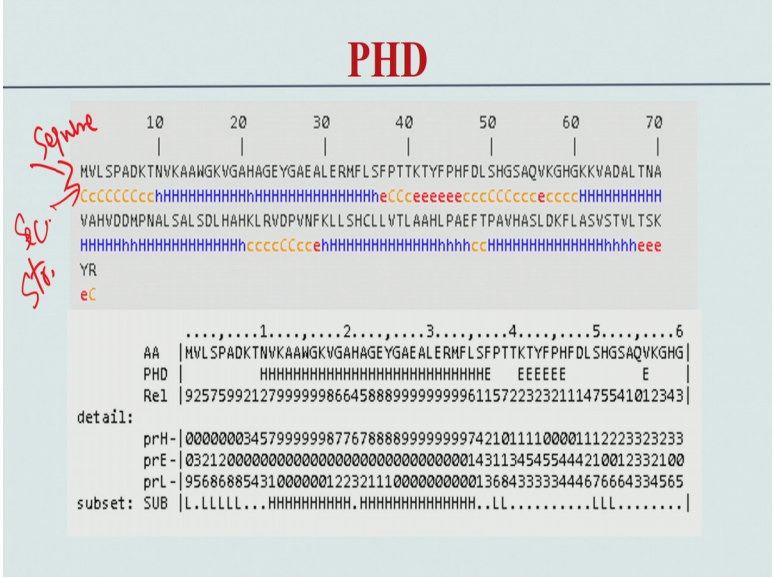
For example, if you give this as 1; this is probability 0, this probability 0 then you can say this residue L is in helix. So, likewise they do it for each; not residue, this residue is for example, if it is Y, then this is you can see these are Y. For each residue, you can get the probability; from this probability you can see whether it can be a helix or strand or coil.

(Refer Slide Time: 11:08)



So, this is the server; so it takes the amino acid sequence as an input. And they takes the information that you obtained from the sequence, then they give the desired output; this is your sequence and here you can see; this is your secondary structure.

(Refer Slide Time: 11:23)

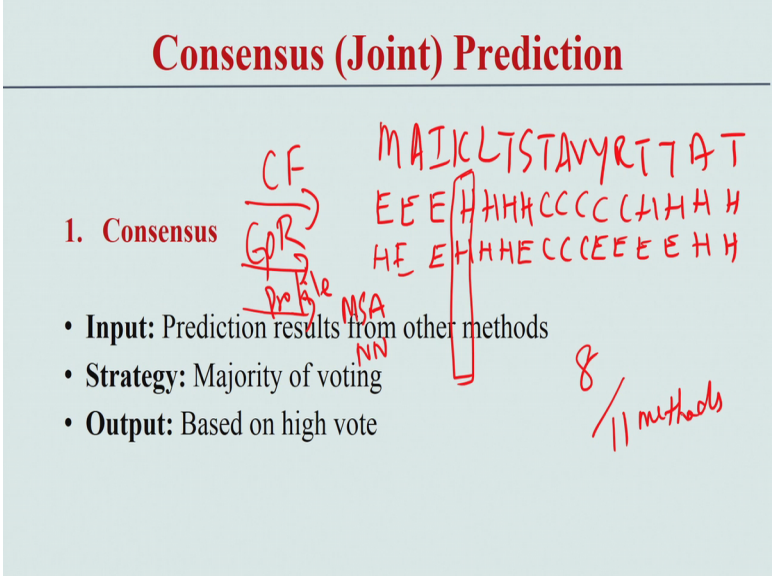


So, the PHD is one of the highest performing methods; so, it could achieve an accuracy of about 70 percent at that time in 1993. They used a neural network to map the information and to predict the secondary structures and highest accuracy.

So, now if you see the growth; from these statistical methods to the machine learning; you can see the growth in the accuracy. And finally, they could be able to combine different information in the neural network to get the highest accuracy. Now, the next one is; so they have different, different methods and what are the methods you take? The performance is around 60 to 70 percent.

So, now they try to use the consensus for example, if you have 10, 15 prediction methods; take any amino acid sequence try to use different methods.

(Refer Slide Time: 12:38)



And predict the secondary structures and see are there any tendency of different methods to predict the same residue in same secondary structure. In this case they take a particular sequence; for example we have one sequence here, use the Chou and Fasman and you get the values; get the secondary structures for example, this is E E E H H H H C C C C C H H H H.

Then we use GOR and we get the secondary structures; for example, it will give like this. Then we use a profile and the multiple sequence alignments; where I will use neural network, we get the values. Then compare these values; for example this region; if all methods could predict helix or majority of voting.

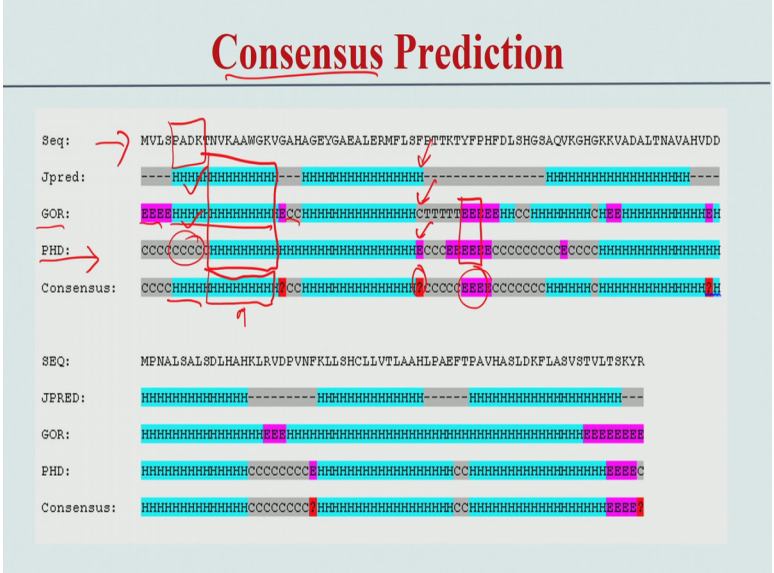
For example, if you have 11 methods and if the 8 methods predicts a particular structure like helix or strand. Then we can assign that residue previous to be in helix or strand; so,

now we can compare experimental data with the known database and see; how far we can improve.

See if we do this consensus method or the joint prediction methods; they found that this method could improve the performance compared with any individual methods; this is one. And the second aspect; they can also do the output of each sequence, we will get the secondary structures and this can be the input of any machine learning.

For example, if we take this; the input information again they put this is also input information. Instead of using the sequence, they give the output obtain from the sequence and they train and see what the possibility of having helix or strand or coil. Then at the method we will train and see this is a first residue could have the probability of helix or strand or coil; these ways we can do.

(Refer Slide Time: 14:56)



So, how to do this? First you can see any sequence; this is your input sequence and they use different methods for example, Jpred; Jpred is based on what?

Student: MSA.

Multiple Sequence Alignment; they also later on use the other information; for example, the machine learning and all. So, mainly it's developed initially based on multiple sequence alignment; so, Jpred predicts these residues are helix; the helical regions then take the GOR; what is GOR?

Student: Theory information.

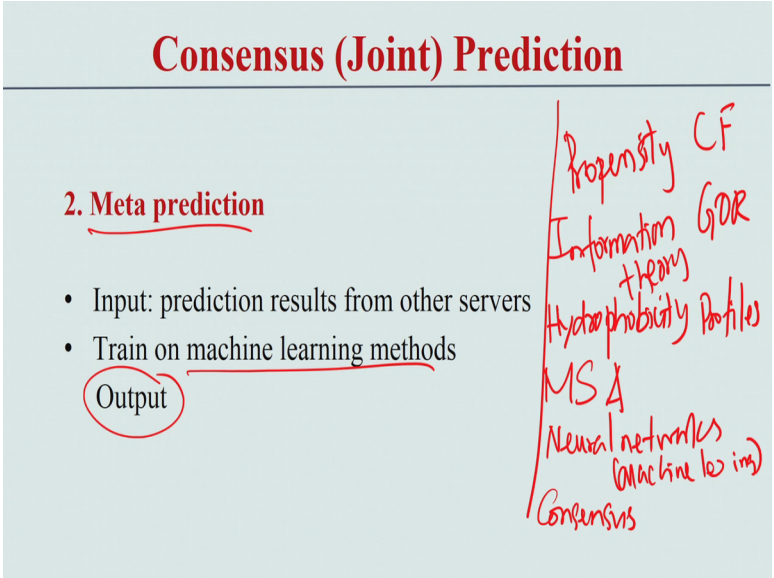
Information theory; this information theory. So, it predicts this as a beta strand and here this is the helix, some coil and the whole segment; we predicted as helix or stand or coil. And this is another method PHD; it is based on neural networks. So, this is the values here for example, I show only three methods; currently several methods are available and you can use as many methods as possible depending upon the performance and then see are there any consensus.

For example, this region if you see all the methods predict as helix. So, there is no conflict; so you can use this region as helix. And this region if you see this region; so, here this predicts as helix and this predict as helix, where as this predicts as coil. But if you take the majority of voting; out of 3; 2 predicted as helix. So, we can take this as helix, but we see the confidence; this maybe predicted a higher confidence than this region because here the probability is less compared with this region; this helical region.

Likewise, if we see here you can see this predicted as E; so, this is also assigned as E. Some cases is totally different now this as helix and this predicted as coil and this predicted in the beta strand. In this case, we do not know which one is this one is confusing; here we is only three cases; this is why it is confusing.

If we use 10 or 11 methods then you will have some sort of numbers and based on the numbers; we can assign the secondary structures and see which segment can be correctly predicted compared with the experimental data. So, this is a complete sequence; so now we will get the values and this method is called the consensus or you can see this is a joint prediction or the based on this voting example based method.

(Refer Slide Time: 17:12)



Second one we can use; the meta prediction there are several methods available. So, what they do? They first take a sequence; for example, if we take the sequence and predict; the secondary structures. They use these information as the input and train this data; using any machine learning and we get the output. In this case, it will consider this type of situations and they try to maximize the results because we know the experimental data to see how to assign the values, These meta predictions usually works better than other methods; any others individual methods which are reported in the literature.

So, there are several methods which individually predict; the main advantage of individual prediction, whether it is we can explain why the method predicts well, why the method fails. Because if the propensity is high and the region is beta strand; we know that this is high values because of that high value, it is wrongly predicted. Even if it is high values and if it is helix; then you can say here high propensity this is the reason, it could correctly identify the helix.

Likewise the information theory or the profiles; we can explain. When you go with the machine learning and go with these meta predictions; we get better results, but the problem is; it is very difficult to explain why this method; this protein prediction is very high, why this accuracy is low and so, on.

Summarizing, we discussed on various methods; what are the various methods we discussed?

Student: propensity

Right first one is the propensity; so, or we can see Chou Fasman and we discussed about the information theory.

Student: GOR.

Garnier; GOR then you put the hydrophobicity profiles; here we can predict with the single values as well as with the average windows and then we discussed about the?

Student: Multiple Sequence Alignment.

Multiple Sequence Alignment; MSA here also two ways you can assess the multiple sequence alignment; either take the parameters from the GOR or you can get the information from the experimental data, so you can predict the secondary structures.

So, also they found that the alignment grows; highly conserved then the prediction performance also increases. So, Jpred is the one or Zpred this first use multiple sequence alignment for predicting secondary structures. And then we discussed about neural networks or you can see the machine learning. Here they take the input information and they try to predict the output; that says secondary structures.

Then finally, we discussed about the consensus; the two types of consensus methods; one is based on ensembles.

Student: Ensemble

The ensemble based methods how it works?

Student: Majority of voting.

Majority of voting; it takes the several method methods and get the voting which one was the highest one; so, that is a secondary structure helix or strand or coil. In the case of the meta predictions; what they do?

Student: They train.

They train the output of the each single method and finally, use the information for predicting the desired output; this is how they do. So, at the present scenario we can see

the variations of the performance and we could see about 75 to 80 percent accuracy; now we can predict the secondary structures.

So, currently due to the advancements in the computing; say the big data; large amount of data and deep learning. They are trying to use deep learning in almost all prediction purposes including protein secondary structure prediction. They are trying to map this information and then try to improve the prediction performance.

So, we discussed about the secondary structures; in the next classes, we will go through the tertiary structures. So, what are the major information we obtain from protein 3D structures? What are the various databases? And how can we use the information available in protein 3D structures? To derive any properties and how can we use these properties with the different obligations for the predictions or have to understand the folding and the binding and so on.

Thanks for your kind attention.