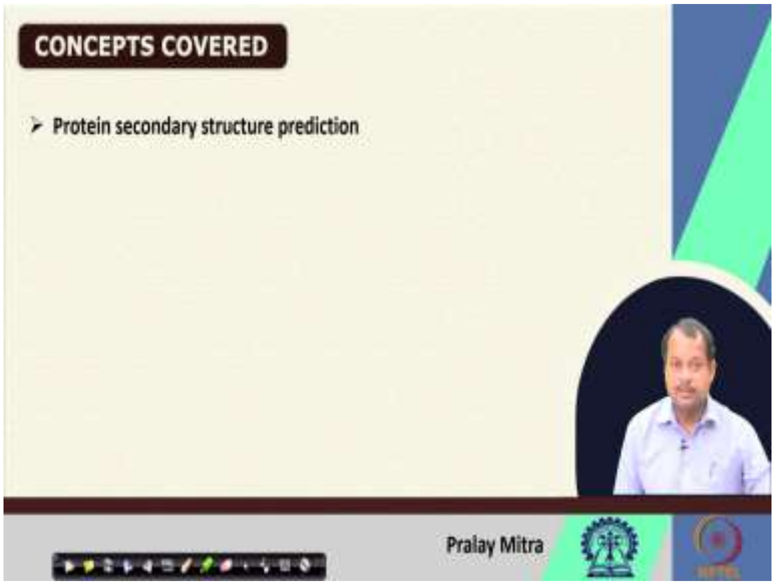
Algorithms for Protein Modelling and Engineering
Professor Pralay Mitra
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur
Lecture 54
Machine Learning to Predict the Secondary Structure from Amino Acid Sequences (Contd.)

Welcome back. So, we are continuing our discussion on predicting the secondary structure from the protein sequence. And for that we mentioned that machine learning technique specifically the neural network technique will be specifically useful for our own purpose.

(Refer Slide Time: 00:34)

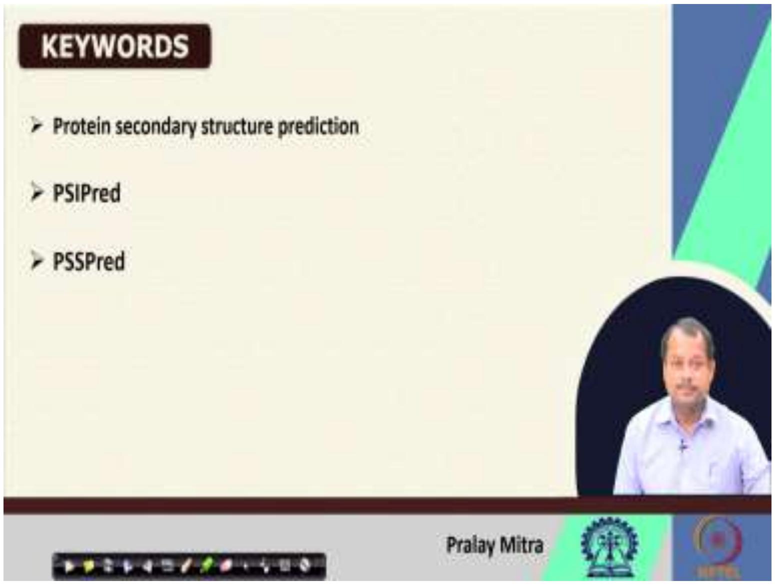


CONCEPTS COVERED

- Protein secondary structure prediction

So, that is why the concept that we will be covering is protein secondary structure prediction.

(Refer Slide Time: 00:38)



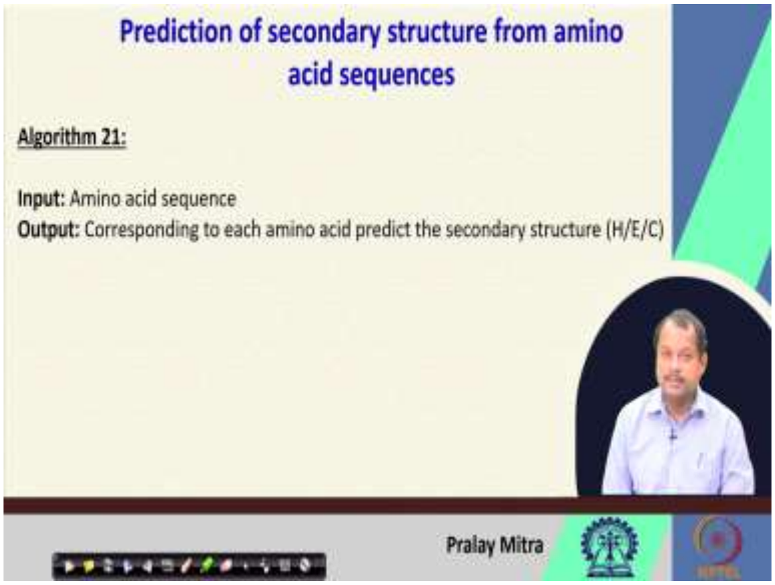
KEYWORDS

- Protein secondary structure prediction
- PSIPred
- PSSPred

Pralay Mitra

And the key word is also same here.

(Refer Slide Time: 00:41)



Prediction of secondary structure from amino acid sequences

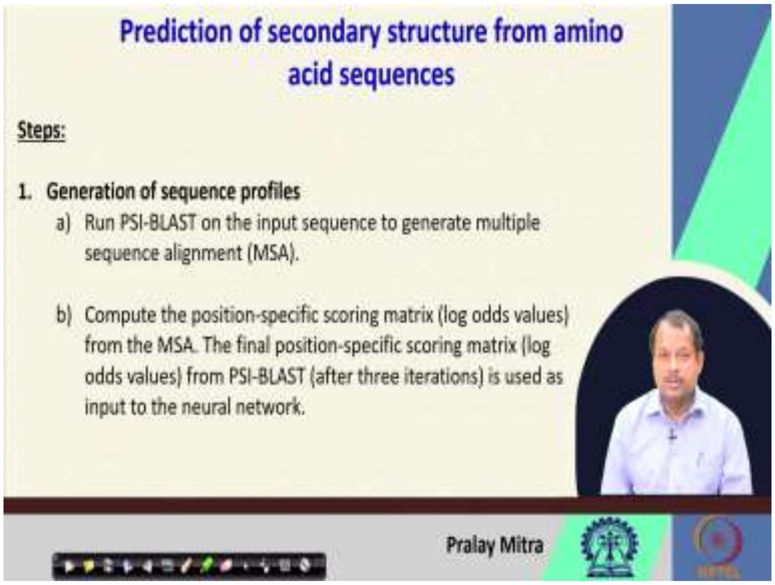
Algorithm 21:

Input: Amino acid sequence
Output: Corresponding to each amino acid predict the secondary structure (H/E/C)

Pralay Mitra

Now, we are discussing our 21st algorithm and that is the input is amino acid sequence and output is corresponding to each amino acid predict the secondary structure, helix, sheet or coil that is our interest, we are not going further details.

(Refer Slide Time: 00:58)



Prediction of secondary structure from amino acid sequences

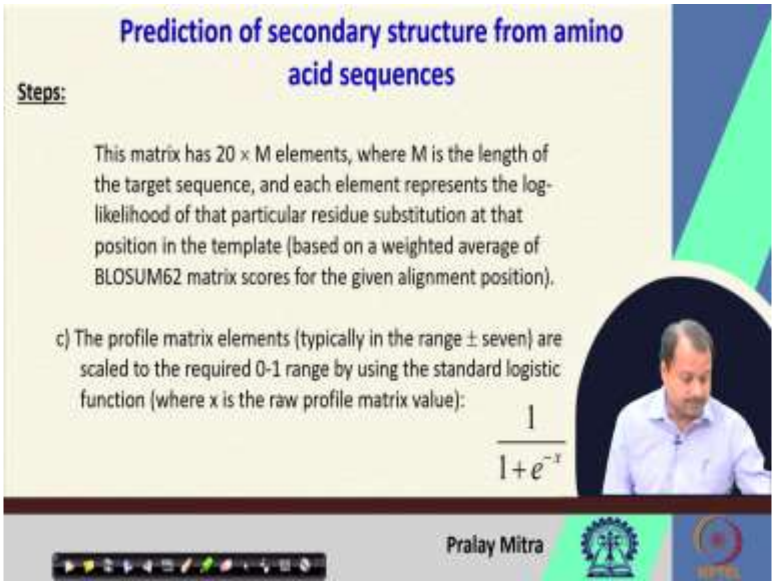
Steps:

1. **Generation of sequence profiles**
   a) Run PSI-BLAST on the input sequence to generate multiple sequence alignment (MSA).

   b) Compute the position-specific scoring matrix (log odds values) from the MSA. The final position-specific scoring matrix (log odds values) from PSI-BLAST (after three iterations) is used as input to the neural network.

So, here on the last lecture also I mentioned that given a protein sequence basically, first we will run psi-blast and three iterations used, then we will compute the position specific scoring metrics which are the log odd values from that psi-blast. So, if we run psi-blaster then usually we got one file, and that file is stored as a temporary file, so if you extract the temporary file, I mean after the running of the psi-blast, if you do not delete that one, then you can extract that one which contents which is on specific scoring matrix.

(Refer Slide Time: 01:37)



Prediction of secondary structure from amino acid sequences

Steps:

This matrix has 20 × M elements, where M is the length of the target sequence, and each element represents the log-likelihood of that particular residue substitution at that position in the template (based on a weighted average of BLOSUM62 matrix scores for the given alignment position).

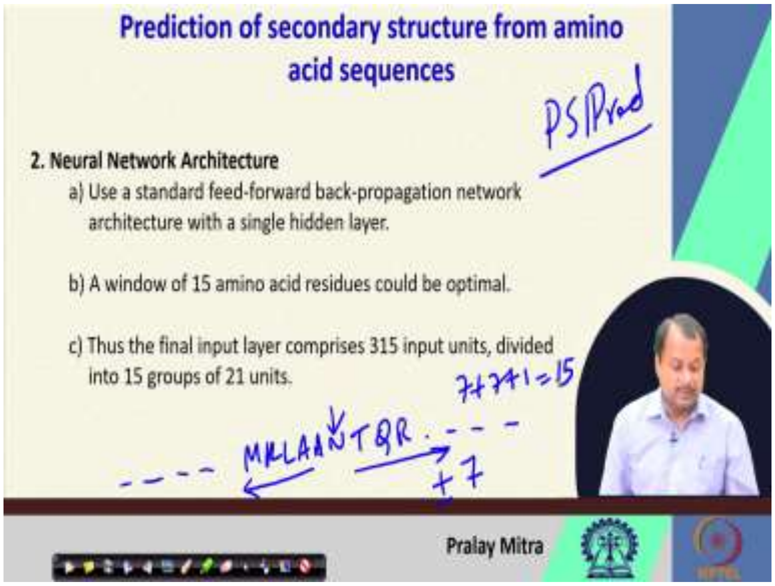
c) The profile matrix elements (typically in the range ± seven) are scaled to the required 0-1 range by using the standard logistic function (where x is the raw profile matrix value):

$$\frac{1}{1+e^{-x}}$$

This matrix has 20 cross M elements, why 20 cross M elements? You know this by this time, because it is a position specific scoring matrix in each say in this case 20 cross M I mention which means in each row there is one amino acid and in each column there are M there are positions of the protein sequence, I am assuming that length of the input protein sequence is M, that is why it is the length.

Where M is the length of the target sequence and each element represents the log likelihood of that particular residue substitution at that position in that template, based upon a weighted average of Blosum62 matrix scores for the given alignment position. Now, the profile matrix element typically in the range plus minus 7, so that is that window we are considering are scaled to the required 0-1 range by using the standard logistic function where x is the row profile matrix value, and we are using as a sigmoid function for this purpose.

(Refer Slide Time: 02:47)



Prediction of secondary structure from amino acid sequences

PSIPred

2. Neural Network Architecture

a) Use a standard feed-forward back-propagation network architecture with a single hidden layer.

b) A window of 15 amino acid residues could be optimal.

c) Thus the final input layer comprises 315 input units, divided into 15 groups of 21 units.

7+7+1=15

- - - - MKLAANTQR . - - -

+7

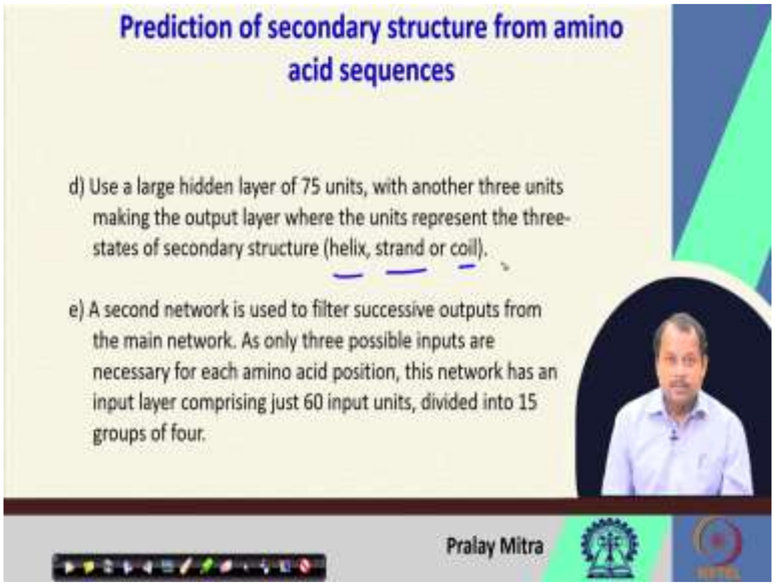
Pralay Mitra

In a neural network architecture part, so definitely there is a scope for a customization, you can definitely do something better, but in most widely used algorithm PSIPred, what is there? So, PSIPred what is mentioned is a standard feed forward back propagation network, so architecture with a single hidden layer, a window of 50 amino acid residues could be optimal.

So, this 50 amino acid residue a 15 amino acid residue indicates that say if I have a KLAANTQR and I am interested to assign the secondary structure for say asparagine N, then on the left and on the right I will go plus minus 7. So, plus minus 7 which means 7 plus 7 plus 1 point of interest or residue or amino acid of interest that is my 15.

So, 50 amino acid residue could be optimal, so that is fixed by doing some trial and error technique, does the final input layer comprising 315 input units it divided into 15 groups of 21 in it each. So, you can see that 21 multiplied with 15 that is basically is that length of the input. Now, in the sorry…

(Refer Slide Time: 04:23)



Prediction of secondary structure from amino acid sequences

d) Use a large hidden layer of 75 units, with another three units making the output layer where the units represent the three-states of secondary structure (helix, strand or coil).

e) A second network is used to filter successive outputs from the main network. As only three possible inputs are necessary for each amino acid position, this network has an input layer comprising just 60 input units, divided into 15 groups of four.

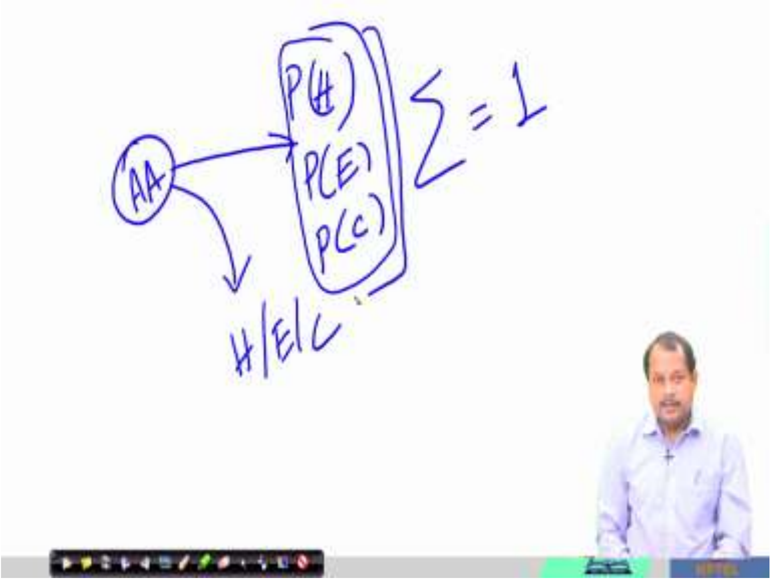
Next, use a large hidden layer of 75 units, with another three units making the output layer, so we need the three states, helix stand and our helix sheet and coil. Making the output layer where the units present that three states of the secondary structure helix stand or coil.

A second can network is used to filter successive outputs from the main network as only three possible inputs are necessary for each amino acid position, this network has an input layer comprising just 60 input units divided into 15 groups of 4. So, 15 groups of 4 and 60 input units that has been used.

So, now it is clear to you that how what is the size of the input layer, how many hidden layers, what is the size of the output layer and how many hidden nodes are there So, everything is not clear to you. So, that is about the, that is about the considering of one hidden layer but you are free to use more than one hidden layer also.

So, if you use then accordingly you have to adjust that in each hidden layer how many nodes will be there, and what will be the mapping for that one. Now, based upon this one if you predict then definitely you will get some scoring value I mean some score value corresponding to three states H, E or C, helix, stand or coil corresponding to corresponding to each amino acid. Now, what you are getting?

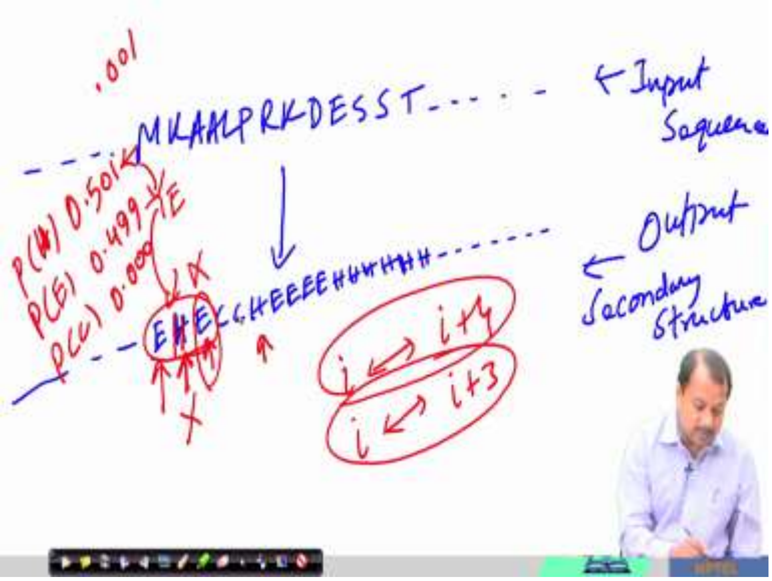
(Refer Slide Time: 06:11)



So, corresponding to each amino acid say one amino acid corresponding to that you are getting, so helix probability, probability of sheet and probability of coil, and I mentioned that if you take the summation over this you are going to get 1. So, that is a total probability is 1. Next, the simple thing which you may think that okay, so I am getting say the probability of three states, then how do I decide which state should I return?

Very simple, we will use the threshold value or we can use the highest probability among these three. What is the highest probability? We will map that 1 to that probability say H, or E, or C, that way this particular neural network technique will give corresponding to one input sequence the secondary structure state for your sequence.

(Refer Slide Time: 07:22)



Now, when you will get then definitely you can think of sorry you can think of that let us assume one hypothetical sequence say MKAALPRKDESST, so that is my input sequence. Output secondary structure say corresponding to M say you can get this. Now, I wish to draw your attention about one thing when this is your prediction look. What it does?

It uses a simple neural network method which I believe is exploiting as much info homologous information and the existing information as possible. Now, there is a scope of improvement or further improvement, there is an I am talking is that if you look at this prediction, although I mentioned that this is one hypothetical prediction, but you cannot rule out the situation where there will be only one single H flank to I, or say on the two neighbor, either E or C, here is one situation, here is another situation, so one residue is forming the secondary structure.

Again, I am telling, one residue is forming the secondary structure; do you think it is valid? Look, what we discussed is that in order to form the secondary structure at the main chain, there must be hydrogen bond between i and i plus 4 or i and i plus 3. And if it is then everything inside this i to i plus 4 or i to i plus 3 will be included as part of the helix, then how come this H will come alone acting as one helix, one amino acid, how it can be helix?

It may possible that can be sequential, but how one amino acid can be a helix? Or here how one amino acid can be one sheet? For sheet also some starch is required. So, one random hydrogen bond cannot say that there is a pattern, because yes the helix and sheet content some pattern in the secondary structure, so where is the pattern? So, pattern is missing, if the pattern is missing then how do I say that this is going to be a valid secondary structure?

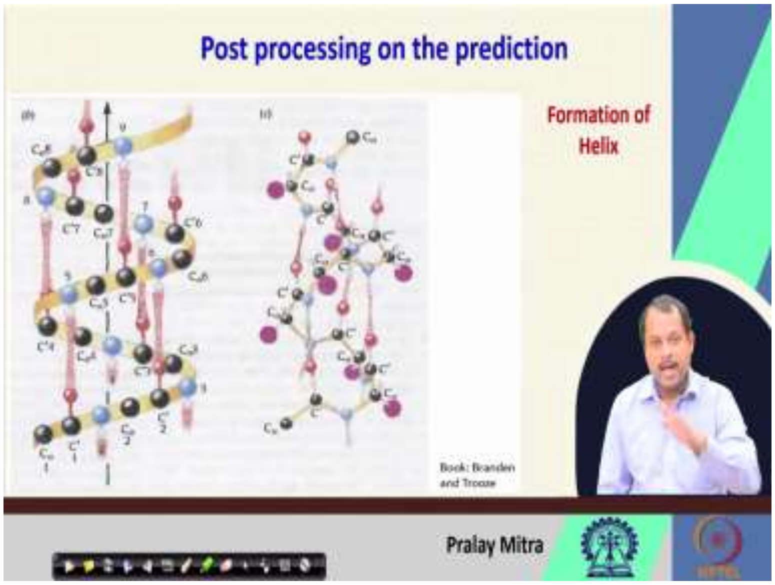
So, taking the observation and that is why time to time we are discussing little bit of chemistry or biology etcetera, because if we exploit that domain knowledge then we will readily say it cannot be an alone sheet, it alone cannot be an alone helix. So, what are the possibilities? Since this H is actually flanked by or say sandwiched in between two E there is a possibility it can also be E.

Although optionally you can look at their probability of value because since out of three probability values P of h, P of E and P of C, I pick the best one or the highest one, it may possible since everything is done through the algorithm, there is no manual intervention or curation. So, if it is manual then perhaps by looking at that we can make some different comment, but since we pick based upon say highest value, so if what if one highest value is a say 5 point 0.501 and this is 0.499, this is absolutely 000.

If you sum it up then you will see it is basically 1. And because this is of high value that is why it is has been picked, but you see what is the difference, so 0.001 very minor very small differences there. So, whether that small differences is say our competition error, so numerical approximation error or say it is an information it is hard to say specifically when you will see like this something like this, then it might be a good idea that why not think that okay this is also going to be also part of E.

So, there is a scope once you will get the prediction by the sequence you go for a post processing, based upon your observation of the biology.

(Refer Slide Time: 12:59)


Post processing on the prediction

Formation of Helix

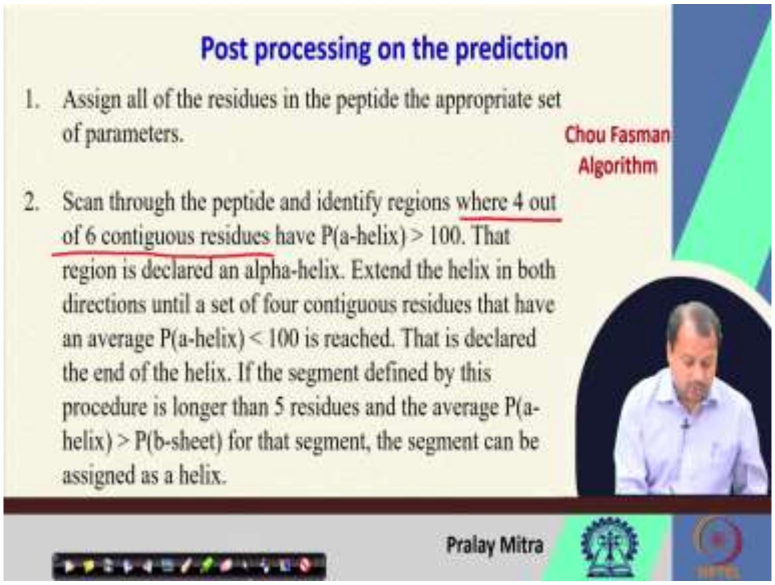
Book: Branden and Tooze

Pralay Mitra

That is what is mentioned here, post processing on the prediction, because we know what is the rule of logic for the formation of the helix. So, I mention when one H was placed between two E, it can be opposite also there are several H, in between suddenly there is one C or E most likely they are not C or E.

So, if it is the situation we understand that there is a need of post processing, but when we will do the post processing definitely we will look at the sequence again and all those singles such occurrences in a stretch of say H or E the regular pattern will be erased then it will be better if we have a rule or logic behind that's one. In order to have that rule or logic we have to go back long back.

(Refer Slide Time: 13:57)



Post processing on the prediction

1. Assign all of the residues in the peptide the appropriate set of parameters.
2. Scan through the peptide and identify regions where 4 out of 6 contiguous residues have P(a-helix) > 100. That region is declared an alpha-helix. Extend the helix in both directions until a set of four contiguous residues that have an average P(a-helix) < 100 is reached. That is declared the end of the helix. If the segment defined by this procedure is longer than 5 residues and the average P(a-helix) > P(b-sheet) for that segment, the segment can be assigned as a helix.

Chou Fasman Algorithm

Pralay Mitra

So, Chou Fasman give the first algorithm on this secondary structure prediction. So, what they have done, we are not going into details of that one because their accuracy is very less compared to the accuracy of any machine learning based secondary structure prediction algorithm. So, currently, you can consider that as just a historical value, so that is why I am not going into details but few steps I would like to mention.

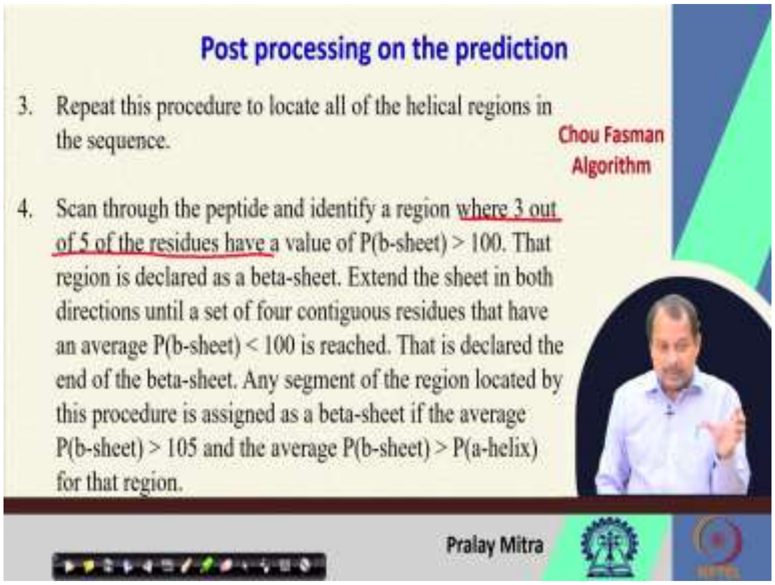
So, that we can have an idea that if I go for post processing, then perhaps what will be my rule or logic, because although with respect to the neural network technique, their accuracy is very less, but if you look at the way they have done it then there accuracy is not something you should ignore. So, what Chou Fasman has been has done by looking at the secondary structure, they computed the probability of the occurrences of the helix probability of occurrence of the sheet probability of occurrence of coil and different turns corresponding to each amino acid, and they have one table.

If you look if you search then you will have two three different variations of the Chou Fasman one is of course given by them and since they computed that probability based upon their data set, so later people computed on larger data set and got some different value, but if you stick on to one table one value then first it says that assign all the residues in the peptide the appropriate set of parameters, that means all the appropriate probability values.

Then scan through the peptide and identify regions where 4 out of 6 contiguous residues have probability of helix greater than 100. So, this probability of helix greater than 100 is his contribution along with one thing which will be useful for us were 4 out of 6 contiguous residues. So, he mentioned that okay, so if there are 6 residues out of which 4 are helix then perhaps that is part of the helix.

Then one more thing which will be important for us, if the segment defined by procedure is longer than 5 residues and average P helix is greater than P beta sheet, the segment can be assigned as helix, which means if there is a tie between the helix and the sheet, so you are getting helix this way next extend the helix in both direction until a set of 4 contiguous residues that have average less than… is reached. So, we are not interested.

(Refer Slide Time: 16:47)



Post processing on the prediction

3. Repeat this procedure to locate all of the helical regions in the sequence.

Chou Fasman Algorithm

4. Scan through the peptide and identify a region where 3 out of 5 of the residues have a value of P(b-sheet) > 100. That region is declared as a beta-sheet. Extend the sheet in both directions until a set of four contiguous residues that have an average P(b-sheet) < 100 is reached. That is declared the end of the beta-sheet. Any segment of the region located by this procedure is assigned as a beta-sheet if the average P(b-sheet) > 105 and the average P(b-sheet) > P(a-helix) for that region.

Pralay Mitra

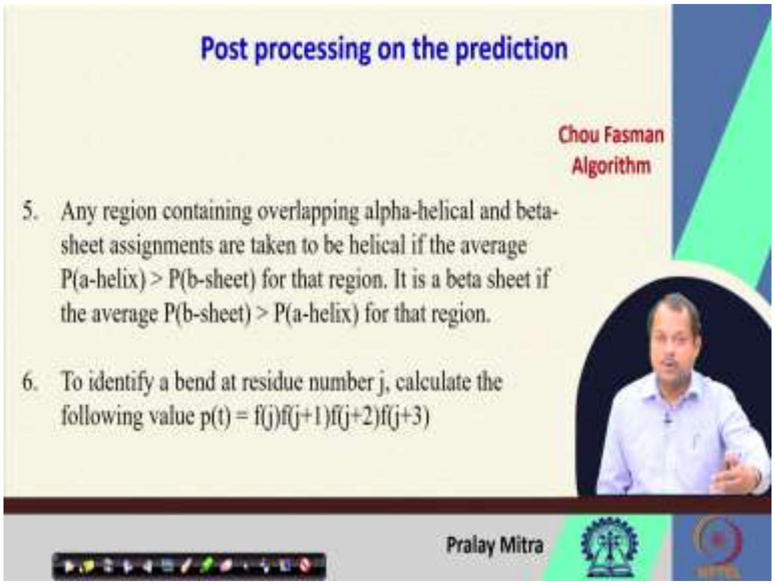
After that one similar to that, so one more role he has established for the basically beta sheet also, it is placed in the exactly same, so same underline is working here also, interesting. So, repeat this procedure to locate all the helical regions, done, so we are not using his algorithm. Now, forth step we are looking at scan through the peptide and identify region of 3 out of 5. So, last time it was 4 out of 6 here 3 out of 5, to have a value of beta sheet greater than 100.

So, you will have that 100, that region is declared a beta sheet. So, if in your prediction in your predicted sequence if you find that in a stretch actually you are having 4 helix but 2 are say if those 2 are at the end so then I cannot conclude, but if you go by sliding window then 6 6 6 6 6 is that like that way and if you got that 4 is helix and in between 1 or 2 is going to be coil or sheet then perhaps that is also part of the helix, it was misinterpreted. So, you can include that.

The same thing can be applied for the beta sheet also, if out of 5, 3 are declared as beta sheet then rest 2 can also be maybe part of the beta sheet. Now, when you are proceeding this way then there is a possibility that okay, so, if there is a tie between say whether it is going to be the helix or beta sheet then mostly you go in the favor of say helix.

And because of that strict hardy morning pattern, you will see individually if you compute the accuracy of say alpha helix, beta sheet and the coil, then you will find the individual accuracy of the helix is far better compared to the beta sheet which stands second, and rest is coil.

(Refer Slide Time: 18:58)



Post processing on the prediction

Chou Fasman
Algorithm

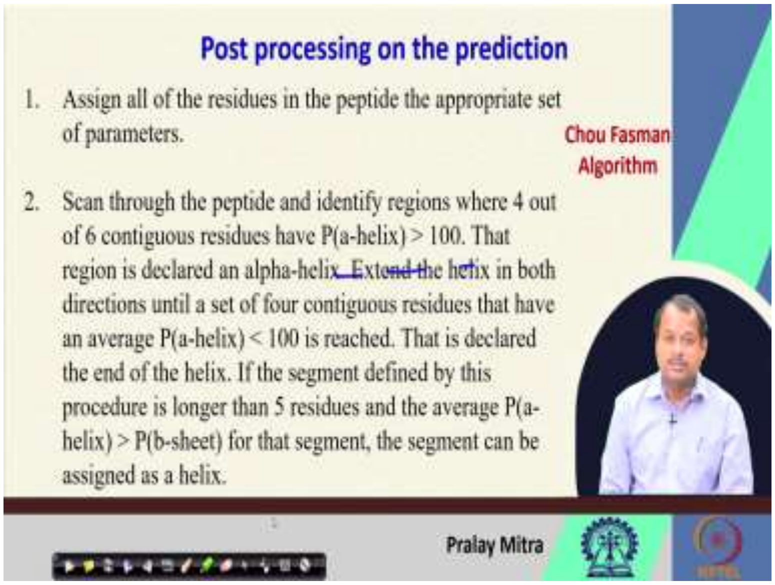
5. Any region containing overlapping alpha-helical and beta-sheet assignments are taken to be helical if the average P(a-helix) > P(b-sheet) for that region. It is a beta sheet if the average P(b-sheet) > P(a-helix) for that region.

6. To identify a bend at residue number j, calculate the following value p(t) = f(j)f(j+1)f(j+2)f(j+3)

Pralay Mitra

Here this particular observation and the Chow Fasman's work which although has say historical value can be useful for you. So, look at the fifth step, any region containing overlapping, so that overlapping information also he considered, alpha helical and beta sheet assignment are taken to a helical if the average alpha P alpha helix sheet greater than P beta set for that region if it is not then go for beta sheet. Now, for you, so this probability P is computed by him.

(Refer Slide Time: 19:41)



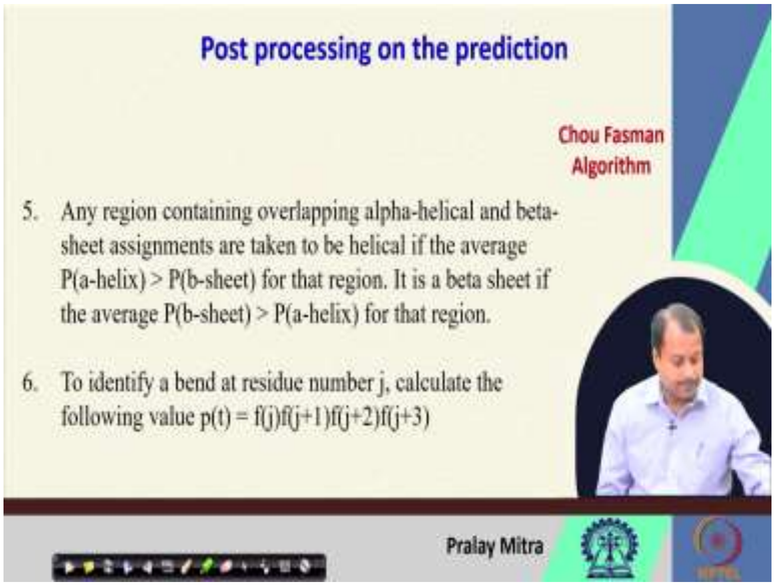
Post processing on the prediction

Chou Fasman Algorithm

1. Assign all of the residues in the peptide the appropriate set of parameters.
2. Scan through the peptide and identify regions where 4 out of 6 contiguous residues have P(a-helix) > 100. That region is declared an alpha-helix. Extend the helix in both directions until a set of four contiguous residues that have an average P(a-helix) < 100 is reached. That is declared the end of the helix. If the segment defined by this procedure is longer than 5 residues and the average P(a-helix) > P(b-sheet) for that segment, the segment can be assigned as a helix.

And then the first step of his algorithm, so he mentioned to he mentioned to assign that one first step.

(Refer Slide Time: 19:48)



Post processing on the prediction

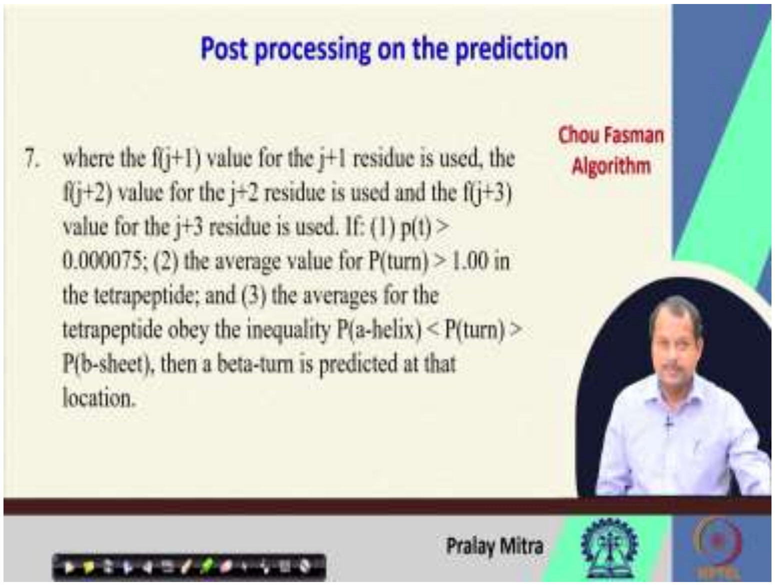
Chou Fasman Algorithm

5. Any region containing overlapping alpha-helical and beta-sheet assignments are taken to be helical if the average P(a-helix) > P(b-sheet) for that region. It is a beta sheet if the average P(b-sheet) > P(a-helix) for that region.
6. To identify a bend at residue number j, calculate the following value p(t) = f(j)f(j+1)f(j+2)f(j+3)

Now, here very much what you can do that you can use your own probability value that is predicted by your neural network technique, you use that probability value and then you check that for the over lapping situations whether it is the probability of the helix which is high or the probability of the beta sheet, then you can consider say the average for that entire stretch and then you can check.

So, he has some further steps, so just for completion since I am writing that it’s Chou Fasman algorithm, so I kept it here, but for post processing purpose actually our interest will be to say to remove the spikes or say small noises which will be incorporated because you know that only one residue per say surrounded by sheet or coil cannot act as a helix, it is true for sheet also.

(Refer Slide Time: 20:51)



Post processing on the prediction

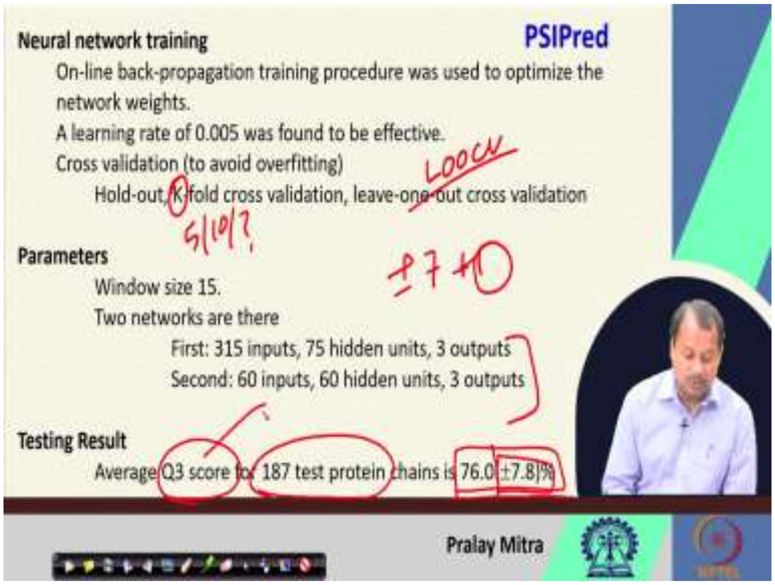
Chou Fasman Algorithm

7. where the $f(j+1)$ value for the $j+1$ residue is used, the $f(j+2)$ value for the $j+2$ residue is used and the $f(j+3)$ value for the $j+3$ residue is used. If: (1) $p(t) > 0.000075$; (2) the average value for $P(turn) > 1.00$ in the tetrapeptide; and (3) the averages for the tetrapeptide obey the inequality $P(a\text{-}helix) < P(turn) > P(b\text{-}sheet)$, then a beta-turn is predicted at that location.

Pralay Mitra

So, this is the final step of the Chou Fasman algorithm.

(Refer Slide Time: 20:55)



Neural network training PSIPred

On-line back-propagation training procedure was used to optimize the network weights.

A learning rate of 0.005 was found to be effective.

Cross validation (to avoid overfitting)

Hold-out, K-fold cross validation, leave-one-out cross validation

LOOCV

5|10|?

Parameters

±7+1

Window size 15.

Two networks are there

First: 315 inputs, 75 hidden units, 3 outputs

Second: 60 inputs, 60 hidden units, 3 outputs

Testing Result

Average Q3 score for 187 test protein chains is 76.0±7.8%

Pralay Mitra

So, apart from the algorithm that we discussed is actually exploited by the PSIPred, the neural network training, so online back propagation training procedure was used to optimize the network weights. A learning rate of 0.005 was found to be effective for them and for the cross validation to avoid over fitting, holdout, key fold cross validation, leave one out cross validation, or LOOCV that we also have done for protein engineering. So, those things they have done.

So, for this K fold, so this K can be 5, 10 or any other thing. Now, the parameter, so window size 15, how it is 15 I mentioned plus minus 7 plus 1. So, this plus 1 is the amino acid for which I am currently doing the prediction plus minus is on the left hand side plus left minus on the right hand side plus, now for the boundary condition definitely you have to take care of that one say for first 7 residues you will not get say minus 7, last seven residues you will not get plus 7.

So, for that you have to treat it separately. Now, two networks are there for PSIPred, the first network takes 315 inputs, 75 hidden units, 3 outputs, another is the 60 inputs, 60 hidden units and 3 outputs and finally, they are taking the final output combining these two information. Next in our testing results, so average Q3 score for 187 at test protein chains is 76.

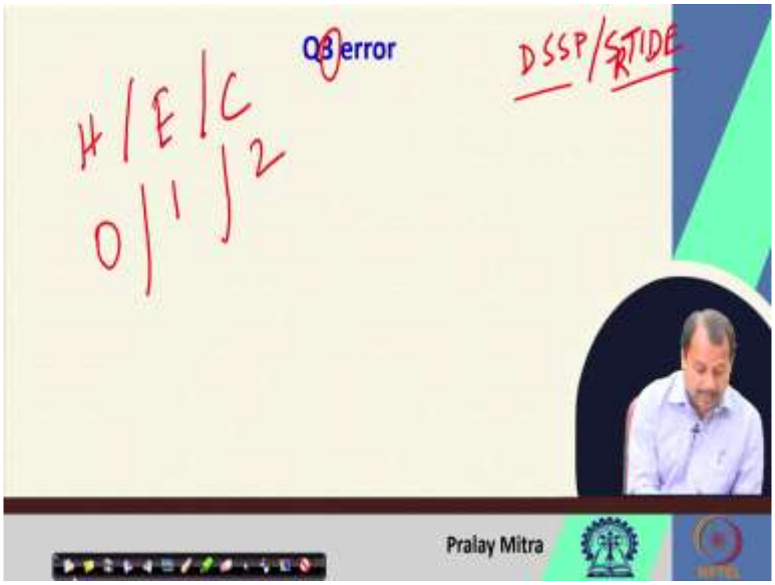
So, this number, so since I am reporting this separate algorithm they are parameters the features that is why I am reporting but you know that this value will very much depend upon the data set that you are using. But when you are using this one, so why I put it for what that, so you have to report the mean and when you are reporting the mean then you have to report the standard deviation.

So, within parenthesis the plus minus 7.8 indicates the standard deviation for their amount of error, that means 187 test products are there. Now, what is this Q3 score? So, that is an interesting thing. So, we will conclude today's discussion after discussing this Q3 score. Now, I believe up to this one it is very much clear to all of you that say when we are predicting the secondary structure from a protein sequence then how to move on.

And it will be utilized for our next topic also, and you will also see that some temporary files which will be generated by this PSIPred specifically the psi-blast which we will generate will be useful for us, because what psi-blast is doing that after identifying the homologous sequences it performs multiple sequence alignment.

Computes position specific scoring matrix which contents the law (())(24:15) values and sometimes that will be weighted by the Blossom62 matrix. We see that this kind of position specific scoring matrix has wide application, one application was also demonstrated in protein design. Now let us move on to this Q3 score, what is that?

(Refer Slide Time: 24:39)

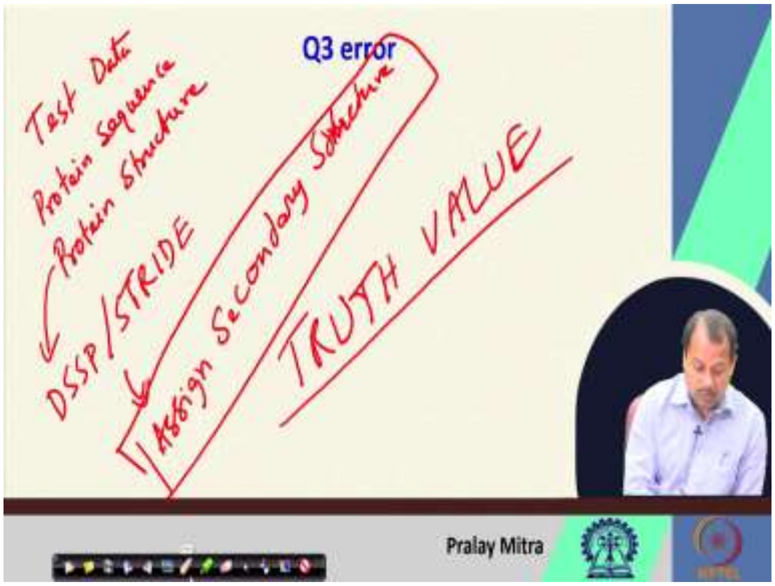

Q3 error

So, Q3 error: So, first of all, when I say Q3, so it is basically related with three states, three states, H, E, C, which you may also report as 0, 1, 2. So, this 0, 1, 2 or H, E, C three states are there, since it is three, so these 3 term has come from there. Now, what do I mean this Q3 or when say you are predicting the secondary structure then how to estimate the error? So, the simple thing is that, what will be the definition of your error?

And that is the first one, because when I am predicting then corresponding to one amino acid sequence I have H, E, or C. Now, that H, E, or C whether that is going to be the correct one or not, who is going to tell me that? For this we will assume perhaps these DSSP or stride which is developed based upon the protein structure is the correct one, we will assume that that is the correct one.

Although while discussing these DSSP and stride we see that from accuracy point of view there is a small variation between these two and also I cannot say that they are 100 percent correct, but you know that when you are computing some error, so that error must be measured with respect to something and here that respect I am considering DSSP or stride so stride sorry.

(Refer Slide Time: 26:53)

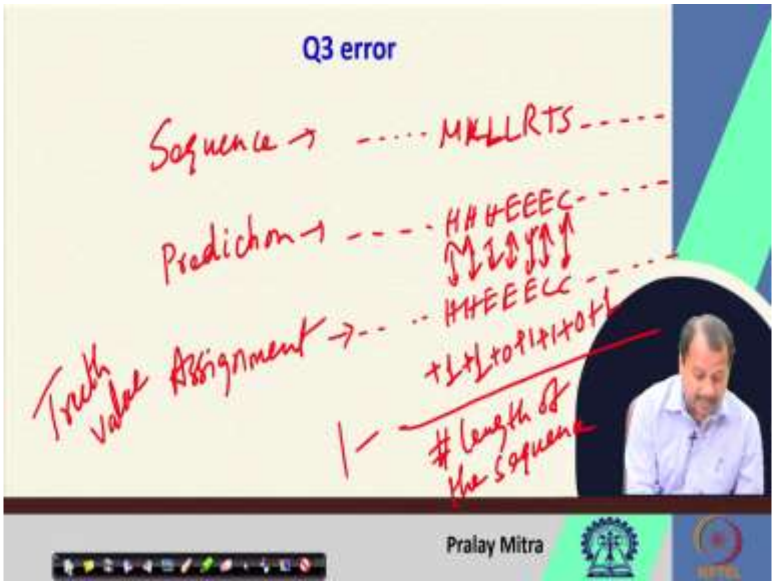


Q3 error

So, now, when this DSSP or stride is there, let us go back to my protein sequence although for the test set or the test data, I will get only the protein sequence, but when say I am benchmarking my algorithm then I must have or I should have some test data for that test data I will have protein sequence and protein structure, I will have protein sequence and protein structure.

Now, from this protein structure either using say DSSP or stride I can assign I can assign secondary structure, let us assume that this is my truth value, if it is my truth value, if it is my truth value, then I will start from here

(Refer Slide Time: 28:07)



Q3 error

Soquence → ..... MKLLRTS-----

Prediction → ----HHHEEEC-----

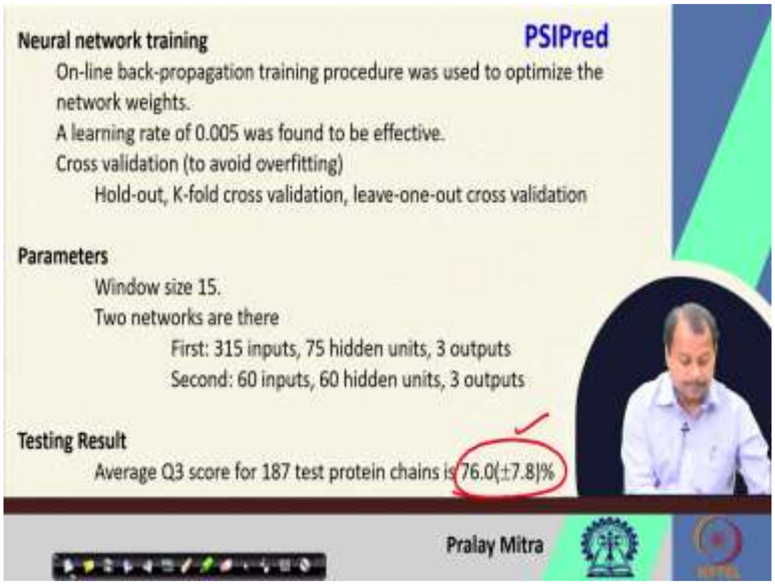
Truth value Assignment →... .. HHEEECC-----

$$1 - \frac{+1+1+0+1+1+0+1}{\#\text{length of the sequence}}$$

Pralay Mitra

So, given one sequence I have a prediction, of course I assume that post processing is also done, and I have one assignment that I am considering as to truth value. Now, I will check if it matches plus 1 plus 1 mismatches plus 0 matches plus 1 matches plus 1 mismatches 0 matches plus 1 normalized by the length, that is the amount of matching. Now, if you say take subtraction from 1 or say if you compute the mismatches only normalized by the length of the sequence then you will get Q3 error.

(Refer Slide Time: 29:42)



Neural network training

PSIPred

On-line back-propagation training procedure was used to optimize the network weights.

A learning rate of 0.005 was found to be effective.

Cross validation (to avoid overfitting)

Hold-out, K-fold cross validation, leave-one-out cross validation

Parameters

Window size 15.

Two networks are there

First: 315 inputs, 75 hidden units, 3 outputs

Second: 60 inputs, 60 hidden units, 3 outputs

Testing Result

Average Q3 score for 187 test protein chains is 76.0(±7.8)%

Pralay Mitra

In the previous slide, so which is mention it is the testing results, this is the accuracy, which means the matching divided by the normalized divided by the length multiplied with 100 that is going to be the value. So, that is it for today's lecture. Thank you very much for your attention.