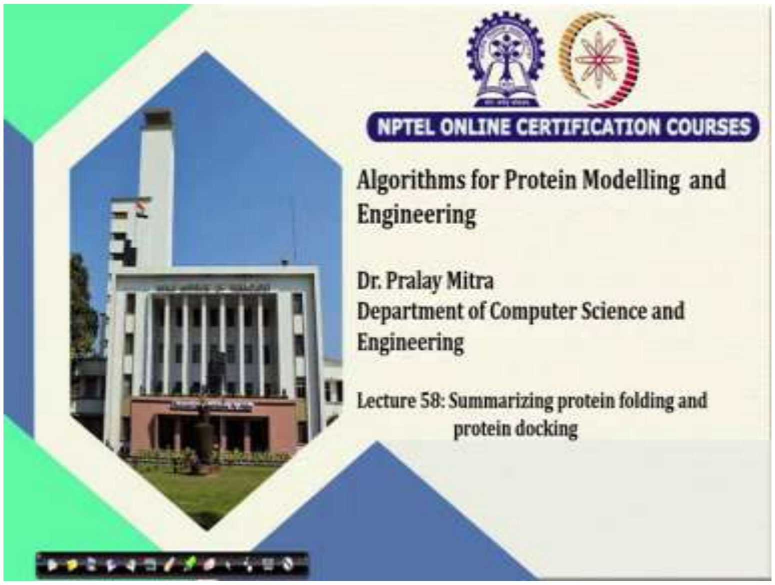
Algorithms for Protein Modelling and Engineering

Professor Pralay Mitra
Department of Computer Science and Engineering,
Indian Institute of Technology Kharagpur

Lecture 58
Summarizing Protein Folding and Protein Docking

Welcome back to this Algorithms for Protein Modeling and Engineering class. So, we reached almost at the end of this course.

(Refer Slide Time: 00:18)



NPTEL ONLINE CERTIFICATION COURSES

Algorithms for Protein Modelling and Engineering

Dr. Pralay Mitra
Department of Computer Science and Engineering

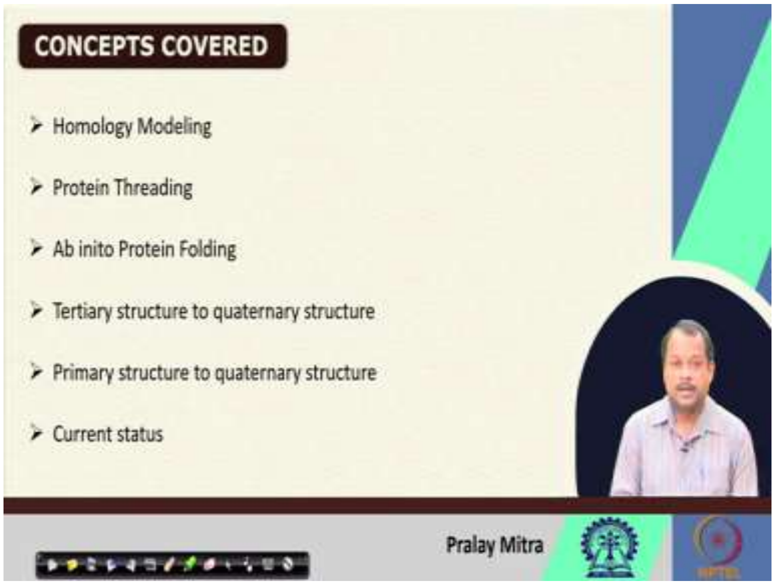
Lecture 58: Summarizing protein folding and protein docking

So, in these few remaining lectures I am planning to summarize what we have discussed so far and also we wish to give some insight, because within this timeframe it is not possible to cover all the softwares or techniques which exist. So, what we did actually we pick some of the algorithms and then we discussed that one, we demonstrated some of the applications, nevertheless some other software's are there.

So, grossly we discussed protein folding, protein docking and protein engineering, where the design, protein modification and protein design is included and along with that one as per the requirement so other small, small things we also have discussed. So, grossly it is protein docking, design and folding.

So, we wish to summarize on these three topics and we wish to tell you that some other tools or techniques which exists that may be useful for you if you just apply for your own purpose, but if you wish to customize or modify then since you know or you have sufficient amount of knowledge or background based upon the classes or the lectures you have done so far then I believe you can able to do that one.

(Refer Slide Time: 01:37)

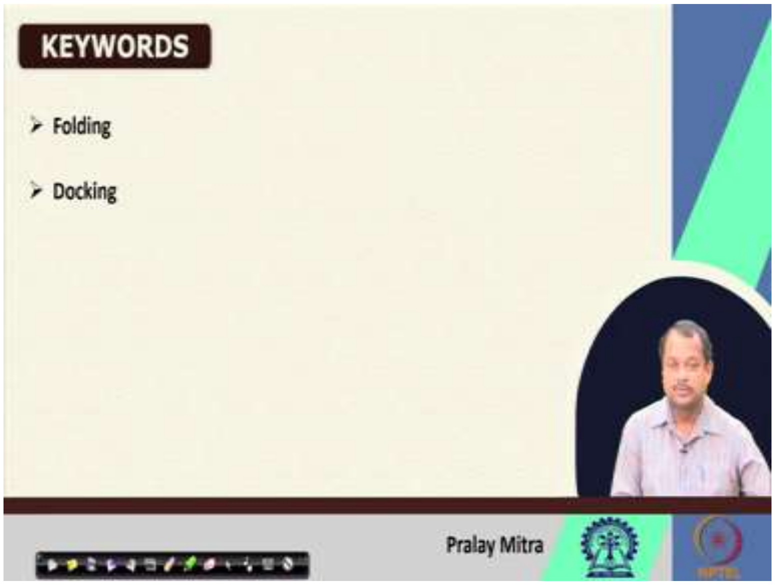


CONCEPTS COVERED

- Homology Modeling
- Protein Threading
- Ab inito Protein Folding
- Tertiary structure to quaternary structure
- Primary structure to quaternary structure
- Current status

So, the concepts that we are planning to cover is homology modeling, protein threading, ab initio protein folding, tertiary structure to quaternary structure, primary structure to quaternary structure and what is the current status in this context.

(Refer Slide Time: 01:54)

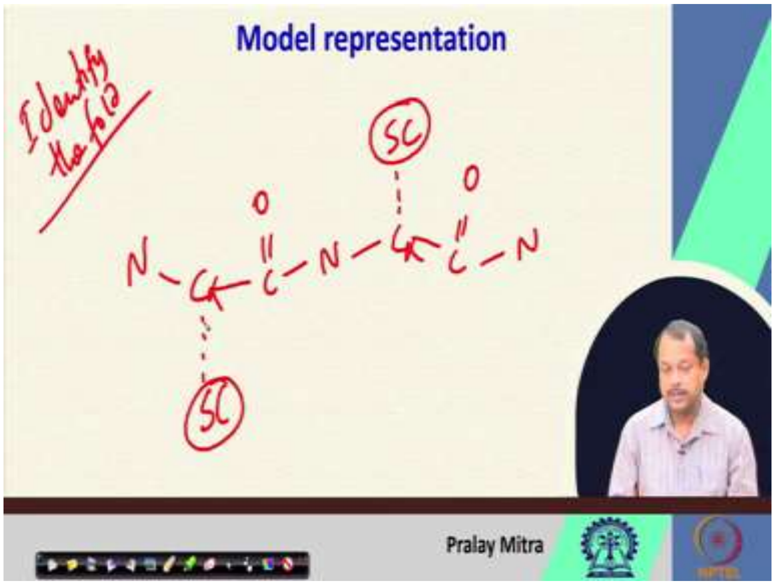


KEYWORDS

- Folding
- Docking

So, keywords I have picked is just folding and docking, so that it is more concise and concrete rather than diffused.

(Refer Slide Time: 02:02)



Model representation

Pralay Mitra

So, let us start with the protein folding. So, I mentioned that when we are doing the protein folding, so our main interest is to identify the fold. And although the protein side chain atoms has some role in the overall stability and the fold of the protein, but it is only the main chain atoms which contributes to the fold of that protein. And also, to improve the computation time or to make it faster, then usually in protein folding problem most of the time people took an approach that it will consider only the main chain atoms and then it basically attach the side chains later after the protein folding is done or the fold has been designed.

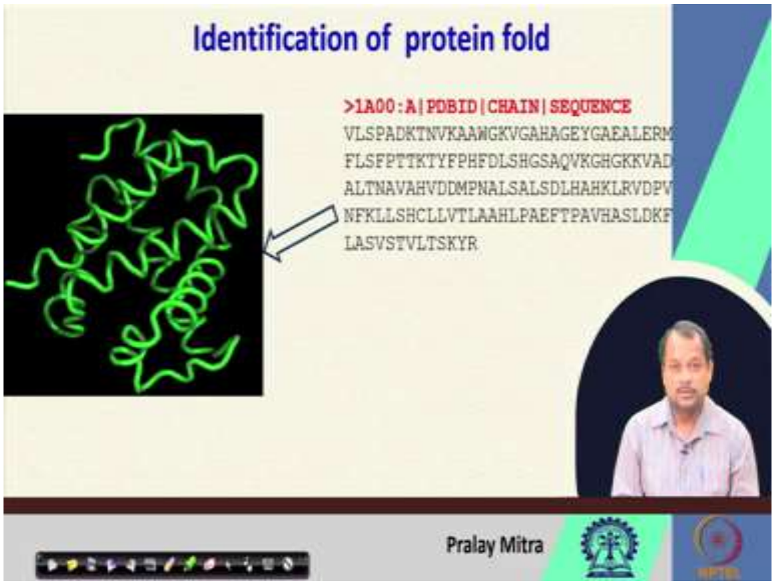
Now, if I say that main chain then it is starting with the central carbon, here there is N, then C double bond O, then N, then C, then one more C double bond O then N so like that it will go. And what is the side chain? So, specifically this is of interest. But at the same time to make it little more practical, so instead of stripping out all the side chains, the reason is very simple, because as by this time you understand that say there are 20 different amino acids and since 20 different amino acids are there and if you look at their side chain you will find that all the side chains are not same. Definitely from physico chemical property or from say biochemical property or say from geometric features you will see that they are completely different.

So, if they are different so it might not be a good idea to just a model the main chain. Because if you model the main chain, then you will actually boils down to the situation where eventually all the proteins are same. Because of the main chain level if you strip up or if you chop it out the side chain, then you will find that all the main chains are same for all the proteins. It is only the length by which it is differing, but most of the, many proteins may have the same length also, their actual the sequence is different, their structure is different. So, that is why instead of chopping it out completely, so this C actually is my C alpha, so I am giving side chain information here.

And in this case dotted line, so let us erase this one to avoid the confusion, so this dotted line indicates that actually there is no covalent bond, although I know that C alpha atom is connected with its side chain by the covalent bond, but here SC is actually representative one atom only. So, this indicates the centroid of all the side chain atoms. And when it is the centroid, then definitely the length of the centroid and the C alpha atom will not indicate the length of any covalent bond that is why it is not a good idea to give a solid line indicating it is a covalent bond. So, put some dotted line. And that way you can also carry on or you can also include the existence of the side chain but not with the same accuracy along with the say main chain.

So, for the main chain accuracy is very high. For my side chain accuracy is not that much high, but still I am bearing that information along with the main chain. With the hope letter I will make it perfect, but for the time it does focus only on the folding part. Now, once the fold is determined, then definitely my job will be to look at the side chain.

(Refer Slide Time: 06:18)



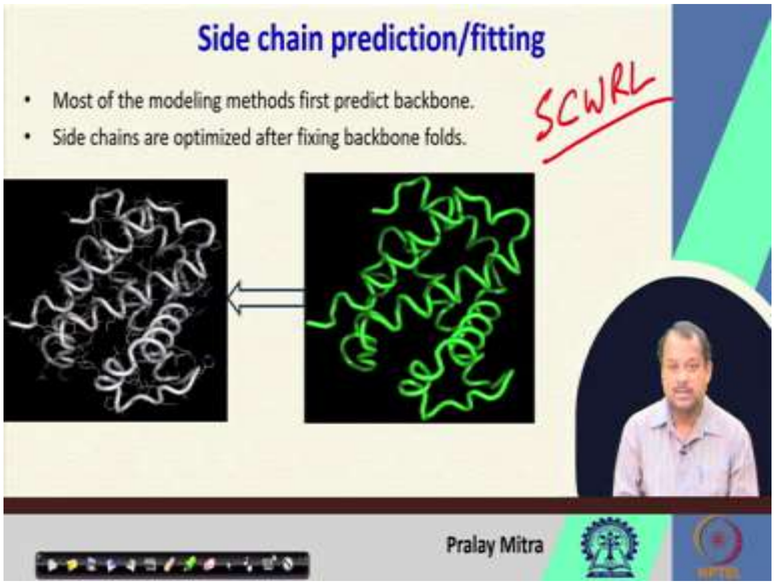
Identification of protein fold

```
>1A00:A|PDBID|CHAIN|SEQUENCE
VLSPADKTNVKAAWGKVGAHAGEYGAEALERM
FLSFPTTKTYFPHFDLSHGSAQVKGHGKKVAD
ALTNAVAHVDDMPNALSALSDLHAHKLRVDPV
NFKLLSHCLLVTLAAHLPAEFTPAVHASLDKF
LASVSTVLTSKYR
```

Pralay Mitra

So, given a protein sequence first I am proposing you identify the fold. So, when it is fold, you consider this green one is just the main chain. Side chain information is not here.

(Refer Slide Time: 06:36)



Side chain prediction/fitting

- Most of the modeling methods first predict backbone.
- Side chains are optimized after fixing backbone folds.

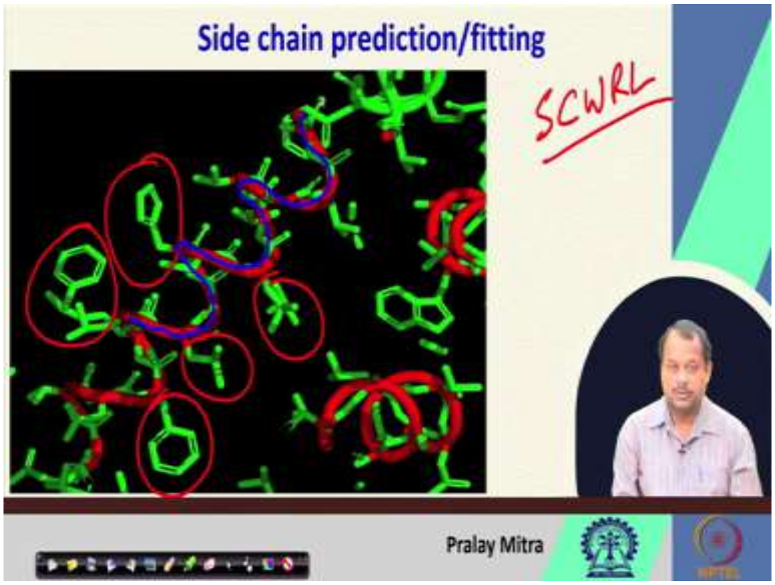
SCWRL

Pralay Mitra

Next, your job will be add the side chain. So, side chain fitting or side chain prediction is another set of algorithms that we did not discuss. So, most widely used one is by Dunbrack Laboratory that is called as a SCWRL. So, side chain fitting algorithm. So, which will take one protein sequence as an input and it can fit the side chain. Similar to the SCWRL there are several other software also which can do the same thing. So, most of the modeling method first predict the backbone. So, side chains are optimized after fixing backbone folds.

So, how, like this. So, here are the lines, the thin one is basically the side chain in this case and the thick one is same as the main chain that is in the gray color on the left hand side and green color on the right hand side.

(Refer Slide Time: 07:37)



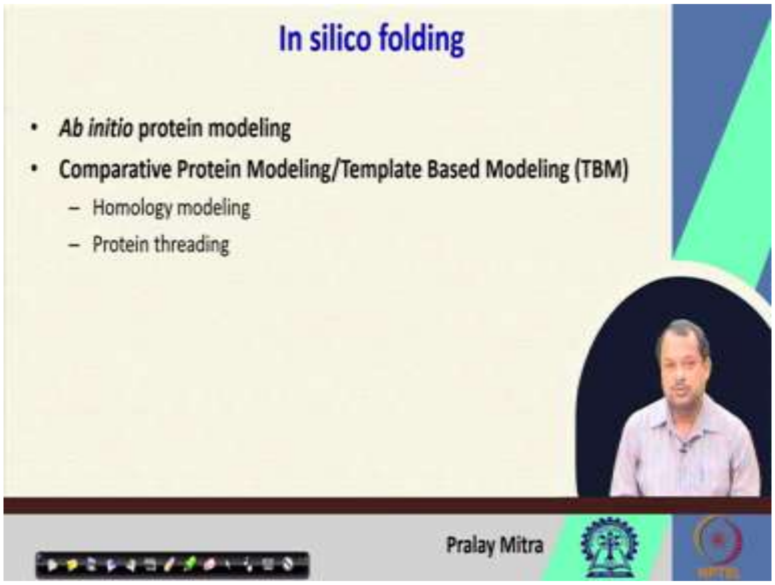
Side chain prediction/fitting

SCWRL

Now, if I look closer to this one, then I will see that the closer view indicates that here the red lines is following the main chain and only the green one is the side chain. So, I can show you some of the side chains. So, here is one side chain, here is another side chain, here is another side chain, so this is another side chain, this is another side chain and this red line actually if I say take which color blue, so blue on red, if you follow my this blue, then he will see this is my main chain.

So, on side of that, what is there is basically my side chain that I am fitting after identifying the fold, so hierarchical way. And if I do that way, then you will see that it will be faster and also it will not be computing or evaluating the sophisticated energy function during the optimization process. So, this is that one of the common practices people used to do when they are dealing with this protein folding problem.

(Refer Slide Time: 08:43)



In silico folding

- *Ab initio* protein modeling
- Comparative Protein Modeling/Template Based Modeling (TBM)
  - Homology modeling
  - Protein threading

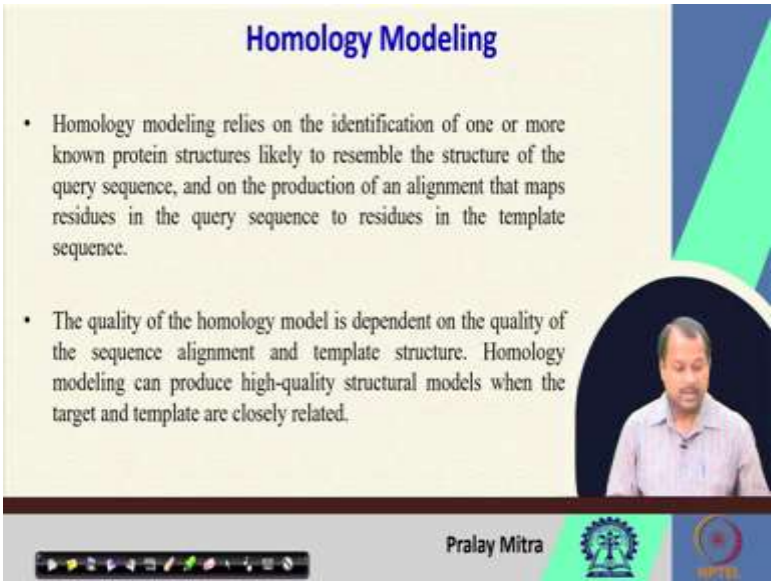
Now, if I talk about the in silico protein folding, then it is grossly divided into few parts. So, one is the ab initio protein modeling, another comparative protein modeling, third one is the template based modeling, which is also called as a TBM, template based modeling. Now, under this comparative protein modeling, it is the homology modeling and template based modeling is the protein threading. So, we did not cover this homology modeling and protein threading, because it has designed long back and now a new paradigm has come, but for the completion so today I will briefly or cursorily go through that one. But ab initio protein modeling, say we discussed in detail.

So, basic difference is that, so the comparative protein modeling relies mostly on the homologous protein sequences. So, you will understand that for a given protein sequence, if I can able to find a lot of homologous protein sequences in the protein databank, which means the structures of those homologous protein sequences are also known because I find it in the protein databank, then use that information as the model or template in order to model the structure or structure of the input sequence, that is the basic idea behind the homology modeling.

Now, in case of ab initio protein modeling, I mentioned that there is no homologous sequences. So, if the homologous sequences are not present then you have no other option, but you have to build it from the scratch. So, what are the possible say orientations at the residue level, at the, then segment level, at the topology level, so based upon that one you build that one and say Monte Carlo simulation technique can be useful for you.

And the template based modeling lies between these two, so ab initio modeling and homology modeling. So, homology modeling is when I have a lot of homologous sequences and homology is more than 70 percent or so, then when say, there is very few or say sequence similarity with the existing structure is say less than 15 percent or 10 percent then you have to go for ab initio. In between this say so when it is say between 20 to 25 to say 50 or 60 then you can go for this protein threading process.

(Refer Slide Time: 11:15)



Homology Modeling

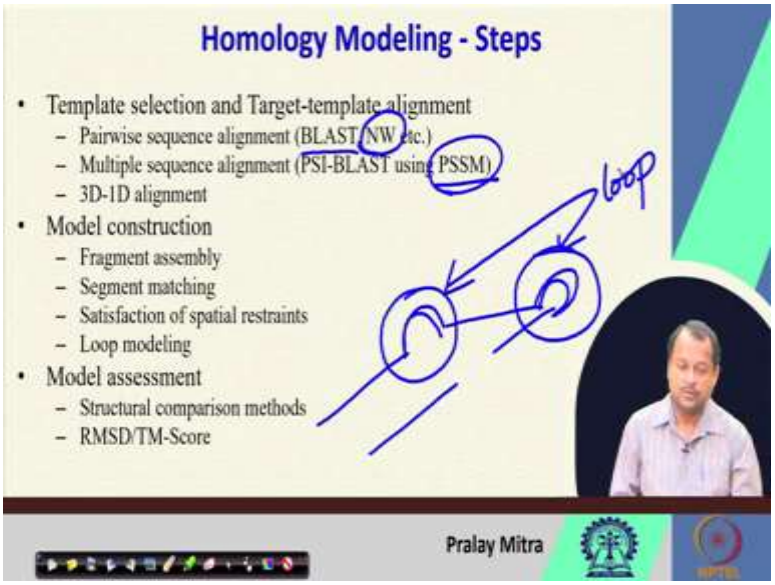
- Homology modeling relies on the identification of one or more known protein structures likely to resemble the structure of the query sequence, and on the production of an alignment that maps residues in the query sequence to residues in the template sequence.
- The quality of the homology model is dependent on the quality of the sequence alignment and template structure. Homology modeling can produce high-quality structural models when the target and template are closely related.

Pralay Mitra

So, let us start with homology modeling. Homology modeling relies on the identification of one or more known protein structures likely to resemble the structure of the query sequence and on the production of an alignment that maps residues in the query sequence to residues in the template sequence. So, the template actually is the structure that I get from the protein databank, whereas the query sequence is the input sequence for which I need to decide what will be the structure.

The quality of the homology model is depend on the quality of the sequence alignment and template structure definitely, absolutely. Given an input protein sequence, if you get very good alignment and corresponding structures are also good quality then your modeling will be very good. But if the sequence similarity is less or your alignment algorithm is not good enough to pick the correct homologous sequences, then definitely your problem is going to fail. So, homology modeling can produce high quality structural models when the target and template are closely related, means highly homologous in nature.

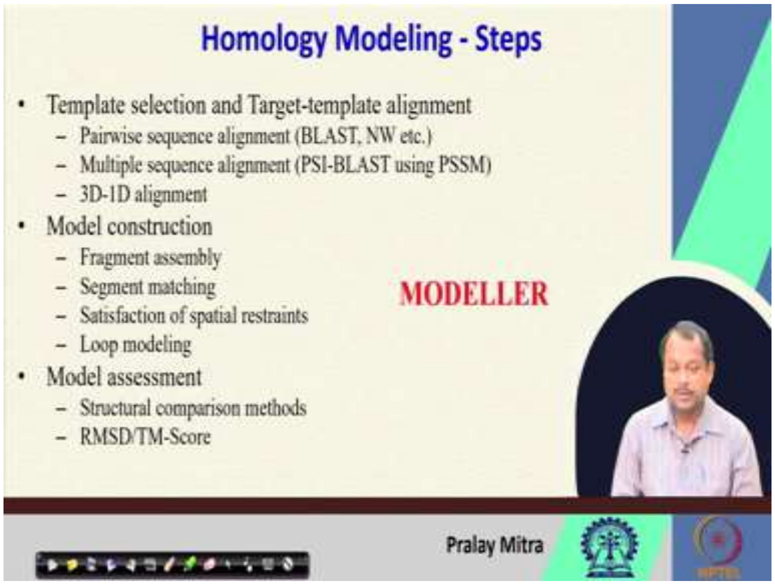
(Refer Slide Time: 12:23)



Homology Modeling - Steps

- Template selection and Target-template alignment
  - Pairwise sequence alignment (BLAST, NW etc.)
  - Multiple sequence alignment (PSI-BLAST using PSSM)
  - 3D-1D alignment
- Model construction
  - Fragment assembly
  - Segment matching
  - Satisfaction of spatial restraints
  - Loop modeling
- Model assessment
  - Structural comparison methods
  - RMSD/TM-Score

Pralay Mitra



Homology Modeling - Steps

- Template selection and Target-template alignment
  - Pairwise sequence alignment (BLAST, NW etc.)
  - Multiple sequence alignment (PSI-BLAST using PSSM)
  - 3D-1D alignment
- Model construction
  - Fragment assembly
  - Segment matching
  - Satisfaction of spatial restraints
  - Loop modeling
- Model assessment
  - Structural comparison methods
  - RMSD/TM-Score

MODELLER

Pralay Mitra

So, the steps in homology modeling is that so template selection and target template alignment you have to do. So, for this mostly Needleman–Wunsch or BLAST algorithm or say in pairwise sequence alignment or for multiple sequence alignment PSI-BLAST using that position specific scoring matrix, here you see that PSSM, position specific scoring matrix that can be done. So, this Needleman–Wunsch or sequence alignment using the dynamic programming that we discussed. This PSSM also we have discussed. So, the BLAST is one heuristic method for the sequence alignment.

And 3D to 1D alignment, you know the target template is, the structure exists for the target template, whereas for the sequence, the query sequence or the input sequence the structure does not exist, so when the structure exists, so that is a three dimensional information. When the sequence is input then that is 1-D. So, if you can have a proper alignment or mapping from that 3D to 1D, accordingly if you model then definitely you will get a better solution.

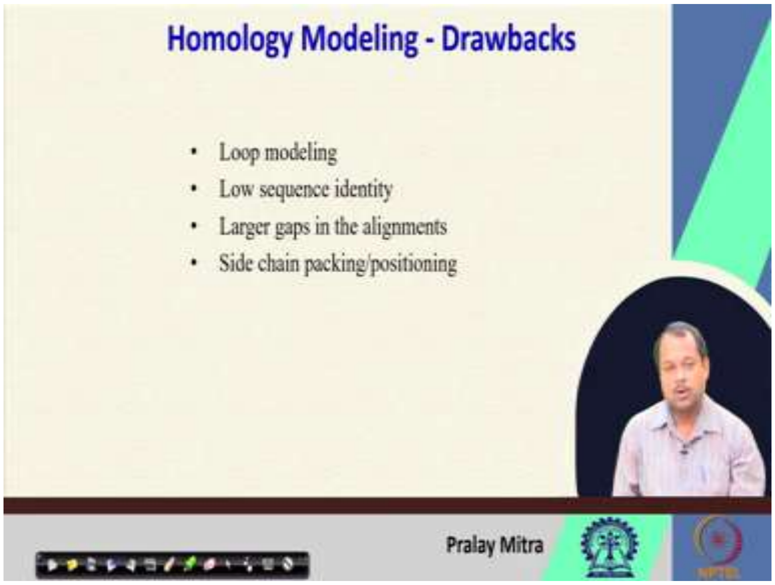
Then model construction. So, when you are performing this one then you have to keep in your mind that everything you will not get exact, so which means that given a query or input sequence, you will not get 100 percent identical sequence which you can readily copy and paste. That is not possible. So, if it is not, then definitely, so you will get some part which is matching, some part is not matching. So, take the matching part. So, that way you will have a lot of such matching parts. Then those matching parts, I will call as a fragment. If I have such, such small, small, small, small fragments then one job in homology modeling is definitely fragment assembly.

Then after assembling those fragments, you have to match the segments, so proper, for the proper assembly. Once the, it satisfies some special restraints say, for example, phi-psi values and then when I am matching, so the matching is proper based upon that one I will get the model. But when I will match most of the time during the fragment assembly, so this will be assembled, this will be assembled, this will be, this will be and some small, small regions may not be assembled, so this region say or this region, if it is then this particular region as you can see also these two are called as the loop region.

So, after say fragment assembly, segment matching and satisfying the spatial constant you got most of the thing but few small things may be missing. So, that missing thing you need to connect. When you need to connect, mostly it will be connected by some loop. And that loop you have no model. Most of the time that more loop modeling will be kind of say ab initio protein folding problem. But since that loop is a very small length nature so the computation time required for that modeling will not be that much.

So, the model assessment, structural comparison method, so after you got the structure then you need to go for some structural comparison method and then say using RMSD or TM-score you can check whether the model you have generated is good or not. Definitely for a given input sequence if you do not know what is the structure then you cannot go for this kind of model assessment, but if you know then definitely you can go. So, one of the most popularly used homology modeling technique is called as the modeler. It is from the Andrej Sali's lab from UCSF.

(Refer Slide Time: 16:14)



Homology Modeling - Drawbacks

- Loop modeling
- Low sequence identity
- Larger gaps in the alignments
- Side chain packing/positioning

However, the there are lot of drawbacks in the homology modeling. Although if you got the structure, perhaps this is the fastest one and most accurate one also, but the loop modeling, low sequence identity, larger gaps in the alignment, side chain packing or positioning those are the drawbacks. So, loop modeling is definitely you have to go for the ab initio. And now based upon the length, how long you have to connect, so that modeling may be very difficult.

Sometimes it has been noted that say if within 10 amino acids or 10 consecutive amino acids the modeling is easy, but if you go beyond 10 say 12, 15, 20 etc, then that might be a problem for homology modeling. Also, the sequence identity if say falls below 70 or so and you are not getting sufficient number of fragments which can be assembled in order to get the total structure then that is also going to be a problem. And if you get even that fragments, but during that sequence alignment, you are finding that there is a larger gap, then that can also create the problem. So, these are the problems you should be careful about.

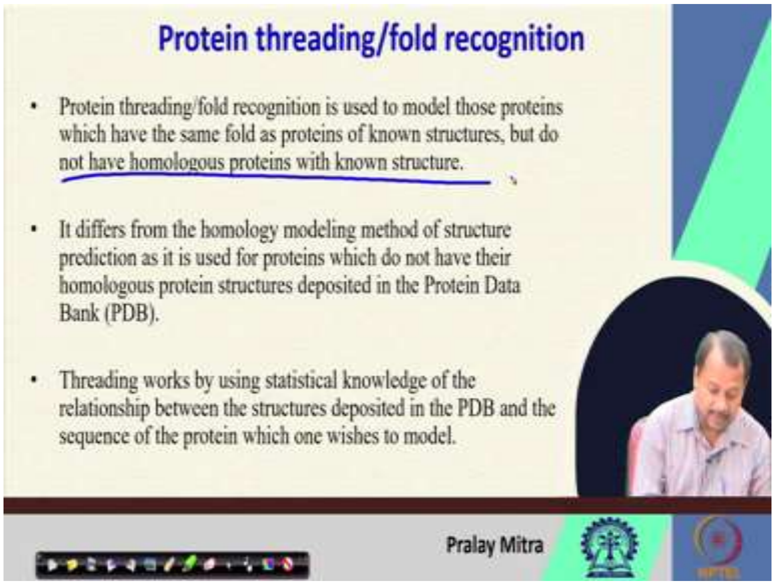
Now, no need to mention that side chain packing or positioning so which will come up after identifying the structure is also a problem. So, as I mentioned most of the time, so people consider that just focus on the fold or the main chain considering side chain is there but do not put much emphasis on the side chain during the first cut or when you are looking for the protein fold.

But while you are doing that one and specifically for the homology modeling, you are picking the fragments separately and the fragment means the original structure you are copying from some position and then you are matching based upon the spatial constant and then fixing that one. So, when you are doing all of this, then that particular fragment, what are the side chains are there was optimized for that particular structure.

Now, you are directly copying that side chain information for your own purpose. Definitely when your total organization or total structure modeling is done then you may see that some side chains are clashing, which means one is penetrating within the another, which is not allowed. So, you have to refine those side chains or you have to reposition or you have to remodel those side chains. So, there are several algorithms. So, even in fire doc in the context of protein docking this kind of problem may arise.

When say we are considering two separately crystallized protein molecules which are coming for the interaction then definitely their side chain packing will not be good. So, what we have to do that we have to basically reposition that one or refine those side chains. So, that kind of problem exists in the homology modeling also. Nevertheless, the last I should not say drawback or problem of the homology modeling, mostly it is the problem or almost all the modeling. You have to take care about that.

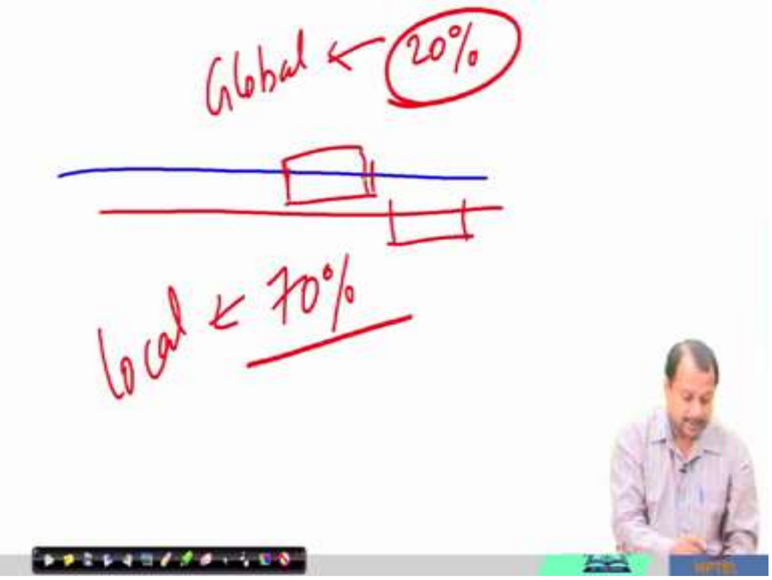
(Refer Slide Time: 19:31)



Protein threading/fold recognition

- Protein threading/fold recognition is used to model those proteins which have the same fold as proteins of known structures, but do not have homologous proteins with known structure.
- It differs from the homology modeling method of structure prediction as it is used for proteins which do not have their homologous protein structures deposited in the Protein Data Bank (PDB).
- Threading works by using statistical knowledge of the relationship between the structures deposited in the PDB and the sequence of the protein which one wishes to model.

Pralay Mitra



The next point we are coming which is called as a protein threading or fold recognition which is the most widely used. Because, so it is not very easy to get a very high sequence similarity and also considering the fact that almost all the folds are available with us, then we can expect that some sort of similarity we will get and based upon that partial similarity we need to build a total complex, we need to build a total protein or model total protein.

So, protein threading or fold recognition is used to model those proteins which have the same fold as proteins of known structures, but do not have homologous proteins with known structure, because if I have homologous proteins with known structure then it is going to be the part of homology modeling. So, protein threading is not required for this. Now, it differs from the homology modeling method of structure prediction as it is used for proteins which do not have their homologous protein structures deposited in the protein databank, that I mentioned.

So, if the homologous structures are not there, so the percentage of similarity is comparatively less then we should go for protein threading or folding. Now, when it is less then you remember in the context of sequence alignment also, specifically when we discussed that Needleman–Wunsch and Smith–Waterman, global alignment and the local alignment then I mentioned that it may be possible that at the global alignment level two sequence similarities are very less.

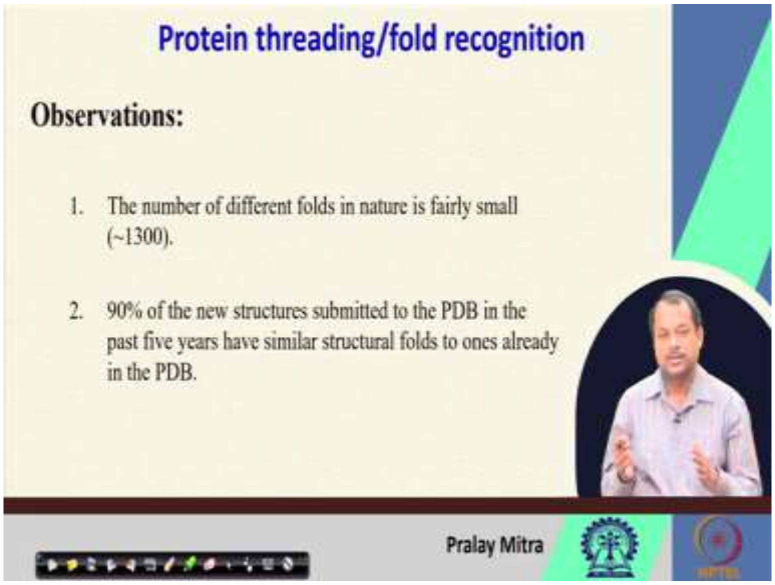
Say, for example, this is one sequence given to you and say this is another sequence given to you, now if I align at the global level I can see that sequence similarity is say 20 percent. Now, if I look for some local region it may possible that that is 70 percent. Now, you remember that when I am computing the similarity or the identity then that total match will be normalized by the length of the protein. Now, because of the normalization say only a small amount has been matched and since I am normalizing by the total length of the protein that is why the global percentage is reducing.

On the other hand, if I just focus on the local part and normalization is also done at the local part, then it may be that or it must be I should say that the percentage of the match will increase. Now, when the homology modeling insist that corresponding to the query sequence or the input sequence you provide me the homologous sequences whose structure is known, but when it is not possible, then protein threading compromises on that, and it says, okay, at the global level is a sequence identity is 20 percent is it possible that at the local level or some local region is matching.

If the local region is matching say this match is matching only with the red of this match, so this match is matching, then give me that match part from this blue. So, from blue give me the matched part, rest of the part I am not interested in. Now, if you think like that way, then what you will get, so from blue we will get one matched part, so from green, from red, from black, so locally you are getting small, small match part.

Next your job will be to thread those match part in order to get the complete assembly or the complete model. Since I need to thread those small, small match parts that is why the name also suggest that it is protein threading or fold recognition problem. So, threading works by using statistical knowledge of the relationship between the structures deposited in the PDB and the sequence of the protein which one uses to model.

(Refer Slide Time: 23:52)



Protein threading/fold recognition

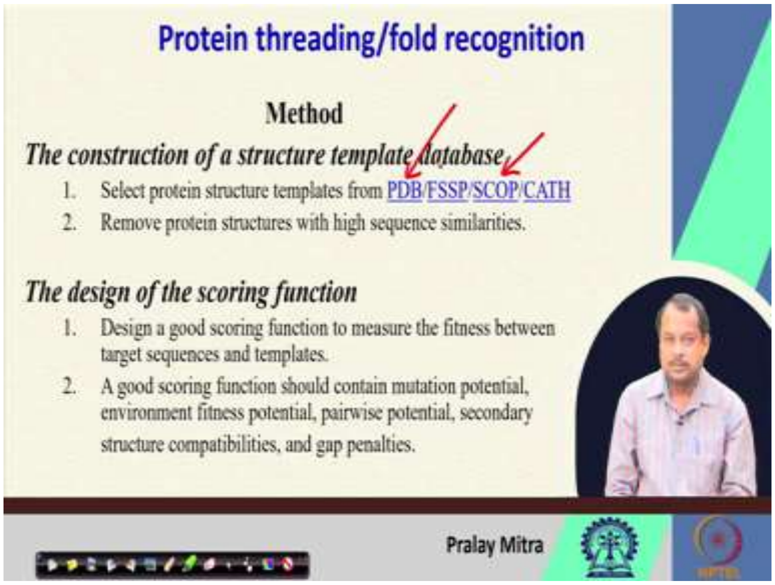
Observations:

1. The number of different folds in nature is fairly small (~1300).
2. 90% of the new structures submitted to the PDB in the past five years have similar structural folds to ones already in the PDB.

So, observation is the number of different folds in nature is fairly small about 1300. 90 percent of the new structures submitted to the PDB in the past five years have similar structural folds to once already in the PDB. Now, if you have these two observations, then you can say that probably the query sequence will not have a new fold, probably I said. But the probability of having a new fold is very small.

Another, if no new structures has been submitted, so some similarity I will get. And if the similarity is not good enough for homology modeling, definitely it is going to be good enough for my protein threading for the local alignment. So, based upon these two observations actually the protein threading or fold recognition algorithms are developed.

(Refer Slide Time: 24:56)



Protein threading/fold recognition

Method

The construction of a structure template database

1. Select protein structure templates from PDB/FSSP/SCOP/CATH
2. Remove protein structures with high sequence similarities.

The design of the scoring function

1. Design a good scoring function to measure the fitness between target sequences and templates.
2. A good scoring function should contain mutation potential, environment fitness potential, pairwise potential, secondary structure compatibilities, and gap penalties.

So, in the method, so first you need to construct a structural template database. So, how you have to do that one? Step one, select protein structure templates from protein databank FSSP, SCOP, CATH etc. Now, SCOP we discussed in detail, PDB we discussed in detail, although the protein structures which are deposited in the PDB is also available in the SCOP, but we discussed that SCOP is trying to classify the proteins based upon their function.

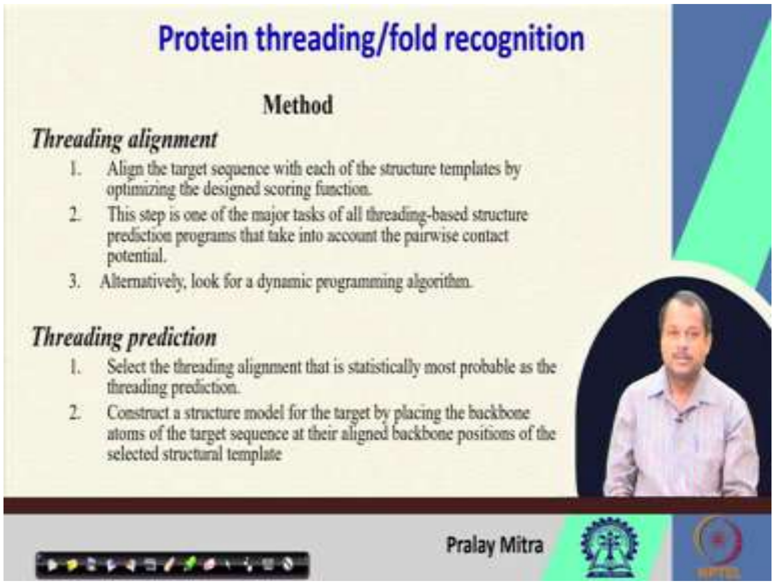
So, if the semi-automated technique that SCOP uses to classify your proteins is also been adopted by you to identify the similar proteins based upon only the sequence not the structure, then that similar structural information, so based upon the sequence only for the new or query sequence, I am following the same semi-automated technique as SCOP following to reach to its new class, sorry, not new class its class.

So, when I will go to that particular say cluster, then in that cluster whatever the proteins are present, now I will find that they are with the same function, which means they are with the same structure. So, you got some information. You use that information as your structural template. Remove protein structures with high sequence similarities.

So, this high sequence similarity although you may feel that, okay, that gives me a very good information that how many proteins are with the same sequence similarity, but when say I am looking for the statistical information so that may be redundant or that may bias me. So, remove all those redundancies. I need only the new information and some overlapping information so that I can keep on threading.

Next, you need to go for a design of a very good scoring function. The design of a good scoring function to measure the fitness between target sequence and template is very much required. Because a good scoring functions should contain mutation potential, environmental fitness potential, pairwise potential, secondary structure compatibility and gap penalties. So, incorporating all those things you need to come up with a very good scoring function, because scoring function as you have noted is integrated part of any algorithm that we are discussing in this protein modeling and engineering.

(Refer Slide Time: 27:55)



Protein threading/fold recognition

Method

Threading alignment

1. Align the target sequence with each of the structure templates by optimizing the designed scoring function.
2. This step is one of the major tasks of all threading-based structure prediction programs that take into account the pairwise contact potential.
3. Alternatively, look for a dynamic programming algorithm.

Threading prediction

1. Select the threading alignment that is statistically most probable as the threading prediction.
2. Construct a structure model for the target by placing the backbone atoms of the target sequence at their aligned backbone positions of the selected structural template

Next, you have to go for threading alignment. So, align the target sequence with each of the structure templates by optimizing the design scoring function. This step is one of the major tasks of all threading based structure prediction programs that take into account the pairwise contact potential. Alternatively look for a dynamic programming algorithm. So, basically in this case, it suggests that dynamic programming algorithm can also be useful for threading alignment.

Now, in threading prediction select the threading alignment that is statistically most probable as the threading prediction. Construct a structure model for the target by placing the backbone atoms of the target sequence at their aligned backbone positions of the selected structural template. So, those are the steps. And based upon those steps, actually, the protein threading or for recognition techniques has been developed.

Definitely each steps demands a further step-wise implementation of the technique, but grossly the methods are developed or build upon this one. So, we will continue this to the next lecture. But thank you now for the time being. Thank you.