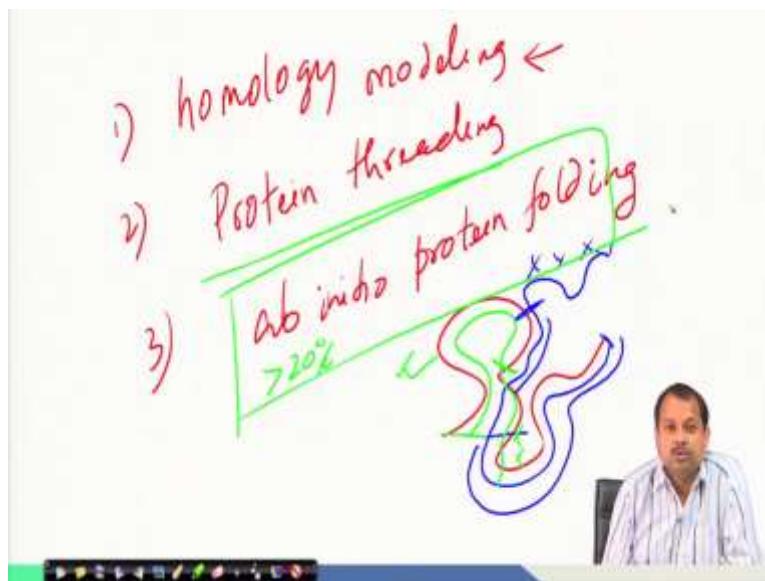


**Algorithms for Protein Modelling and Engineering**  
**Professor Pralay Mitra**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology, Kharagpur**  
**Lecture: 24**  
**Ab Initio Protein Folding**

Welcome back. So, in this lecture, we will see the application of the replica exchange Monte Carlo technique that we discussed on the last lecture for protein folding and in protein folding there are say three different kinds of techniques for protein folding.

So, one is called as the homology modeling, another is called as the protein threading and third one is called as the Ab initio protein folding. So, when we will discuss a computational framework for all those then in detail I will mentioned but for at least today's discussion. So, these three techniques, I would like to mention little.

(Refer Slide Time: 01:02)



So, homology modeling that I mentioned, then protein threading and Ab initio. So, in summary, what this three technique used to do in case of homology modeling. So, it looks for homologous protein sequences. So, given one protein sequence as an input it will search say protein databank.

That protein databank consists of a number of protein structures and when structures are their protein sequences are also there. So, it search for the existence of the similar protein sequence in

the protein databank if it finds that heat, then it takes that particular structure as a reference and based upon that one it model.

So, that way, homology modeling is kind of easy technique for protein structure prediction And two protein sequences. So, when say it is greater than 70 percent sequence similarity, then we call that as the homologous sequences and based upon that one it take place. In case of protein threading the situation is not like 70 percent sequence similarity.

But sequence similarity is something in between say 25 to 60 or 70 percent in that case, when it is percentage, then definitely you understand the similarity divided by the total length of the protein is considered as the overall protein. Similarity now, when say I am talking about the protein threading then overall the similarity maybe say 50 percent.

But, if I consider one region or one part, then the similarity may be very high. So, if I give you one example, say this is one structure let us assume and blue color. So, let us assume this is another structure. Now, if I consider say what is the similarity assuming that structure is different when sequence is also different with that assumption?

If I assume then you can see that at the global level if I see the sequence similarity again I am assuming the sequence and structure is one to one correspondence because of that one because schematic on the structure will be easy compared to the sequence that is why I am assuming this one.

Now, you see that the similarity say if I consider for two proteins complete proteins then it will be less but if I consider that similarity up to this, it will very high. If I consider say up to this then it is almost 100 percent. Now, protein threading uses that concept what it does that when it finds that does not exist to one single protein structure.

Whose with whom my sequence identity is a more than 70 percent or both the proteins are same in nature Then it looks for all the protein structures who are say in between say 25 to say 60 or 70 percent. Because I know that nobody is there, who are more than 70 percent and when it is overall say 25 to 60 or 70 percent then it tries to identify is there a small regions.

Where the sequence entity is more, if yes, then it chops out this region, this region but do not consider this region for blue protein, if there is another protein say green. So, for which

similarity is something like this, then it finds that with the green. This region is fine, but not this region.

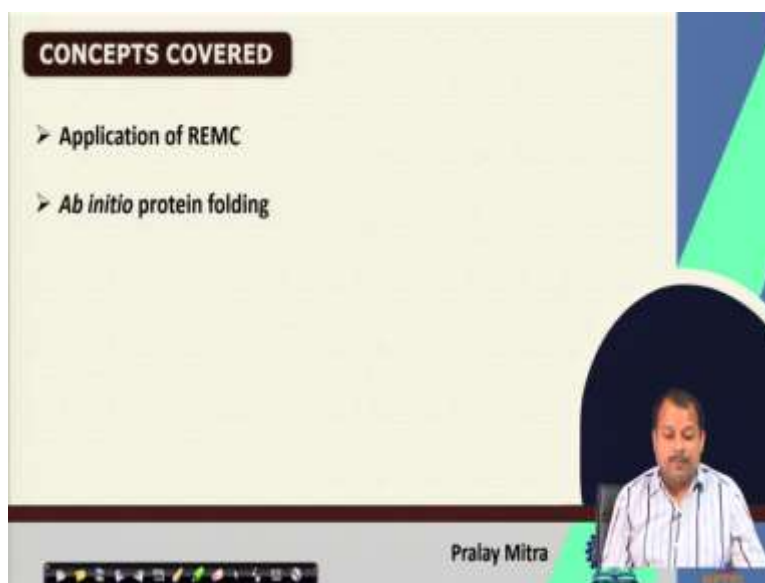
That way piecewise information it gathers and then combines that piecewise information in order to get the protein threading the difficult hard or toughest one is the Ab initio protein folding when the sequence searches that does not exist any protein sequence whose structure exists and we throw the given input sequence is with the sequence similarity greater than 20 percent also mean you have no other option, but you have to build from the scratch.

When you are building from the scratch, then what you are doing, you remember the protein representation we talked about. So, you generate all the protein representations corresponding to the protein sequence and you store in some database. So, we will present residue level we will present atomic level we will present segment level we will present topology level we will present global moment also and we will place in some database.

Now, what we will do when I will say in my replica exchange Monte Carlo will call for one conformation then once the residue level starting from the topology level then segment level then residue level then atomic level conformation will come assume the existence of one energy function.

That energy function will evaluate that energy and that conformation energy will come then through the replica exchange Monte Carlo we will decide whether we will return that or not, I mean we accept that or not that decision we will take at that position. So, we are going to discuss today's class this ab initio protein folding.

(Refer Slide Time: 07:12)



**CONCEPTS COVERED**

- Application of REMC
- *Ab initio* protein folding

Pralay Mitra

The slide features a dark blue header with the title 'CONCEPTS COVERED' in white. Below the title, two bullet points are listed. A video inset in the bottom right corner shows a man with a beard and glasses, wearing a striped shirt, speaking. A control bar with various icons is visible at the bottom of the slide.



**KEYWORDS**

- REMC
- *Ab initio*
- Protein folding

Pralay Mitra

The slide features a dark blue header with the title 'KEYWORDS' in white. Below the title, three bullet points are listed. A video inset in the bottom right corner shows the same man from the previous slide. A control bar with various icons is visible at the bottom of the slide.

So, the concept, we will cover the application of REMC and ab initio protein folding and keywords are REMC, ab initio and protein folding.

(Refer Slide Time: 7:23)

**Ab initio protein folding**

**Algorithm 8:**

**Input:** Protein sequence  
**Output:** Full atomic model

**Steps:**

1. Preprocessing steps
  - a. Create sequence profile
  - b. Create distance profile
  - c. Predict sequence based secondary structure, solvent accessibility, and torsional angles.

Handwritten annotations: Two circles labeled A and B with radii  $r_1$  and  $r_2$ . A bracket indicates  $\text{dist}(A, B) < r_1 + r_2$ .

Pralay Mitra

So, in ab initio protein folding input is just a protein sequence and output is full atomic model and when it is a full atomic model. So, it must follow the Ramachandran map or Ramachandran plot. Now, the steps are grossly divided into three parts, first one is the pre processing step. So, in this step what we are going to do that in protein sequence I got and I also mentioned that since it is ab initio protein folding.

So, there is no way that we are getting the protein structures whose sequence is similar with our input protein sequence with more than 25 percent it is not if it is not then also at the beginning we will create some sequence profile. So, out of how many protein sequences there are some similarities even if it is small then exist those small small sequence similarity information I am taking and I am creating one sequence profile based upon that one.

I am creating one sequence profile based upon that. Next I am creating the distance profile. So, regarding this distance you should understand that when we are writing the program, then theoretically two atoms with say radius  $r_1$  and  $r_2$  may penetrate which means during our modeling it may be possible that  $r_1 + r_2$  is less than  $\text{dist}(A, B)$  if I say A and B.

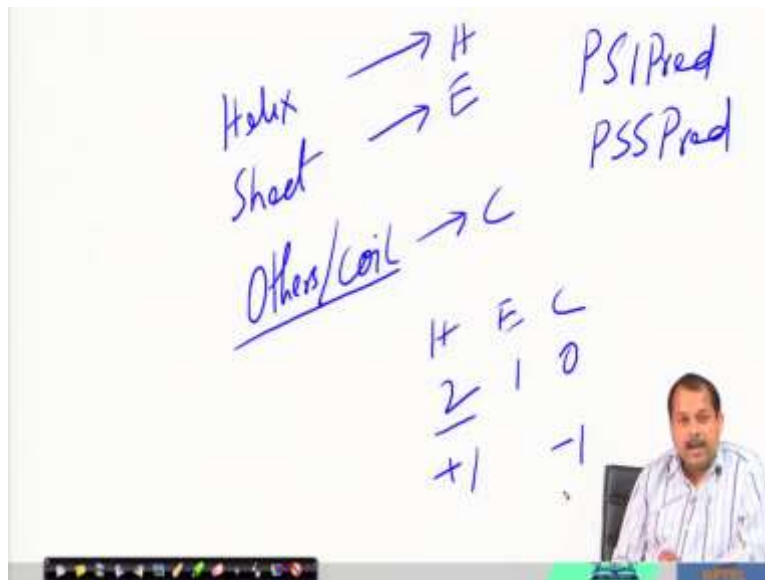
So, which means it will penetrate sorry this will be this distance between A and B is less compared to  $r_1 + r_2$ . Theoretically these two cannot penetrate with each other. So, this kind of

distance profile also we have to create. Next we have to predict sequence based secondary structure solvent accessibility and torsional angles.

So, these are separate algorithm. We will assume the existence of that one. Mostly the neural network based techniques are there which will actually take one protein sequence and as an input and predict that what will be the, its secondary structure solvent accessibility and torsional angles.

If we have that information based upon the protein sequence, then what we will do that during our energy function, we will incorporate that information and if it is within that range, then we will give some score if it is not within that range then we will penalize that one, the simplest one I can tell you, let us assume this secondary structure.

(Refer Slide Time: 10:40)



## Ab initio protein folding

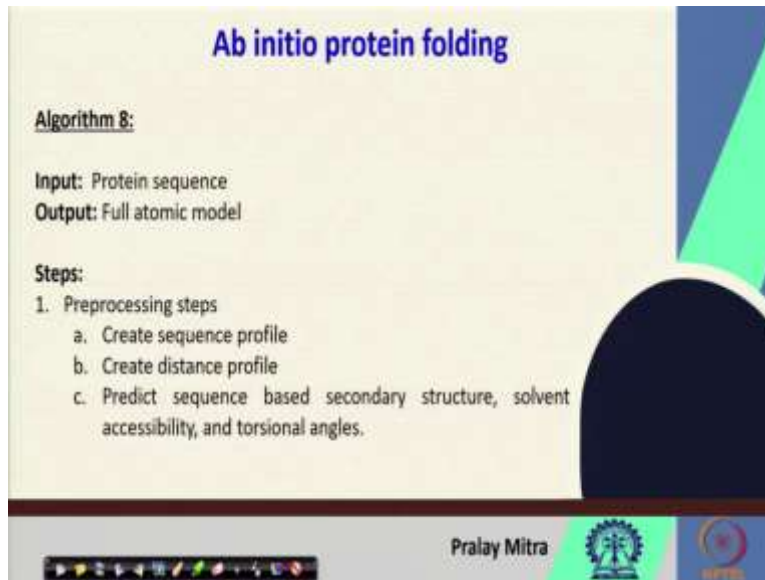
**Algorithm 8:**

**Input:** Protein sequence  
**Output:** Full atomic model

**Steps:**

1. Preprocessing steps
  - a. Create sequence profile
  - b. Create distance profile
  - c. Predict sequence based secondary structure, solvent accessibility, and torsional angles.

Pralay Mitra



So, for the secondary structure there are three secondary structure helix, sheet and others or coil you can consider. So, for this I am calling this. Now, please note there are several variations of helix sheet or several variations of say this irregular structure which is not part of helix or sheet we are not going into that detail.

So, grossly we are classifying in the three class H E and C helix sheet and coil. Now, if we predict from the sequence using the neural network technique. So, there are some techniques called as the Psi Pred or PSS Pred, which can take one protein sequence as an input and predicts the secondary structure.

Now, if we predicts say H E C in terms of HEC, then this can easily I can translate to say 0 1 2, 3 different states. Now, during our random selection of one conformation during the score function of that random conformation, if I find that this will make a helix and I am predicting also as a helix then I will give on penalties say plus 1 if I see that.

That conformation that I have picked tells that it is going to be a coil whereas my prediction says that it is less then I will get some negative that way if I have some empirical energy function then I can use that. So, the same thing is also used in case of this ab initio protein folding for that I need to do a pre processing step.

(Refer Slide Time: 12:40)

**Ab initio protein folding**

2. Perform REMC for  $N$  number of replicas
  - a. Decide on the number of replicas
  - b. Decide on the maximum and minimum temperature, and the offset temperature between different replicas.
  - c. Design an energy function for REMC simulation
3. Post processing steps
  - a. Cluster the accepted conformations.
  - b. Output the models based on their clustering rank.

Handwritten notes:  $M < Num$  (twice)

Pralay Mitra

Next comes the actual the core of the simulation that is the replica exchange Monte Carlo. So, perform REMC for  $N$  number of replicas. So, here in I am mentioning variable because I am going to tell this number of steps in a generic way. Now, regarding this  $N$ , energy function, et cetra et cetra, you can design your own or you can plug from others and then you can have your own customized algorithm. Now decide on the number of replicas decide on the maximum and minimum temperature and the offset temperature between different replicas.

Design and energy function for REMC simulation, those three things you have to design. Now, this core part will give you a number of different conformation. Because if there are  $N$  number of replicas in each replica, there was some conformations switched was accepted. So, list of such conformation for first replicas second, third, fourth, fifth up to  $N$  is there you are taking the union of that one finally.

You are getting one list which is the correct one you are having with a number of conformations. So, which is the correct one? So, in order to check so, you have to have some post processing steps. So, there can be two different post processing steps even one is kind of integrated with your REMC another after REMC processes over.



So, in your REMC if you incorporate that my energy function is one of the best and I wish to completely rely on my energy function. So, whoever was the lowest energy you please output that one fine, simplest way you can go ahead with that. But it is always suggested that instead of one conformation you please output in number of conformation say 5, 10, 20, something you out of them.

So, that way also what you can do that you can say rank them based upon your score function you have to after accepting those steps and when REMC is over. You are taking the union of all the accepted conformations from different replicas after accepting those corresponding to each conformation there is some energy function based upon that energy function you showed them and you output the conformation and their energy function in the sorted list and you place them as per they are ranked that is one situation.

Another situation is that during the process, I mean during REMC how many times you are encountering one situation if you look at the protein energy landscape, then you will understand that it is more unlikely that the top side of the funnel which means that which are close to unfolded state in that region, more and more conformations will be accepted.

If it is accepted, it is accepted either because its energy is less or it is accepted because of the fact that you are applying metro list criteria, but whatever maybe the reason the chances are high that lowest or lower energy conformation will be accepted that way, it might be a good idea that from the union of all the conformations that you have accepted you go for one clustering first.

The clustering will tell that which are very much related conformations, how I can understand the relatedness of that one that I will tell you later, but if you assume that if somehow I can understand that these two conformations are related, then you put them together and that way you can reduce the number of accepted conformations.

What I am trying to say that if you have two conformation and so, one is like this, another is like this. Now, you see these two conformation are very much same. So, instead of having both you can have only one or you can have only one say with the minimum energy that way you can reduce from initial union of say in Num you can reduce to M where M is less than Num.

And when you are doing that one and when you are doing that one then you are clustering them and after clustering them one approach could be that you cluster So, each cluster has some cardinality. Now, your assumption might be that the cluster whose cardinality is higher is possibly having the more and more conformations with the minimum energy and I also agree with you. So, in the clustering step, what do you need to do in the clustering step?

So, first you have to cluster based upon those structures, how to cluster that we will discuss later please hold out. But assume that one such technique is there. After clustering, first of all number will reduce, if not then also you will have all the number of clusters there instead of individual conformations, you have to rank the clusters and when you will rank the clusters I am telling not by energy you go by the cardinality of the cluster.

The advantage is that when you go by the energy then you are assuming your energy function is perfect, but when I am saying going by the ranking of the clusters where the cardinality is considered then I am assuming whatever maybe my energy function it may not be perfect it is approximately correct and within some range who are having since those are explored more then chances are high that is near to the native state. So, based upon that assumption, we will cluster and we will select based upon their rank.

(Refer Slide Time: 19:53)

**Ab initio protein folding**

- 40 replicas
- 200 cycles are run for each protein by default
- 10 different REMC simulations with different starting random numbers are initiated.
- The Lehmer random number generator is used for random number generation (256 different streams with a long period  $2.15 \times 10^9$  in each stream).
- In total, 5000 decoys randomly selected from the last 150 cycles of the 10 low-temperature replicas from the 10 simulations are gathered and clustered (using SPICKER).

QUARK (Yu and Zhang(2012) Proteins 1715-1735)

Pralay Mitra

The slide features a video inset of a man in a light blue shirt speaking. At the bottom, there are navigation icons and logos for IIT Bombay and the Department of Chemical Engineering.

Next in ab initio protein folding. So, one such method exists whose name is quad it is published in proteins 2012 they have used 40 replicas, 200 cycles are run for each protein by default local

steps. 10 different REMC simulations with different starting random numbers are initiated. The Lehmer random number generator is used for random number generation 256 different streams with a long period  $2.15 \times 10^9$  in each stream.

So, you understand that the reputation or the pattern will reappear after  $2.15 \times 10^9$  which is very high. And that was our proposal when we discussed the random number generator also. And we mentioned that in Monte Carlo simulation even for the REMC, which is an extension Monte Carlo simulation basically random number plays a crucial role.

So, it must be random as much as possible. In total, 5000 decoys randomly selected from the last 150 cycles of that 10 low temperature replicas from the 10 simulation are gathered and clustered using SPICKER that is another algorithm for clustering, but this algorithm also used to do the same thing and that thing is basically you cluster them based upon their structure and you rank them.

(Refer Slide Time: 21:31)

**Ab initio protein folding**

The Lehmer random number generator, also known as Park–Miller random number generator is a type of linear congruential generator (LCG) that operates in multiplicative group of integers modulo  $n$ . Thus, we can write:

$$X_{k+1} = a \times X_k \bmod m$$

where,  $m$  is a prime number or a power of a prime number,  $a$  is an element of high multiplicative order modulo  $m$ , and the seed  $X_0$  is coprime to  $m$ .

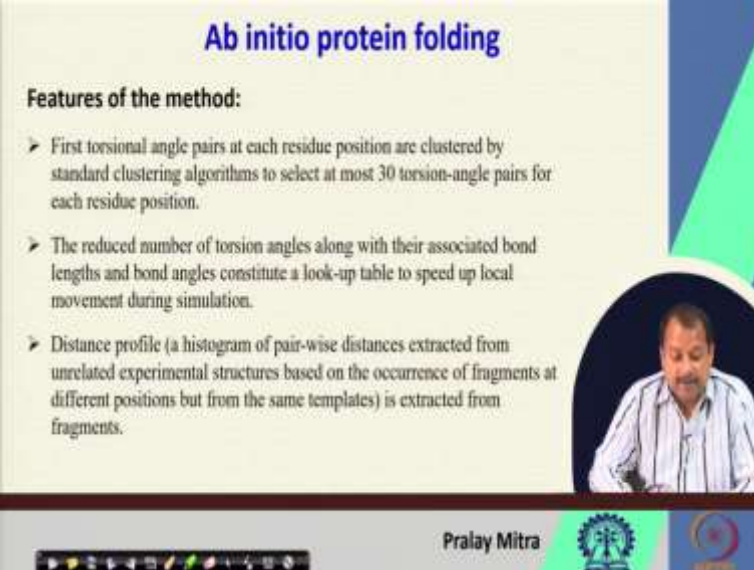
Pralay Mitra

Now, what is this Lehmer random number generator for you implementation purpose I am giving you the equation it is also called as the linear congruential generator. So, that operates in a multiplicative group of integers modulo  $n$ . Thus, we can write  $X_{k+1}$  equals to  $a$  multiplied with  $X_k \bmod m$ , where  $X_k$  is the random number in the previous stage with that one, you are multiplying  $a$  and then you are taking  $m$ .

$m$  is a prime number or a power of prime number  $a$  is an element of high multiplicative order modulo  $m$  and the seed  $X_0$  is coprime to  $m$ . So, you are starting from  $X_0$ . So, that is a coprime to  $m$  then you are generating  $X_1$  equals to  $a$  say high multiplicative order modulo  $m$  multiplied with  $X_0$  then taking mod  $m$ ,  $m$  is a prime number and that way you have to go but you should remember like the previous quad implementation.

That it should be very long in nature what it mention that 256 different streams with a long period  $2.15 \times 10^9$  in each stream that you have to maintain. So, this is the Lehmer random number generator. So, as of now, we finished the discussion of the algorithm which is customized for the quad. So, one thing is remaining that is the energy function.

(Refer Slide Time: 23:07)



**Ab initio protein folding**

**Features of the method:**

- First torsional angle pairs at each residue position are clustered by standard clustering algorithms to select at most 30 torsion-angle pairs for each residue position.
- The reduced number of torsion angles along with their associated bond lengths and bond angles constitute a look-up table to speed up local movement during simulation.
- Distance profile (a histogram of pair-wise distances extracted from unrelated experimental structures based on the occurrence of fragments at different positions but from the same templates) is extracted from fragments.

Pralay Mitra

The slide features a blue and green geometric design on the right side and a small inset video of the speaker, Pralay Mitra, in the bottom right corner. A navigation bar is visible at the bottom of the slide.

So, in energy function, we will go but before that, let me tell you some of the features of this work method. So, first torsional angle pair at each residue position are clustered by standard clustering algorithms to select at most 30 torsion angle pairs for each residue position.

The reduced number of torsion angles along with their associated bond lengths and bond angles constitute a lookup table to speed up local movement during simulation and distance profile a histogram of pairwise distances extracted from unrelated experimental structures based on the occurrence of fragments at different positions, but from the same template is extracted from fragments.



(Refer Slide Time: 23:55)

**Ab initio protein folding**

**Features of the method (contd.):**

- The predicted SA is used in the energy term. The predicted three-state SS types will guide the simulation to generate decoy structures with the similar SS types.
- If one template fragment is successfully placed into the decoy by the fragment substitution movement, this segment will have the same SS types as the fragment structure.
- The predicted probabilities of  $\beta$ -turn positions will be used to guide one movement for  $\beta$ -turn formation.

Handwritten annotations:   
← Solvent accessibility (pointing to the first bullet)   
→ Secondary Structure H/E/C (pointing to the second and third bullets)

Pralay Mitra

Next, continue the predicted SA, SA means solvent accessibility is used in the energy term. As I mentioned that if you have some profile information like distance profile, solvent accessibility, secondary structure torsional angle information, it is better to combine that with your score function.

If you combine that one with your score function, then what you can get that if it is following that profile, then you can award some score or reward something if it is not following that one then you can penalize that one. So, it is possible. So, the predicted SA is used in the energy term predicted three-state SS the secondary structure that I mentioned as HEC.

So, this is solvent accessibility. And this is your secondary structure. Helix, sheet or coil type will guide the simulation to generate decoy structures with the similar types. If one template fragment is successfully placed into that decoy by the fragment substitution movement then segment will have the same secondary structure type as the fragment structure.

Now, you remember that segment level, topology level, residue level and atomic level information we discussed that information stored in database. So, when you are picking then you please follow this template information. The predicted probabilities of beta turn positions will be used to guide one movement for  $\beta$ -turn formation.

(Refer Slide Time: 26:06)

**Ab initio protein folding**

Energy parameters

1. Backbone atomic pair-wise potential ✓
2. Side-chain center pair-wise potentials ✓
3. Excluded volume ✓
4. Hydrogen Bonding
5. Solvent Accessibility
6. Backbone torsion potential
7. Fragment-based distance profile
8. Radius of gyration
9. Strand-helix-strand packing
10. Helix packing
11. Strand packing

SC

QUARK (Yu and Zhang[2012] Proteins 1715-1735)

Pralay Mitra

**Ab initio protein folding**

Selection of Dataset

Pralay Mitra

So, next the energy parameters which is used in the quad. So, all the parameters are empirical parameters. So, it is computed based upon the known protein structures. So, that is why I am not going into detailed discussion of that energy function you can pick any energy function for your purpose, I am just focusing on the algorithm development and the customization.

But for the completeness I am mentioning that what are the components in the energy function will be so, backbone atomic pairwise potential then side chain center pair wise potentials then

excluded volume or the first three is atom level and residue level. So, backbone atomic pairwise potential.

So, following the layer potential or whether somebody is penetrating or not that part will be taken care by this backbone atomic pairwise potential side chain center pairwise potential. So, you remember that instead of complete side chain for the sake of computational improvement, computational efficiency, we represented the complete side chain by one single atom that is SC which assume the centroid of all the atoms that is SC.

So, sidechain center pairwise potential between them the pairwise potential you have to calculate the excluded volume you have to calculate. Now, at the residue level hydrogen bonding, so, if there is a donor and acceptor and also there is a capability of forming the hydrogen bond then that hydrogen bonding energy will be calculated how many number of hydrogen bonding are formed based upon that one on contribution will come solvent accessibility how much is on the surface how much is at the core.

So, that information will be considered backbone torsion potential whether it is following the profile of the torsional angle that we computed during pre-processing stage that we have to consider fragment based distance profile. So, we computed the fragment and for the known protein structures, previously, we computed the fragment which are possible and we created one profile that profile is stored.

So, we will compute that profile along with the random conformation that we picked right now, whether they are matching or not based upon that one we will consider this fragment based distance profile. Next radius of gyration at the protein level or the whole fold level, radius of gyration, strand helix strand packing, helix packing and strand packing. So, this strand also indicates the sheet. So, strand helix strand packing how packed there.

So, when there is a packing which means, between two helix two strand or one helix, one strand, there will be hydrogen bonding and again if there is a hydrogen bonding then they will contribute to the hydrogen bonding part. So, they are interrelated. So, that way these are the features is considered in quark. Now, you can consider these features you can add more features and they computed empirically, you can compute by yourself or you can use their score function or lay and you improve your algorithm that is what we are discussing now.



So, that is all about the application of the REMC that we discussed in the context of this protein folding. Next the selection of the data set Since we are working with ab initio protein folding, so, my suggestion is that you should select the proteins with which there is no you do not find any homologous protein sequence or structure in the PDB.

So, if you take that one, then you can go for ab initio protein folding otherwise during your fragment calculation or distance profile calculation, you will get a lot of help or support from the PDB and he will not able to judge the goodness of your algorithm or goodness of your energy function. So, regarding this detailed discussion on the data set, how to design the data set. We will discuss more in other lectures also. So, that is it. Thank you very much.