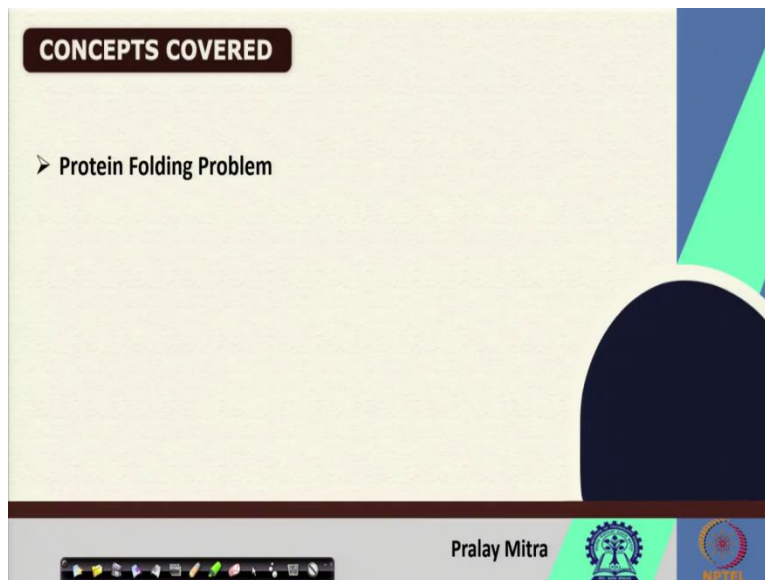


**Algorithms for Protein Modelling and Engineering**  
**Professor Pralay Mitra**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology, Kharagpur**  
**Lecture 19**  
**Protein Folding**

Welcome back. So, today I will start introducing you protein folding problem. So, that is one of the most challenging problem and in this domain and recently alpha fold did a very good job in this in solving this protein folding problem. So, those things will come later, but let us start introducing this problem with the problem definition.


And if we wish to apply the Monte Carlo technique, then how we can apply that one because some of the existing tools or software which solves protein folding problem are in using this Monte Carlo Simulation based technique or a variation of that one that is (( ))(1:01) exchange Monte Carlo that we plan to discuss on the next week.

(Refer Slide Time: 01:08)




**KEYWORDS**

➤ Protein Folding



Pralay Mitra



**Protein folding problem**

**Problem Statement:** Given a protein sequence (primary structure), determine the three dimensional (3D) structure (or fold) of the given protein (tertiary structure) without the need for experimental validation.

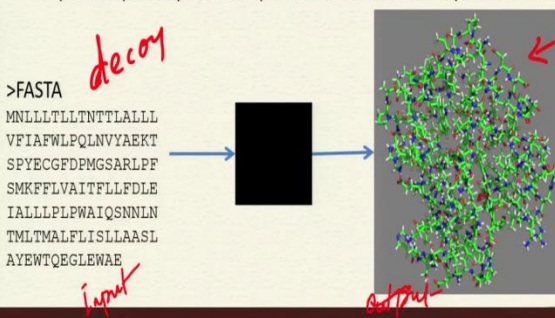

*decoy*

*input*


```
>FASTA
MNLTLTLLTNTTLALLL
VFIAFWLPQLNVYAEKT
SPYECGFDPMGARLPP
SMKFFLVAITFLLFDLE
IALLLPLFWAIQSNNLN
TMLTMALFLISLLAASL
AYEWTQEGLEWAE
```

*output*

*stable*

Pralay Mitra



So, the concept that will be covered is protein design problem. We will start with the problem statement, we will discuss different representation of the protein and then we will discuss also about the Monte Carlo algorithm in this particular concept. So, the problem statement says that given a protein sequence which is nothing but the primary structure that we discussed in the introductory week.

So, from that protein sequence, you have to determine the 3-dimensional structure or the fold of the protein which we introduced as the tertiary structure. But, for this determination or computational modeling, we should not look for experimental validation. Because if we go for

the experimental validation, then that is time consuming and we are trying to provide an alternative of that experimentally expensive way or process of determining the protein tertiary structure or protein 3-dimensional structure or the protein fold. So, that is why the protein folding problem the computational protein folding problem says that given the input protein sequence, you need to determine the 3-dimensional structure or fold of a protein without the need for experimental validation or verification.

So, when I say proteins a primary structure or protein sequence, then quickly it should come in your mind that FASTA or the sequence where the single character indicates one amino acids. Here is one such example, next with this input, I wish to map or I used to generate 3-dimensional structure which will be looking like this given here. So, this is my input and this is my output.

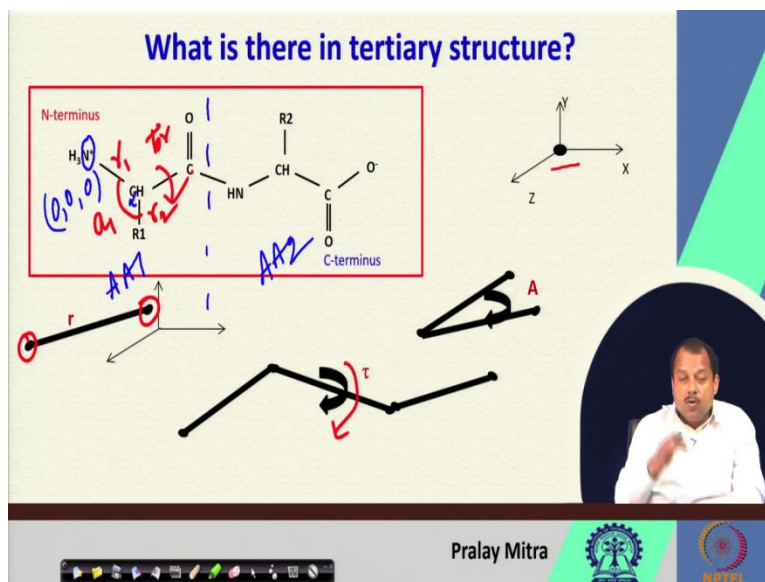
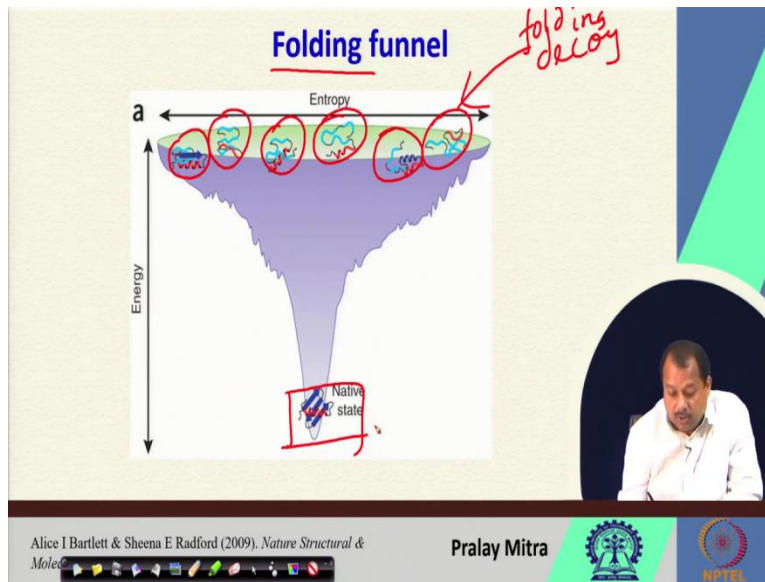
And what we need to do we need to design these black box. These black box will aid in transforming or modeling the FASTA or the protein sequence information to the protein structure information or it will give some three dimensional shape or fold. Definitely these structures should be stable one.

I can give you a number of such possibilities. But not all possibilities are valid. And I am not interested about all the possibilities rather those which are valid. Now regarding this validity. If you remember on the last lecture last slide. I introduce you the protein folding funnel or energy landscape the name funnel because it is looking like a funnel.

That funnel shape contains all the possible structures. Again, we can use the word decoy where different conformation or different structures look like a possible solution but may not be, we need to evaluate their energy and when the energy will say that it is at the tip downward then it is the stable one.

Again, we are making it simple at this point of time, because if we introduce a number of other environment variable or changes that one then the structure is going to change because the folding funnel will also going to change. But for the time being we wish to keep it simple. So, given a protein sequence, we need to map to the three dimensional structure of the protein sequence and for that, we need to design this black box. We will do this one with the help of the Monte Carlo technique and for that, we need a protein representation. How do we represent that one?

(Refer Slide Time: 06:01)



So, this folding funnel we discussed about. Now, different decoys. So, these again I call that as a decoy since it is in the context of the folding. So, I can say these are folding decoys. We are interested about this native state or we are interested to generate one instance whose stability or whose energy score will lie in this region that is our interest.

But in order to do that one we need to devise some techniques. So, that we can generate all the different possible conformations. Now, the first thing probably you are thinking that. So, you are given with only the sequence. So, how do you know that what will be the three dimensional

structure. I mean corresponding to each amino acid there are atoms and also I know the covalent bonding.

So, given those information, how can I generate that one? So, those things we are going to discuss. So, for that let us start with our initial introductory classes, where I mentioned that if there is only one point then I need to fix x y z three coordinate to fix its position in three dimensional space. If there is one atom like this. Now, if the atom is there, I can assume that as if this one, sorry there are more than one atoms.

So, let us assume without any loss of generality. Let us assume that this N is that point and again without any loss of generality. I can say this is my origin 0, 0, 0. If that is my 0, 0, 0 then it belongs to one amino acid and this is another amino acid AA1 and AA2. So, I know that there is a covalent bond with which it is the CH or this is hydrogen along with.

So, this is the C alpha. I know from the chemistry that what will be the ball length and also I know that if there are two points, one point here, another point here, if there are two points, then mentioning x y z for N and then r for this I can mention. What will be the coordinate for C alpha? Now, as of now, if I just mention r for C alpha then I understand the situation will be it is just a stick.

So, if this is my 0, 0, 0 and this is the r. So, it can be anywhere here, that is not a problem, it can be anywhere you just mention that is 0,0,0 this is r or using that one you can say this is 0,0,0 this is r 0,0. So, along the x-axis only that is not a problem. But after that one, you have to fix others also. Because once you will have this r then for this, there are three say this one, this one and this one.

Then you have to if I say this is r1, this is r2. Then this will be say angle a1 that you have to fix because there are three points. When there will be four say 1, 2, 3, 4 then about this there is another torsional angle. So, that torsional angle is present here. So, this torsional angle is also given here. Now, this thing will keep on repeatating in order to make it a complete structure in three dimensional space. Now, how should we proceed.

(Refer Slide Time: 11:21)

### Model representation

**Centroid**

1. Hydrogen removed
2. Except CB all other side chain atoms are representative

N-terminus

H<sub>3</sub>N<sup>+</sup>

CH

C

O

HN

CH

C

O<sup>-</sup>

R1

R2

C-terminus

N-terminus

H<sub>3</sub>N<sup>+</sup>

CA<sub>i</sub>

CB<sub>i</sub>

SC<sub>i</sub>

C<sub>i</sub>

O<sub>i</sub>

N<sub>(i+1)</sub>

CA<sub>(i+1)</sub>

CB<sub>(i+1)</sub>

SC<sub>(i+1)</sub>

C<sub>(i+1)</sub>

O<sub>(i+1)</sub>

C-terminus

Pralay Mitra

### Model representation

**Cartesian Coordinate System**

**Torsional Angle System**

$\phi_i, \psi_i, \omega_i$

N-terminus

H<sub>3</sub>N<sup>+</sup>

CA<sub>i</sub>

CB<sub>i</sub>

SC<sub>i</sub>

C<sub>i</sub>

O<sub>i</sub>

N<sub>(i+1)</sub>

CA<sub>(i+1)</sub>

CB<sub>(i+1)</sub>

SC<sub>(i+1)</sub>

C<sub>(i+1)</sub>

O<sub>(i+1)</sub>

C-terminus

N-terminus

H<sub>3</sub>N<sup>+</sup>

CA<sub>i</sub>

CB<sub>i</sub>

SC<sub>i</sub>

C<sub>i</sub>

O<sub>i</sub>

N<sub>(i+1)</sub>

CA<sub>(i+1)</sub>

CB<sub>(i+1)</sub>

SC<sub>(i+1)</sub>

C<sub>(i+1)</sub>

O<sub>(i+1)</sub>

C-terminus

Pralay Mitra

## Conformational movements

- Residue-level movements
- Segment-level movements
- Topology-level movements
- Global movements

Pralay Mitra

This is my kind of presentation for the protein. So, in amino acid there are two amino acids. Now, this I am making a simplified representation like this for our purpose. So, what is extra here you see that in this case R1 was there R2 was there that was a side chain and I mentioned that if I go that way the first C alpha that I will encounter, sorry first carbon that I will encounter is beta.

So, that is CB, CB. Now, based upon what kind of amino acid it is. So, it will vary that how many atoms will be there as the side chain in the simplified representation. What I am assuming that along with CA or C alpha there is C beta which is connected by covalent bond this solid line, solid line after that one all the side chains are representative one, they are not actual one and by that representative what I will do.

I will compute their centroid all the atoms side chain atoms except CB or C beta, which I already considered. Now, the centroid of all other atoms will be denoted as SC and that is why you see that there is a dotted line not solid line indicating there is no covalent bond. Because after CB there may be C gamma or maybe another item at the gamma position.

But I am not concerning that one I am calculating the centroid of all other atoms and I am putting one representative atom as if there since that is representative. So, definitely there is no guarantee and mostly that will be beyond the length of a covalent bond. That is why there is no covalent bond dotted lines are there.

Now, the computation time or overhead will increase as the size of the protein will increase. But we are helpless. If there are say 230 number of amino acids. We have to model 230 number of amino acids that way we cannot reduce anything. But what we can do that we can eliminate all the hydrogen atom for the time being, after modeling this part, after modeling non hydrogen atoms.

Then we can incorporate hydrogen atoms appropriately and since it will be the incorporation of the hydrogen atom. So, that program will not be much difficult computation point of view. So, that we can fix later. So, for the time being I that is why you can see here I excluded all the hydrogen atoms except for this part. Because that is the starting position that is why but if you want to you can exclude this one also.

So, in this reduce representation first thing, I removed all the hydrogen. Number two except CB or beta position atom all other side chain atoms are representative, they are not correct one. I mean they are not physically correct position that is the centroid position. Now, this is my first assumption or representation. Now, after this one I am going for model representation in two different systems, one is the cartesian coordinate system, another is that torsional angle system.

In the cartesian coordinate system as the name suggests. So, corresponding to each. So, here this C is  $SC_i$  and  $SC_{i+1}$  these two was missing. Now, corresponding to this cartesian coordinate system corresponding to each atom there is  $x_i$ ,  $y_i$ ,  $z_i$  corresponding to each atom. So, for this this this this this this this this this this this and this nitrogen. Let us assume hydrogen we are not considering for the torsional angle representation.

We are having how many torsional angles  $\phi$ ,  $\psi$ ,  $\omega$ . So, this position this position this position this position this position this position now, for this position there is a little problem. So, for this position for this one let me give with a different color for this position, torsional angle at this position problem is that. I have to consider say this or this but without this SC consideration I cannot go however, this SC is the centroid position and is not physically correct position.

That is a problem that is why what I am planning to do right now that let us not include this side chain torsional angles rather let us fix the torsional angle of the main chain. Main chain means this N C  $\alpha$  C N C  $\alpha$  C and then in will come that way it will go and those three torsional

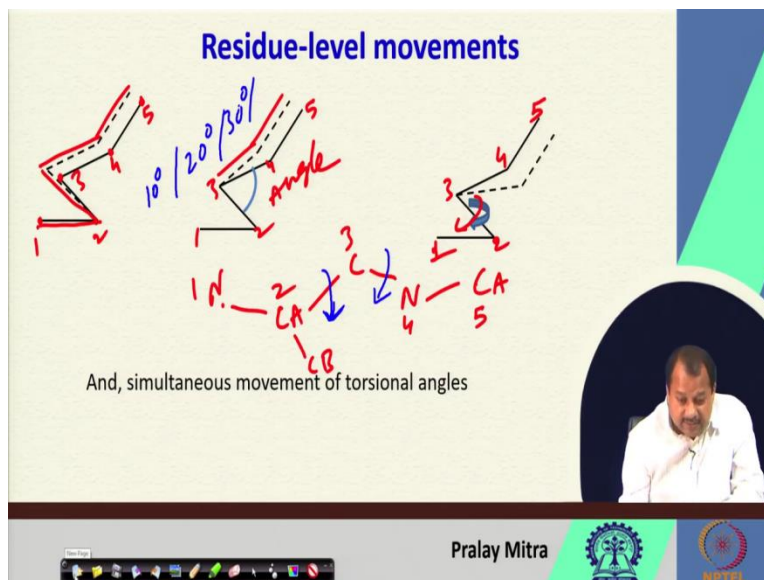


angles phi psi omega and that way if it goes then phi i, psi i, omega i then i plus 1, i plus 2, i plus 3 that way it will keep on going.

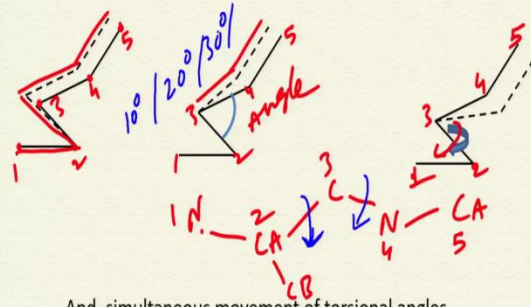
So, let us fast fix this one and this cartesian coordinate system once those are fixed, then rest can be fixed and also if I can fix the backbone, then the fold of the protein structure will be fixed. So, I need to fix the backbone or this phi psi omega along with the other cartesian coordinate information there. Next, we are going for the model representation where we will include the different conformational movements.

So, first one is the residue level movement. Then we will consider the segment level movements then we will consider topology level movements then we will consider global movements. So, this global movements we will consider later not in the current lecture. So, when we will discuss the protein folding algorithm then we will discuss that global movement but first three residue level movements, segment level movements and topology level movements we will be discussing.

(Refer Slide Time: 19:10)



## Residue-level movements



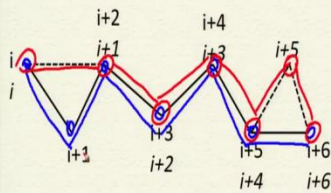
And, simultaneous movement of torsional angles



Pralay Mitra

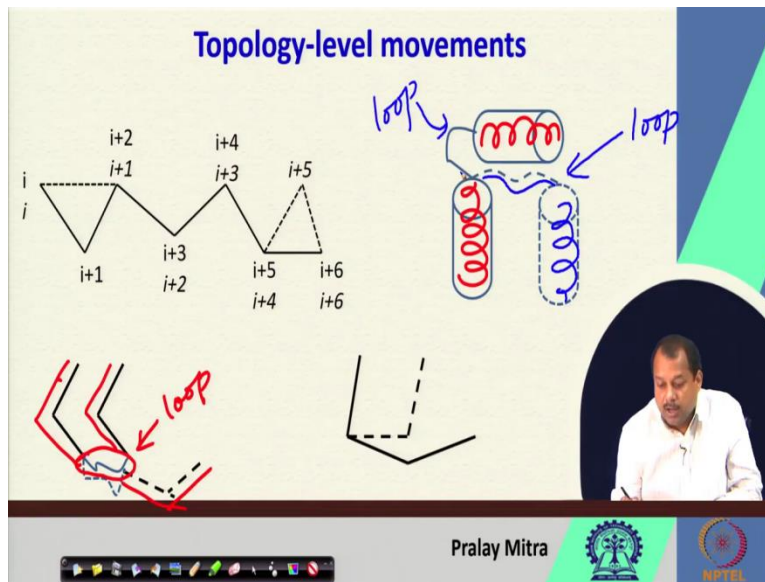


## Topology-level movements



Pralay Mitra





So, when I say it is residue level then you should remember that residue means what? N C alpha C C beta is there after that one is the side chain, etcetera. Then N that way it will go. Now, you see that there is one point here, another point here this is a covalent bond, this solid line. Another point here, another point here another point here.

So, considering the situation that there can be some bond length. So, this is one representation. So, this solid line and this solid line, this solid line plus these dotted lines is another representation. So, this way when we have 1 say 1, 2, 3, 4, 5, 5 consecutive atoms. So, 1234 another C alpha and 5, 5 consecutive atoms, then we can have this bond information, then we can have this angle information again 12345 and you can see that if the angle changes.

So, the angle because of the change of the angle at this position, this blue color angle position. So, these 3 up to 3, 123 these 3 positions will be same, but position of the 4 and 5 will change. So, this dotted line indicates the change position. So, that position will change. If I assume that torsional angle will change like this and this is 1234.

So, good enough to have one torsional angle because of that one the fifth atom will also change and this dotted line indicates that because of the change of the torsional angle, that will also change and simultaneous movement of the torsional angle which means that. So, this N C alpha C N C alpha. Now, if I take a different color so, this is one torsional angle that you accept this 1234 and 2345 is another torsional angle.

So, this simultaneous change of the torsional angle will also be there. Now, recall the situation of the protein complex modeling that we discussed and then we mentioned that whenever there is a change, then you decide what will be the incremental changes. So, in this case also you decide that the angle when there will be a change.

So, whether it will be 10 degree 20 degree 30 degree similar to this torsional angle also and simultaneous movement of the torsional angles also and for each movement I suggest to generate one lookup table or hash table and storing some database. So, that when we will integrate that one in some simulation technique those will be called and their fitness or after calling that one their energy will be scored.

And then that energy will be considered or that energy we will check for the acceptance or rejection of that particular conformation. Now, this is about the residue level movement. If I go for the segment level movement, then first thing I need to decide that what will be the length of my segment. You can consider that  $C\alpha_i$  and  $C\alpha_{i+1}$  or you can consider  $CA\alpha_i$  to  $C\alpha_{i+2}$   $c\alpha_{i+2}$ .

So, three segment also you can consider. Now, when we consider that one then this can structure is one particular confirmation or one instance. Now, another is that if say this will be. So, what is happening this N is pulling down and because of pulling down N. So, there is a small change in the others also keeping  $C\alpha$  here and  $C\alpha_{i+1}$  here fix this can be pulled up this can be pulling down at this position or at this position.

And you are considering that one. Now, try to understand the situation last slide I did residue level movement. So, only the translation angle change torsional angle change simultaneous change of the torsional angle was there at the very core. Now, I am taking instead of one residue two or three consecutive residues and for those residues. I am doing the movement and because of the stretching pulling down pulling up, etcetera or say out of the plane or down to the plane that way.

When we are doing that one because of that what is happening there is a change at this segment level. So that is my segment level moment. Now, next is my topology level moment and I am talking about the topology when I am talking about the topology now it is  $i$ ,  $i+1$ ,  $i+2$ . So, at the amino acids level. So, zooming out residue, segment, now topology and when it is that

then you see what I am doing here that  $i$  so  $(i)$ (25:29) indicates one particular conformation and along the dotted line and solid line along with the simple straight indicates another conformation. So, this  $i$  it can be one conformation say  $i$  this then this then this then this then this another is that say  $i$  after that one this is my  $i$  plus 1. So, when it is going.

So,  $i$   $i$  plus 1,  $i$  plus 2,  $i$ ,  $i$  plus 1,  $i$  plus 2,  $i$  plus 3,  $i$  plus 4,  $i$  plus 5,  $i$  plus 6 another situation is with red color. So, this is my  $i$  this is my  $i$  plus 1, this is my  $i$  plus 2, this is my  $i$  plus 3, this is my  $i$  plus 4, this is my  $i$  plus 5, this is my  $i$  plus 6. So, using red color, it is this one, this one, this one, this one, this one, this one following the dotted line.

So, that is so,  $i$  is one residue or amino acids  $i$  plus 1 another,  $i$  plus 2 another. So, that way I am moving or using this one, it will be even more clear for you that this is one situation. So, the solid line here between two black say parallel one. So, there is a blueish line solid line indicates one conformation, dotted line indicates another conformation and this region is a call as the say loop or non regular.

There is no regularity this region by this region it is connected this is regular this is regular, but it is connected using non regular or this is regular and this is another regular. Now, another situation is something like this, here it is even clear. Now, each point indicates that is one amino acid and one straight line indicates the connection again it is not the covalent bonding right now here.

So, two points being connected one is the C alpha  $i$ , C alpha  $i$  plus 1, C alpha  $i$  plus 2, C alpha  $i$  plus 3 and what is their relative position in three dimensional space that is being considered here. Because in between C alpha  $i$  and C alpha  $i$  plus 1 that N C those things I already considered in residue level and the segment level. So, right now I am not considering. Right now, I am considering only at the amino acid level.

Now, here you will see that this is say one secondary structure. Let us assume this is helix. This is another helix. Now, this is another possible helix. Now, as I mentioned that these dotted line here this is my loop region. This is my loop. So, you will see that loop is not that much rigid or there is no regularity you need to mention that is why it can move, it can move a bit more compared to the regular structure.

So, this is one topology the following the red and the solid that loop and following the dotted loop and the red and blue. I am having another topology. So, these two are two different topology. That way we are having different level of topologies.

(Refer Slide Time: 29:46)

**Conformational database**

- Need a database to store all such movements
- Need a hash table/look up table for quick retrieval of the movements from the databases.

Thank you

Pralay Mitra

Now, once you will have this residue level, this segment level and the topology level chain conformations. So when I say conformation that indicates the different when I say conformation that indicates the different the orientation. So, you can put it in a database now need a database to store all such movements that database does not mean that it is it will be stored in the secondary structure.

It can be very much residing in the RAM. But for the time being you need to encode that one and maybe in some file in some format in some data structure. And there should be a hash table or lookup table. So, that quickly you can retrieve those and then fit and check that whether the correct one you are getting or not. So, that is regarding the database and once you will have the all the different conformation in place.

And you have a lookup table or hash table and using that one you can access that one then you are ready to integrate that one in the Monte Carlo simulation techniques. So, what we discussed in this lecture starting with the definition of the classical protein folding problem. Where input is a protein sequence and you need to output a three dimensional structure or fold of that protein without the need of the experimental validation.

Then we see that when no information is provided to you kind of you are starting with an ab initio then beta to reduce the problem size by eliminating hydrogen atoms and then do not consider the side chain up to C beta you consider and rest of the atoms you represent it using some centroid using some dummy atom.

In some representation in some software tool. Even the C beta is also not considered up to C alpha and all the side chain that is R1, R2, etcetera is represented by some dummy atom using some and position at their centroid. Now, we are generating the different conformation at the residue level, at the segment level and at the topology level, then we plan to store that one.

And we also plan to design some hash function or a hash table or lookup table so that quickly we can access that one. So that during the Monte Carlo simulation process at each iteration when we will look for generating the new instance of that new instance then quickly we can access that one we can fit. We can check their energy value and accept or reject. Thank you so much.