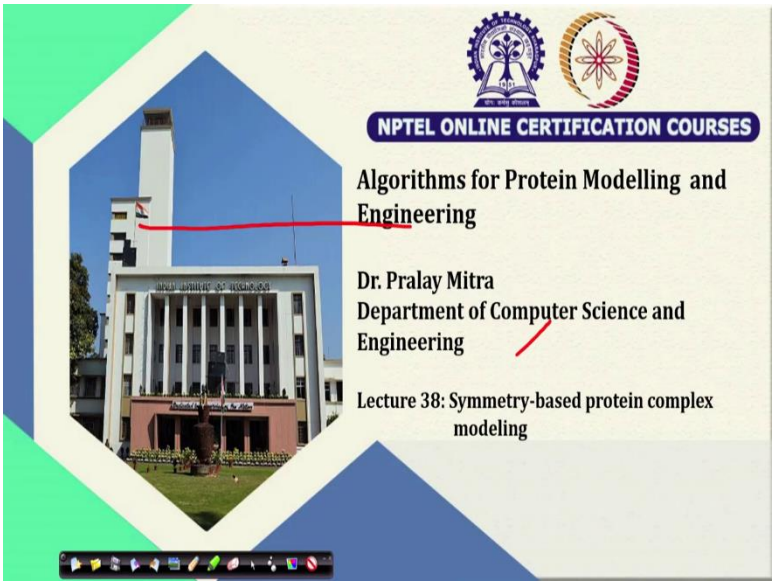
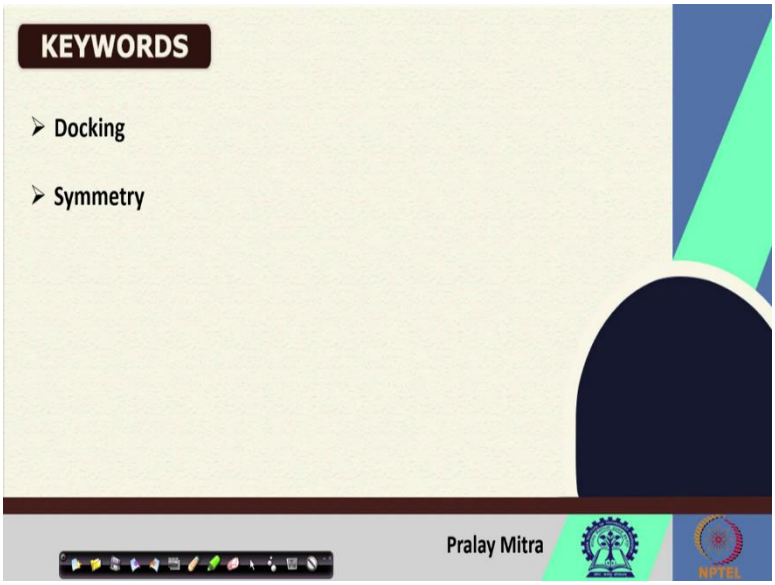
Algorithms for Protein Modelling and Engineering
Professor Pralay Mitra
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur
Lecture 38
Symmetry-Based Protein Complex Modeling

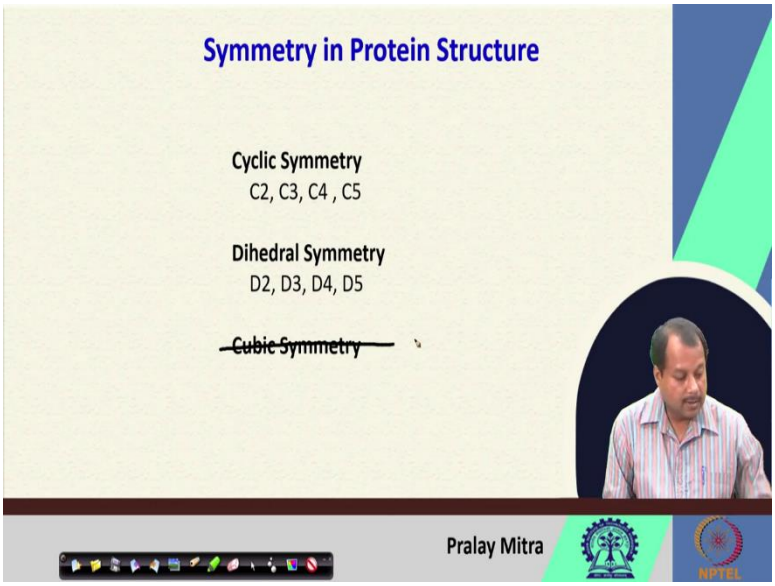
Welcome back. So, in this particular lecture, I am planning so discuss one symmetry-based protein complex modeling technique. As I mentioned that symmetry is present in protein complex. So, it has been observed, and there are a lot of analysis based upon the symmetry. We discussed about a symmetry in the context of the protein, we have the cyclic, dihedral and cubic symmetry, but only the cyclic and dihedral symmetry exist, and we focused on that one.

One application, we demonstrated, where we are inferring the assembly of proteins from the crystals lattice, based upon the symmetry. So, there is not only symmetry, but actually, there is missing learning classifier and the symmetry together they are actually forming a competitional framework using which actually, we are basically predicting that what is the most possible biological assembly, and what is the structure of that one.

(Refer Slide Time: 01:26)

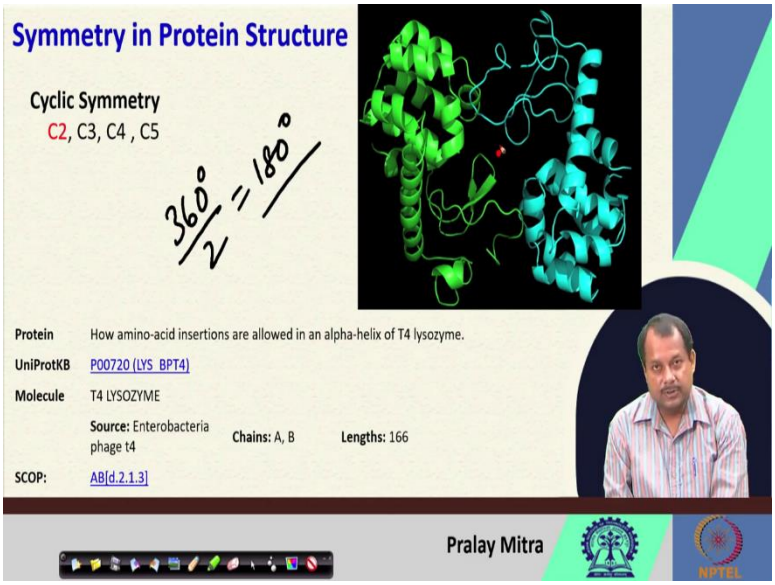






Now, today, I will be talking about one protein modeling specifically protein docking technique. So, based upon the symmetry that is called as the SymmDock or symmetry-based docking. Now, to give you the recapitulation regarding the symmetry, cyclic symmetry C2, C3, C4, C5 etc exist, and dihedral symmetry D2, D3, D4, D5 exist and the cubic symmetry exist. But we are not interested about this cubic symmetry. So, only the cyclic symmetry and dihedral symmetry is of our interest. And we will be discussing those. We discussed those.

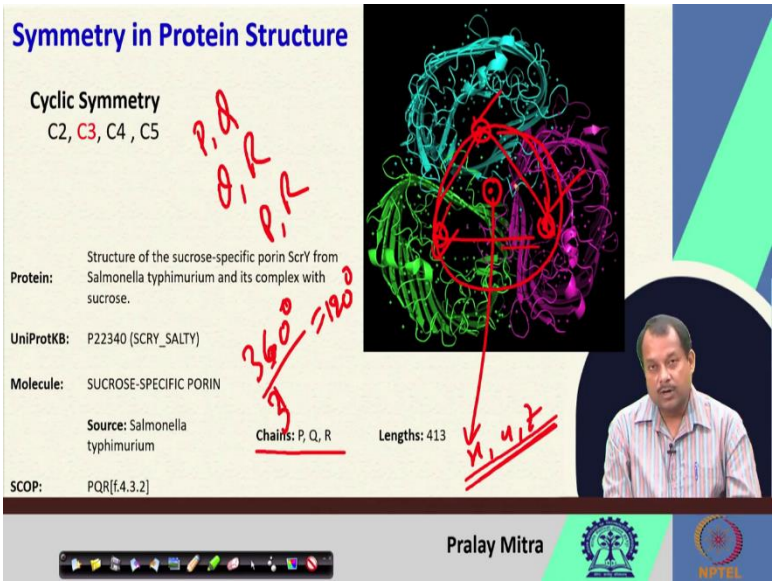
(Refer Slide Time: 02:17)

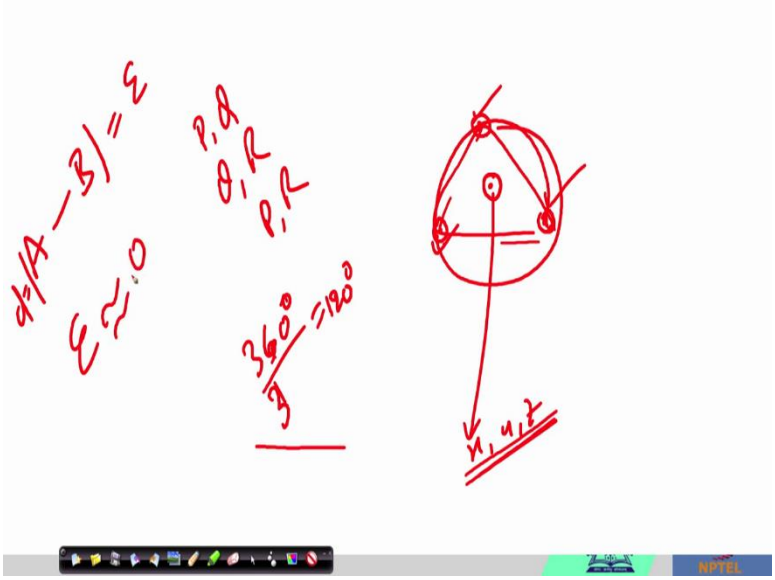


Now, to see again some of the structures, this is relevant, because then only we will understand the limitation of the symmetry-based protein modeling and also the advantages, both, we will discuss. Now, when we see the symmetry in protein structure this is the same slide that we have discussed on the last week. It is C2, is the symmetric. C2, means, so C2 means, it is basically 360 degree divided by 2 which means 180 degree. So, 180-degree rotation about the symmetry axis, definitely, you have to identify the symmetry axis. So, some not here actually it will be somewhere here.

So, if I draw one perpendicular axis, one axis perpendicular to this plane through the red circle point then we will, we can give 180-degree rotation about this plane or about that axis, about the axis actually, then, we will see that green will go to cyan, cyan will go to green and that way we will not able to discriminate or distinguish that whether they are same structure or not. So, there is symmetry. Now, for this it is simple, identification of the symmetric axis and accordingly we can do.

(Refer Slide Time: 03:52)





Now, instead of C2 if I look at say C3, the situation will be more or less same. Here a probably I missed to tell you that the dots, small dots, those are actually some concept water molecules, which are reported in the crystal structure and also displayed here. So, this is C3. Now, if it is C3. So, I will get some axis somewhere here and then I can compute.

Now, the challenge for this identifying the symmetry, whether the symmetry exists or not and what kind of symmetry it is, is first of all identification of the symmetric axis. If I do not able to identify the symmetric axis, then it will be difficult for me to identify what kind of symmetry is there. No doubt, this is basically my C3, as you can see here.

Now, you see that the red color, if I assume, is my point through which one axis or one axis is passing perpendicular to this plane. If it is passing then above this I am rotating. Now, how

do I find this one? So, for this, perhaps, it will be very easy for you. Why? First of all, you identified that there are three chains, that information he will readily get.

So, this is what I am going to analyze. Then you will understand how to identify the axis, what are the different ways to identify the axis, and then, conclude that what kind of symmetry is there. So, in this particular case what you can do, is that, so, you can easily identify how many chains are there that is P, Q, R.

This you will get just by analyzing just by analyzing your protein structure, that there are three chains, P, Q, R then you can go for PROIs alignment to check whether it is a homomer or not. So, that is also required. So, in stepwise way, if I do, then first, I have to identify how many chains are there.

Then I have to check, whether, they are homomer or not. Because we know that, if it heterodimer, then in heterodimer complex the symmetry does not exist. So, in the second stage, I will do the sequence alignment. I know the dynamic programming method through which I can do the sequence alignment. So, after aligning the sequences, I came to the conclusion that it is indeed the homotrimer. So, up to this, it is fine.

So, first step, I identify the number of chains, that is three nature, then, I do a pairwise alignment, all pair alignment and based upon that one, I conclude that it is a homotrimer. But I also mentioned while doing the say pairwise alignment, if the identity is not 100 percent, but say, but more than say 95 or 96 percent you check whether that is because of the missing coordinate or actually they are different. So, what is the reason? So, based upon that one you have to conclude.

If there are because of some missing coordinate information. If there are some missing coordinate information then you cannot have that information, missing coordinate information, then you, if there are some missing coordinates then you may not get that P, Q is equivalent to Q, R or it is equivalent to P, R etc. But if not, then actually, you will come to the conclusion that it is homodimer, so that is the step two.

After doing that one, for this, one simple thing you can check that what is the centroid of the chains. So, this is one chain, this is another chain, this is third chain. Now, what is the centroid? If the centroid is here or if you identify the centroid then you need to pass the axis through that centroid, because I am looking for the cyclic symmetry. If I am interested to pass an axis through the centroid then I got one point.

So, when I got one point then basically what I get is one coordinate say X, Y, Z but I do not get the access. So, even if I consider the line of equation, then, through a single point infinite lines can pass. So, I need at least some other information, so that, I can draw or I can pass one line. So, in order to do that one, what I can think of that, I can identify some points.

So, one is computing the centroid for all the three chains, and that is here, another is computing the centroid of each chains. Now, if you compute the centroid of each chains, and the centroid of this one then definitely this is X, Y, Z your axis will pass through these X, Y, Z, but how this can will be off helpful.

So, what you can think. So, if the centroid, this centroid and the centroid, if you draw one circle through this then your line should be perpendicular to the plane containing these three points or containing the circle, which are on the circumference of the circle. Now, if you compute, or if you identify so, I agree, that given one particular point say X, Y, Z in 3D infinite number of straight line can pass through that one.

Now, out of those infinite straight lines if you can identify one straight line, yes, that will be only one straight line, which will pass through X, Y, Z and will be perpendicular with respect to one plane, that is going to be your axis. And perpendicular to which plane? How do you decide about the plane? So, you are given with the three points, so through these three points you can consider.

So, since these three points will be non-collinear, so through these three points you can have one plane. So, since you will have one plane through these three points then, you can basically compute the equation of the plane then your job will be to identify one straight line which is perpendicular to the plane and passing through the point X, Y, Z, you will get the axis of your symmetry.

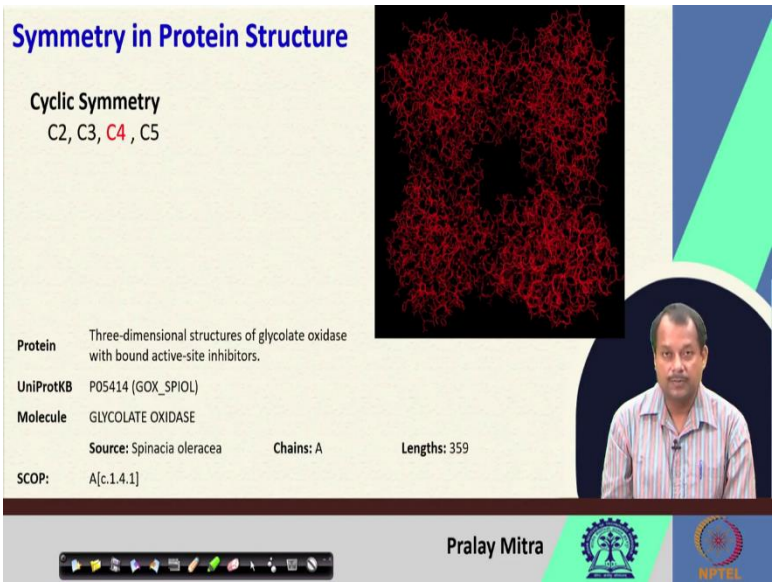
Now, about that axis since three actually chains are there. So, it might be and also that three chains are homotrimeric chains, so straight away what you can do the 360-degree divided by 3, which means, 120 degree. So, about this axis either in clockwise or anti-clockwise you can give 120-degree rotation.

You know the rotation matrix from our previous classes. So, we give another one rotation and after the rotation and before the rotation, what will be the deviation in terms of the RMSD, we discussed about RMSD. So, alignment is not required because I am assuming that after the rotation it will be aligned. If not, then, it is not actually symmetric in nature. So, after the

orientation, after the rotation what will be the structure, and before the rotation what was the structure, between these two you compute the RMSD.

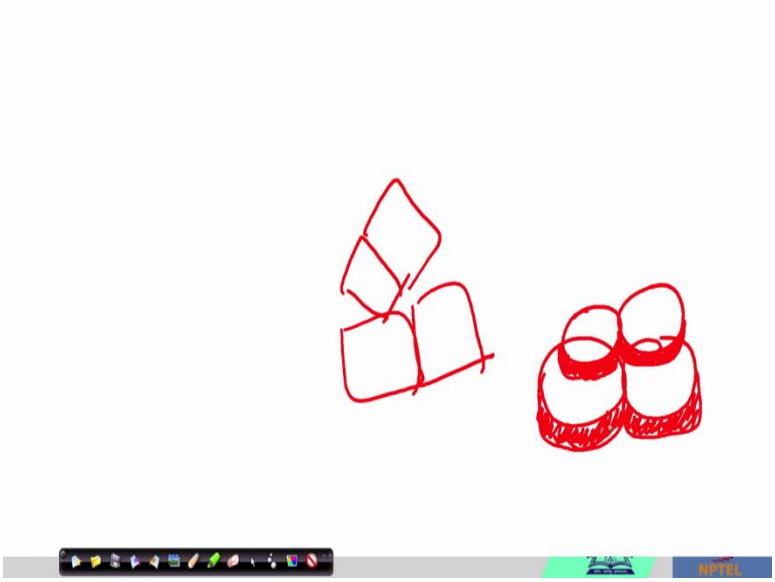
Now, if I assume, that initially my structure or the configuration was A and now it is B then you compute the distance between these two. And if I say, that this distance is going to be epsilon, if epsilon is very close to 0 then you say a symmetry exists. Otherwise symmetry does not exist or you identified wrong axis both are possible.

(Refer Slide Time: 12:45)



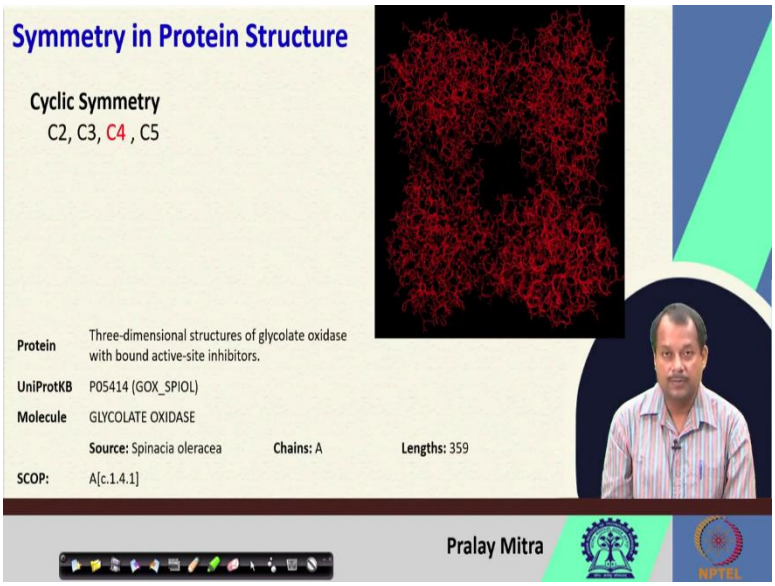
So, this is with the C3, I believe, it will be same for C4 four also where you have four chains. And of course, the four chains are homologous. It will be trivial it will be trivial in this case because there is only one chain out of that one, four chains has been constructed using some technique. So, it is trivial, so you can generate that one. So, that is, about the cyclic symmetry but it is not that much trivial for the dihedral cases.

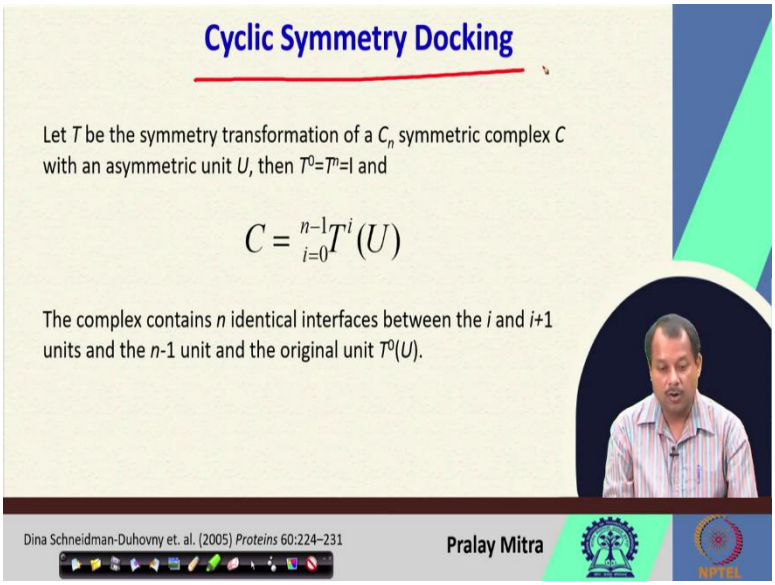
(Refer Slide Time: 13:13)



Because, you know, for the dihedral cases what we did that, say, this is one, this is one, and on top of this these two. So, since, it is actually 3D and this is one, and on top of this, this is one, so that is why I am not able to draw it correctly, but if it is something. So, you can consider that something, then it is not that much trivial or easy.

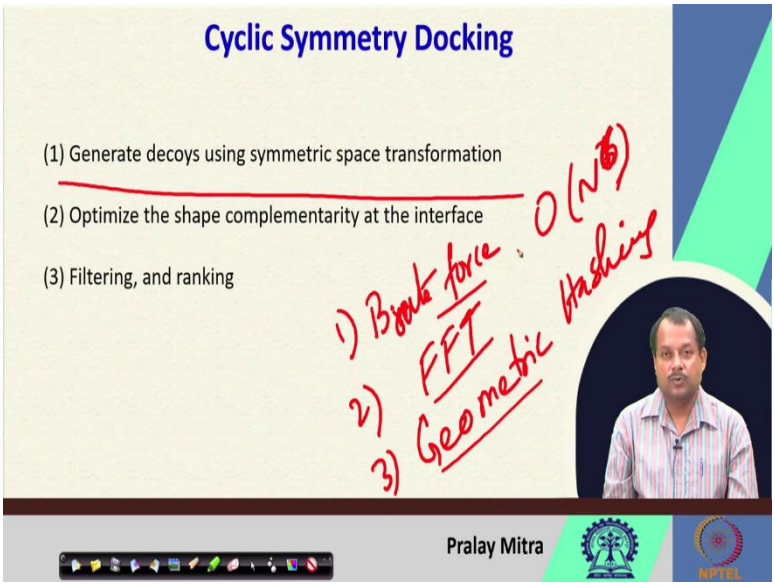
(Refer Slide Time: 14:26)





That is why the protein modeling or protein docking technique that exploits the symmetry information uses only cyclic symmetry information. So that way, it is going to be cyclic symmetry docking. And the definition is, let T be the symmetry transformer of a Cn symmetric complex C with an asymmetric unit U, then T0 equals T1 equals two I and C equals i equals to 0 through n minus 1, T to the power i, u. So, the complex contains n identical interfaces between the i and i plus 1 units, and the n minus 1 unit and the original T to the power 0 u unit. So, that is a theoretical definition of that way, the cyclic dock, symmetric docking is being used.

(Refer Slide Time: 15:35)



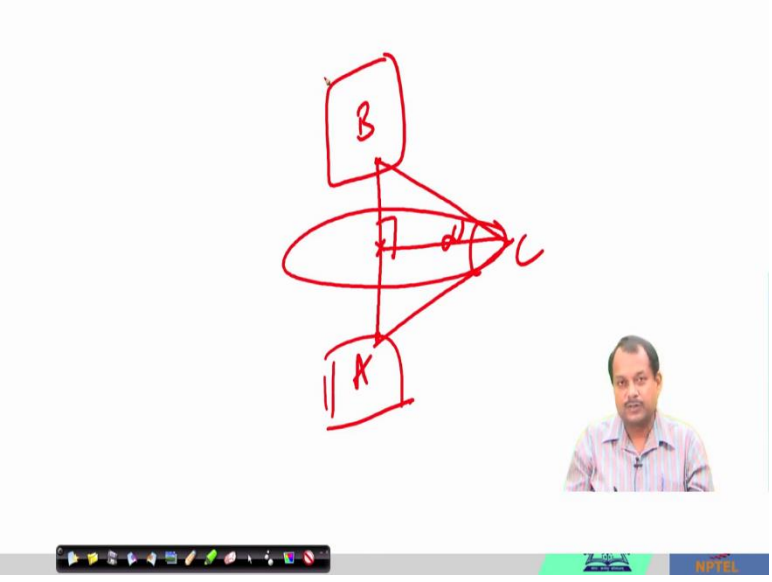
There are three steps in cyclic symmetry docking, generate decoys using symmetric space transformation, optimize the shape complementarity at the interface, filtering and ranking.

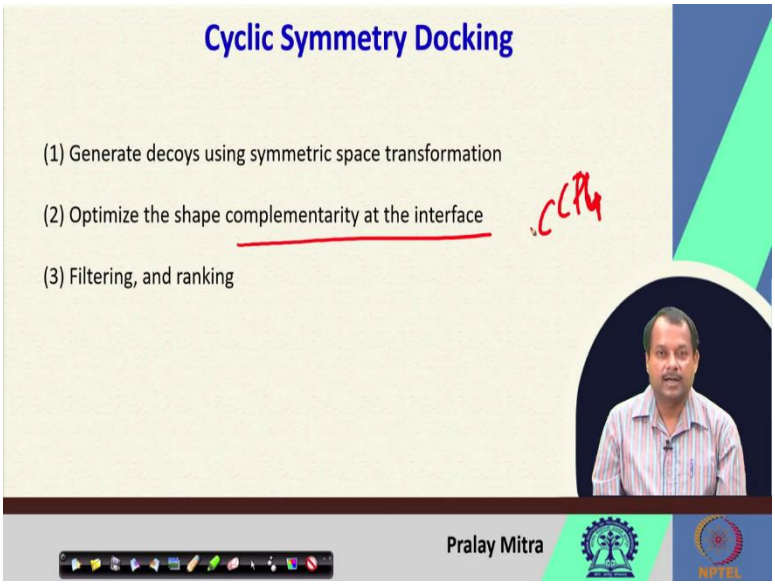
Can you please tell me since we have done a lot of discussion on these docking and protein complex modeling that what is new?

I mean, what is the unique contribution of the cyclic symmetry? In which step? If you look carefully then you will understand it is only at the generation phase. So, far what we have done so, we did brute force where complexity was N to the power 6, 6 nested for loops was there. Along the X axis, along the Y axis, along the Z axis, about the X axis about the Y axis about the Z axis.

So, three translates, three rotation and all are nested that is why N to power 6 that way you can generate all possible orientations or decoys for the docking. Next, we modified this one for implementing using fast Fourier transform where the complexity has reduced in cube log Ql Next, we discussed about geometric hashing where memory requirement is very high, but if I assume that nowadays RAM is cheap, very cheap, and most of the desktop systems are acute with say at least 18 to 16 GB RAM then it is not a problem. So, you can define large hash table and you can use that one for your own purpose. So, there is geometric hashing FFT brute force technique we discussed. Now, here, we are using the cyclic symmetry docking. So, how we can generate that one? The same way we define the axis.

(Refer Slide Time: 17:45)





So, basically, in order to generate that one, so, what you need to do so, let us assume that there are two points. Then you assume the existence of a plane, the angle, sorry the straight line between A and B is passing through the center of say circle, then, here there is a point this is the right angle I mean 90 degree and this is my alpha.

So, if it is, then basically what you are doing if A is given to you, you can generate B. Either you can take, what symmetry will be there as an input from the user or you can generate, as many symmetry information as you want. Say C2, C3, C4 using that one you can generate so many complexes or you can take the symmetry information as an input from the user and accordingly you can generate.

Now, when you will generate then a lot of complexes will be generated. I mean, so from here you are generating this one, using symmetry this one, symmetry this one symmetry, that way you consider, as if, so say a given one protein structure and also symmetry is also given as an input. Now, if you use that symmetry to generate all the possibilities then as if you are creating a crystal lattice.

If you are creating that crystal lattice then following the previous algorithm that we have discussed for inferring the biological assembly from the crystal lattice, so what you can do, you can combine that, you can exploit that information basically in order to identify, which interface is going to be the correct one or what you can do which is suggested here also that optimize the shape complementarity at step 2 it is suggested, optimize the shape complementarity at the interface.

The third point is, filtering and ranking, it is same for the previous or other docking techniques. So, shape complimentary, you can you can optimize at the interface, for that, you can have say NSC, you can have only SC surface complimentary, which comes with CCP4 package CCP4. So, you can calculate NSC, SC or are there any surface complimentary atomic contract vector. By any method you do the scoring, in order to identify, which is the biologically relevant one, and then you output. But the primary contribution for the SymmDock is that way it is generating the complexes using the symmetry information.

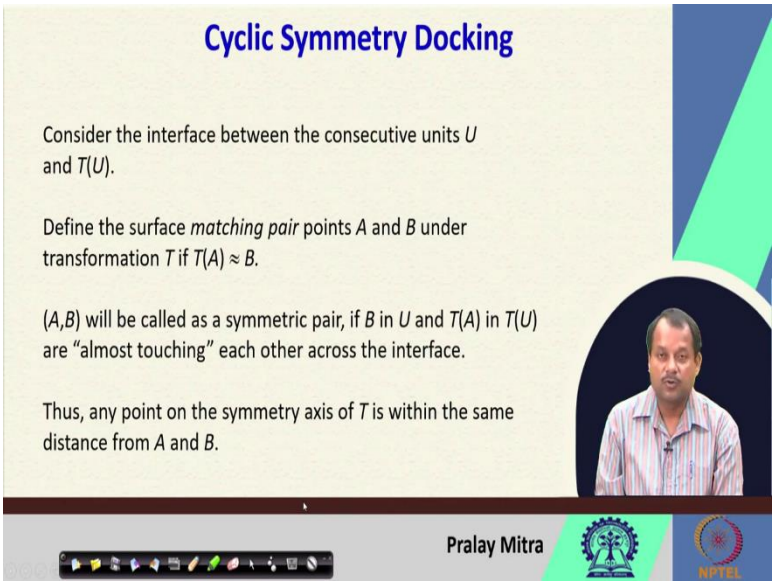
So, symmetry exists in the protein, if it is a homomer and if it is a homomer then you can exploit that symmetry in commerce and to generate the complexes and what is the advantage. You see, that, first of all, when it is a heterodimer, then the symmetry-related docking is not of any use, you forget about it.

But if it is a homomer then you know that I know that symmetry exists. Now, using brute force technique or Fast Fourier technique or say or say geometric hashing-based technique, if I generate a number of complexes which are not symmetric in nature, then it is not going to stay because symmetry exists if it is homodimer.

Say, for example, you are given two protein molecules as an input one is dimer and other is a monomer. But, the dimer chains are that is a homodimer and single unit of or subunit of that homodimer is exactly matching with the third one, which means, after the docking it is going to be that trimer. If it is, then, your brute force technique, your FFT technique, your geometric hashing, FT dock, patch dock that is going to generate a number of possibilities.

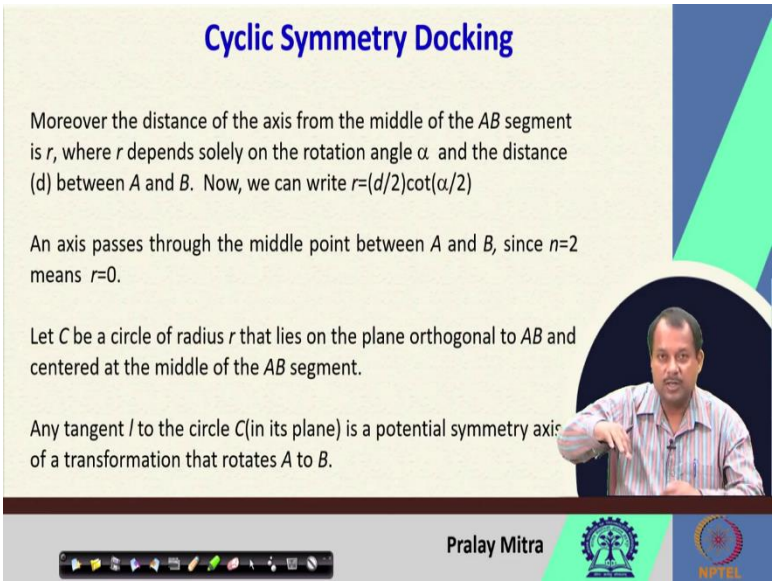
But cyclic symmetry-based docking if you go then it will get all the symmetry information. And since symmetry exists and the trimer is going to be C3, since its homotrimer mostly it will be going to the C3, so it will be very easy to generate the complexes, and that too it will be very fast and also the number of possibilities will be very less. So, chances are very high that you will get the correct solution. So, that is also the advantage of this symmetry-based method.

(Refer Slide Time: 22:30)



Cyclic symmetry docking. Consider the interface between the constitutive units U and TU. Define the surface matching pair points A and B under transformers and T if T A equivalent to B. AB will be called as a symmetric pair if B in U and TA in TU you are almost touching each other across the interface thus, any point on the symmetry axis of T is within the same distance from A and B, that is what I drawn just few minutes ago.

(Refer Slide Time: 23:09)

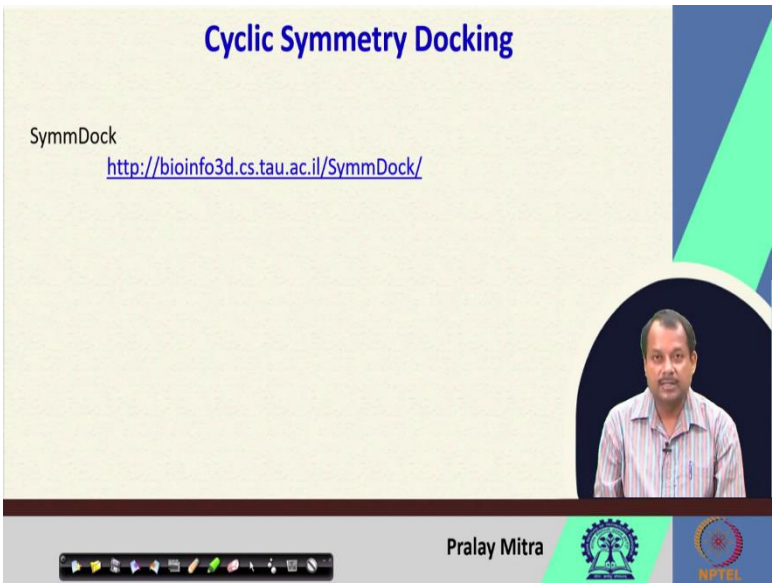


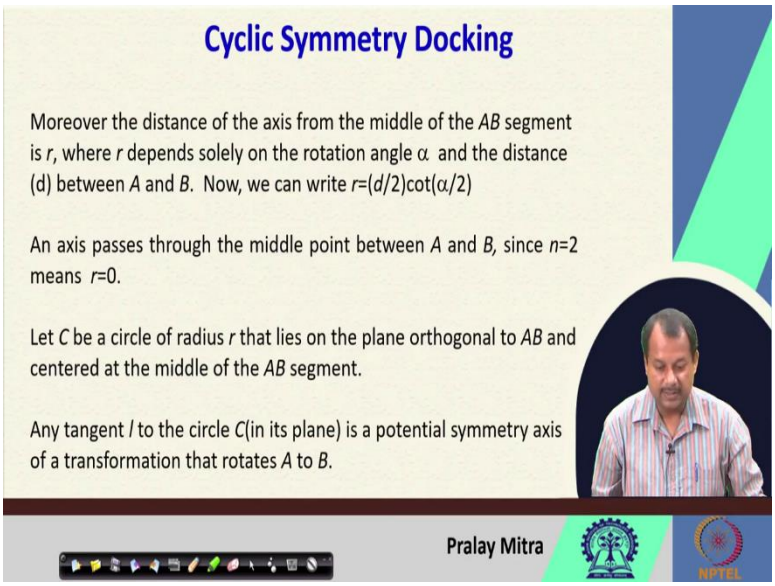
Moreover, the distance of the axis from the middle of the AB segment is R where R depends solely on the rotation angle alpha and the distance d between A and B. Now, we can write, R equals to d by 2 cot alpha by 2. And that way, if n equals to 2 means r equals to 0. Let C be a circle of radius r that lies on the plane orthogonal to AB and centered at the middle of the AB

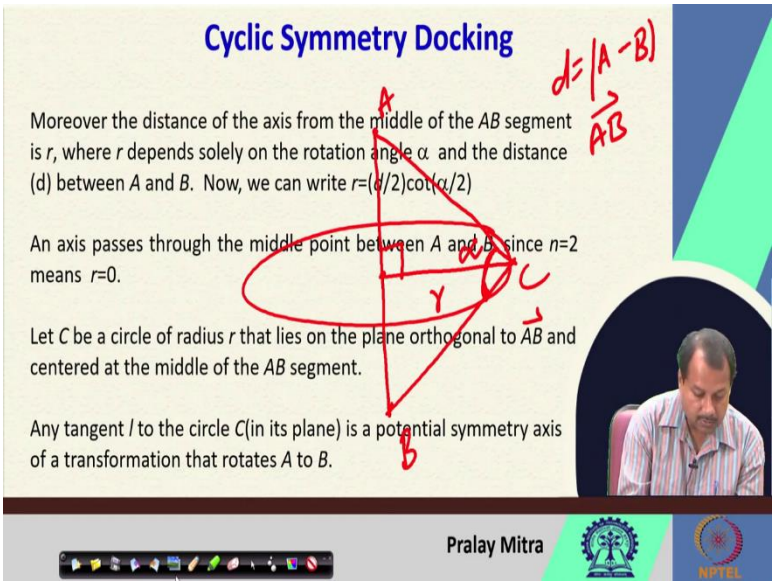
segment any tangent I to the circle C in its plane is a potential symmetry axis of a transformer some that rotates about A to B.

And when it will rotate A to B then you create one instance of A at B and A was already there. Now, you created a new instance at B so AB will be one complex. So, it is very simple, and intuitive also.

(Refer Slide Time: 24:07)







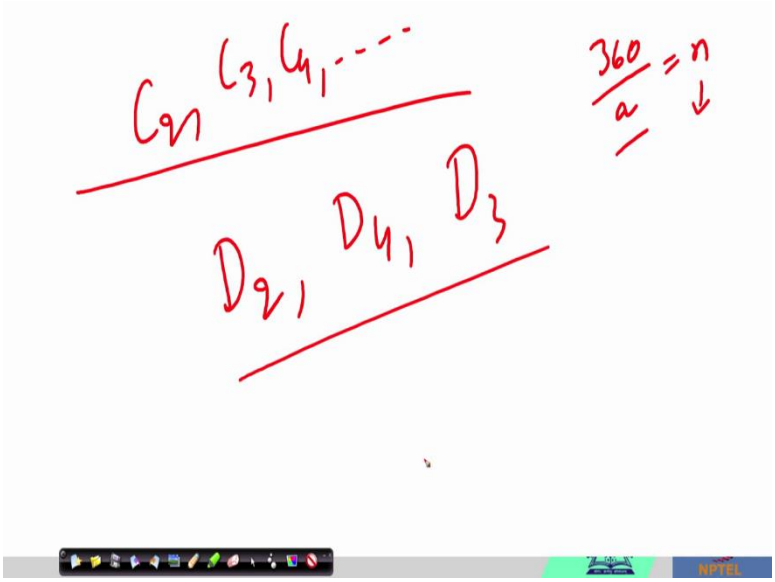
So, this particular technique is basically hosted at the Tel Aviv University. So, with the name SymmDock. So, it is basically symmetry-based docking. That is the work of Ops and NOC. So, if I revisit this one, what it is telling. So, again, to draw that structure the distance of the axis from the middle of the AB segment is r middle of the AB segment is r where r depends solely on the rotation angle alpha, this is my rotation angle alpha, and the distance d between A and B.

So, d equals to A and B, I am writing this way. Actually, it is the distance between A and B or you can write it as say AB vector kind of thing. Now, we can write r equals to d by 2 cot alpha by 2. So, d by 2 is half of this one, so one part up to this circle and cot alpha by 2 cot is inverse of 10. Alpha by 2 so, half of this one and that way you are getting what is the r. An axis passes through the middle point between A and B. So, between A and B since n equals to 2, so we are considering only Cn where n equals to 2 which means the dimeric situation I mean the C2 symmetry means r equals to 0.

Let C by a circle of radius r that lies on the plane orthogonal to AB and center at the middle of the AB segment. So, in this case, let us make it a vector AB and centered at the middle of the AB segment. Any tangent l to the circle C in its plane is a potential symmetry axis of a transformer sum that rotates A to B. So, that is basically the hypothesis.

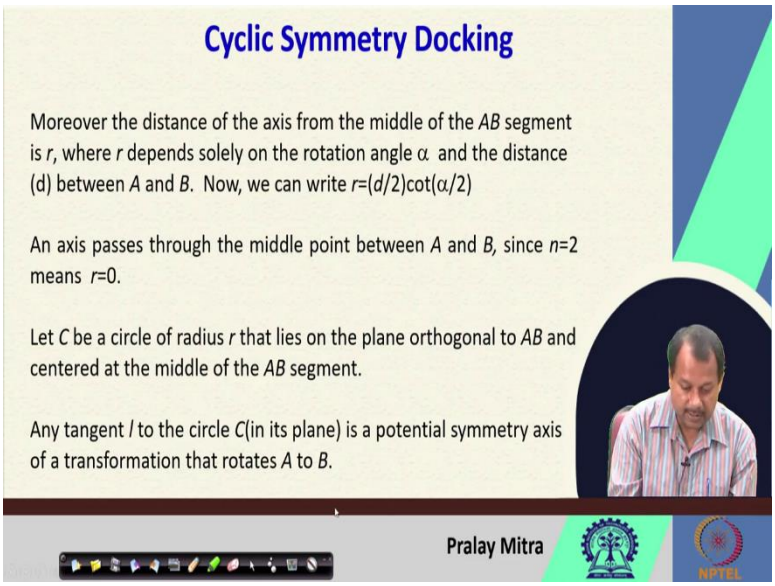
Based upon this, it has been built, and you can create your own say symmetry. So, it has talked about n equals to 2, but you can go for n equals to 3,4,5 and you can create that one so that, you can give that rotation and you will generate basically the symmetry, which are most likely corresponding to this. So, this is the symmetry information about the cyclic symmetry that you are talking about.

(Refer Slide Time: 27:35)



Now, if I pose you one open problem that it is talking about C2, C3, C4 say up to whatever C you wish to mention, but it should be that 360 by that some value say a and n. Now, since n is going to be the integer so you have to pick a accordingly. So, you have to pick a accordingly so, that n is going to be the integer. Now, the question is it possible that similar to that D4, D2, D3 that symmetry you can also create. The question is, if yes, then how? If no, why? So, it is an open question. So, you can think about it.

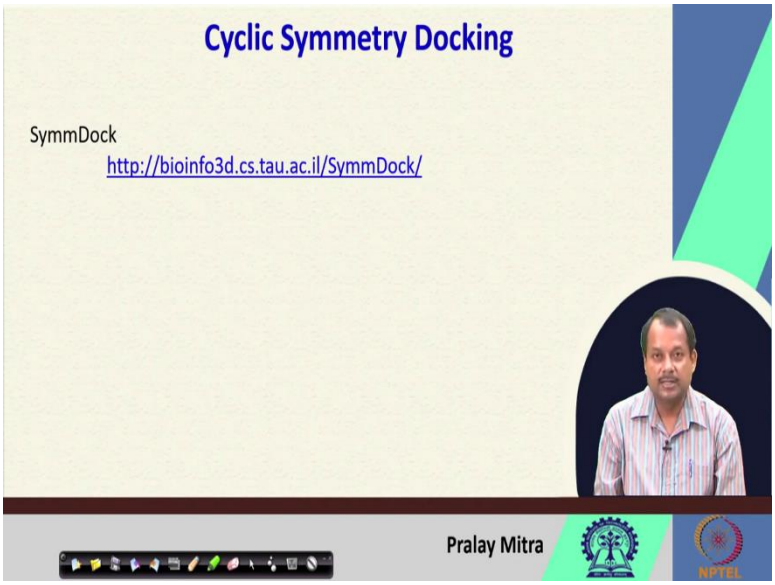
(Refer Slide Time: 28:28)



So, that way we discussed about two specific application of symmetry in protein. One is when, we are inferring the assembly from the crystal structure, and another way when we are basically working with protein docking. Then instead of generating all the possibilities since

since we know that, symmetry will exist, then we will generate only symmetry-related information. And then from there the standard way of doing say scoring, filtering, ranking, which is required for the protein docking that will follow. So, that is it.

(Refer Slide Time: 29:05)



This is the website I mentioned that, SymmDock in the Tel Aviv University, they provide the support, and you can use that SymmDock where you have to provide that symmetry information. What will be the value of N? So, for C2 it is 2 for C 3 it is 3, for C4 it is 4 like that way you can go up to 100 you can go I believe, and you will upload one structure then they will give you the complex. Thank you very much.