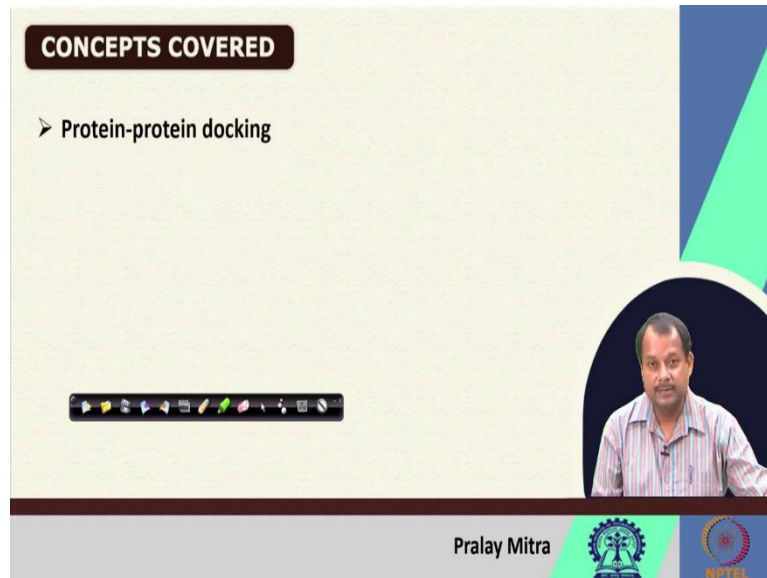


**Algorithms For Protein Modelling and Engineering**  
**Professor: Pralay Mitra**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology Kharagpur**  
**Lecture 40**  
**Some Protein Docking Methods (Contd.)**

(Refer Slide Time: 00:24)



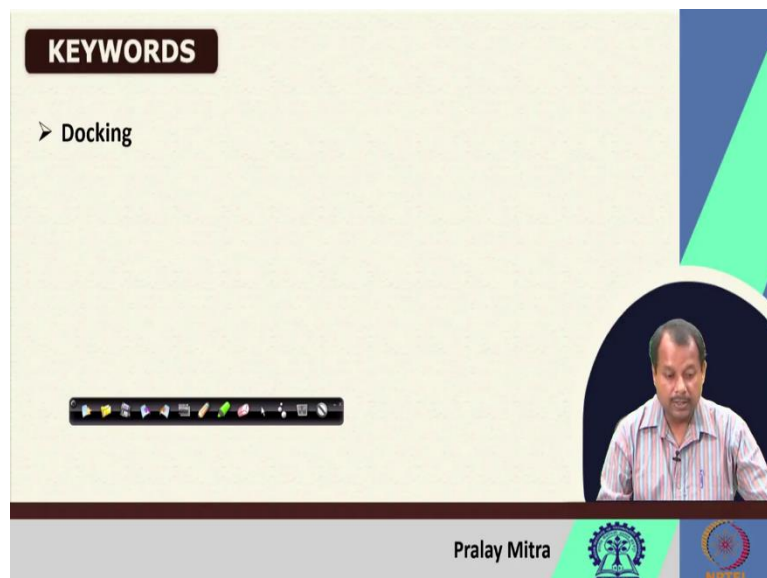
**CONCEPTS COVERED**

- Protein-protein docking

Pralay Mitra

NPTEL

The slide features a light beige background with a dark blue and green geometric design on the right side. A circular inset in the bottom right shows a man in a striped shirt. A navigation bar with various icons is located in the middle left. The footer contains the name 'Pralay Mitra' and the NPTEL logo.



**KEYWORDS**

- Docking

Pralay Mitra

NPTEL

This slide has the same layout as the previous one, with a light beige background and a dark blue and green geometric design on the right. A circular inset in the bottom right shows the same man. A navigation bar with various icons is located in the middle left. The footer contains the name 'Pralay Mitra' and the NPTEL logo.

Welcome back. So, we are continuing some discussion with the protein docking methods. So, we started to discuss one technique, and in that technique, we mentioned about the we mentioned about the bound and unbound datasets. So, this is a protein-protein docking concept I will be continuing here also.



Now, if we analyze that particular protein structure at the interface level, and we categorized the residues into four classes charged, aromatic, hydrophobic and polar, now, the color code indicates that those two are not matching. So, here, you can see that if I pick blue color, then the rate is here and it is going to the green. So, this matching is not correct, and also, green is going to the red. So, two and hydrophobic little matches there but not much.

Mostly it is not charged interaction, it is not aromatic clusters, it is not hydrophobic interaction, it is not even the introduction of the polar residues. On the other hand, using the protein docking technique, if you get this solution the last protein docking technique reported one structure something like this. And in this structure what you see that there is a correspondence of yellow, yellow, I mean, hydrophobic and hydrophobic, red and partially to red not completely, and to red that is the charged, charged and then yellow and yellow, and to some extent green and green which means aromatic and aromatic.

At this point you may argue that okay when it is charged-charged at interaction, so charge this may be acidic or basic. I mean glutamic acid, aspartic acid or lysine and arginine. Now, how do I make sure that it is not going to be the acid-acid interaction. Because you know that acid-acid interaction is not a stable one, it will repulse. Only the acid and base interaction will interact. So that is the charge-charge interaction.

How do I know that this interaction is going to be the acid and base interaction, not acid-acid or charged-charged interaction? For that you have to go back and check the equation for the score function. In that case, the columbic potential, the contribution of the columbic potential that is  $Q_i Q_j$ , divided by  $4 \pi \epsilon$  then  $R_{ij}$  that we computed. Now, these are the partial charges.

If it is the partial charges of the acid-acid or base-base then the same charges will give, will actually give some repulsive forces because of that wall, it will incur some penalty in that equation of the non-bonded energy. So, that is why, it is not actually the charged-charged interaction of that same acid or acid-acid interaction or not base-base interaction it is the interaction between acid and charge. So, this is one example prediction. And when you are doing the docking it is not only that you just do the docking do the ranking and check and lip, but sometimes you have to do the analysis also. Because what is the biological importance or what is the biological relevance of your work that is of interest to everybody.

(Refer Slide Time: 06:07)

**ZRANK**

$$\text{Score} = w_{vdW\_a} E_{vdW\_a} + w_{vdW\_r} E_{vdW\_r} + w_{elec\_sra} E_{elec\_sra} + w_{elec\_srr} E_{elec\_srr} + w_{elec\_lra} E_{elec\_lra} + w_{elec\_lrr} E_{elec\_lrr} + w_{ds} E_{ds}$$

$E_{vdW}(i,j) = \epsilon_{ij} \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{12} - 2 \left( \frac{\sigma_{ij}}{r_{ij}} \right)^6$  van der Waal interaction  
 $E_{elec}(i,j) = 332 \frac{q_i q_j}{r_{ij}^2}$  Electrostatic Interaction  
 $E_{ds}(i,j) = a_{ij}$  Desolvation energy

Pralay Mitra

Here is one of the most popular protein docking technique, which is the ZRANK. So, it uses the same technique that we have discussed for generating the protein docking. So, I am not going to discuss that generation part. Now, after the genetic in the scoring function, so, they use some sort of linear combination of a number of parameters that we are going to discuss.

So here vdW, vdW, vdW a and vdW r, so side chain and the main chain electrostatic interactions, then sorry. So, actually, it will be something, this is one part, this is the side chain and the main chain, then electrostatic then side chain and main chain then electrostatic lra, then lrr and some disulfide interactions. So, this is recently it has updated, but this is one of the primary scoring function for that ZRANK.

Now, this score function they have derived empirically. So, this you can consider as an empirical score function. So, they have defined these parameters so, vdW, ele, electrostatic and ds, so they have defined this desolvation energy, electrostatic interaction and van der wall interaction they have defined. I will come to that one.

But after defining that one what they have done for a known data set where the dimers are known. Say maybe, it has been taken from the protein databank the structure has been taken from the protein databank, and then, consulting the PiQSi or say PISA or some literature evidences they identified and they make sure that it is actually indeed the correct biological association.

After having that one they computed all those features. After computing that one what will be the score function and accordingly these values has been fixed or learnt. So, this is easy you

can do that one, and then finally, they have, they did one linear combination of this individual score function to have one final score function. So, this final score function they have used.

After that one this technique is same as the previous technique that we have discussed on the last lecture that you rank them, and basically, you sort them based upon the score function and you have the rank so you pick based upon one each rank. Now, regarding the individual score functions, so, they have used van der wall interaction, it is same as used the last time. So, this is also called as Lenard Jones potential or say 612 potential because of this 6 and 12 here, so, it is  $r$  to the power 6 and  $r$  to the power 12 here, the  $\sigma_{ij}$  to the power 12 or  $\sigma_{ij}$  to the power 6 is the constant most of the time.

So, this constant and actually when it will be multiplied with  $\epsilon_{ij}$  this will come here, so that value you have to get. So, last time it was said that, you can determine that value from the parameter file of any molecular dynamic software program, but separately also empirically basically you can also determine this one. It is possible. Next, in the electrostatic part it is the columbic interaction  $q_{ij}$  divided by  $r_{ij}^2$  with some constant factor multiplied and this is the desolvation energy which is actually their component of the main score function. Rest of the parts are same.

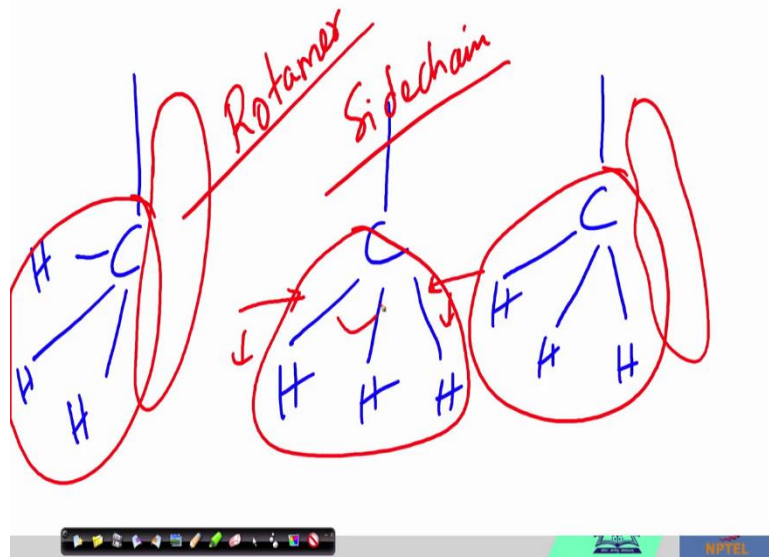
(Refer Slide Time: 10:00)

**PatchDock and FireDock** 725

- **PatchDock:** Molecular Docking Algorithm Based On Shape Complementarity Principles
- **FireDock:** Includes three main steps:
  - (1) Side-chain optimization: The side-chain flexibility of the receptor and the ligand is modeled by a rotamer library. The optimal combination of rotamers for the interface residues is found by solving an integer LP problem.
  - (2) Rigid-body minimization: This minimization stage is performed by a MC technique that attempts to optimize an approximate binding energy by refining the orientation of the ligand structure.
  - (3) Scoring and ranking: This final ranking stage attempts to identify the near-native refined solutions. The ranking is performed according to a binding energy function that includes a variety of energy terms: desolvation energy, van der Waals interactions, partial electrostatics, hydrogen and disulfide bonds,  $\pi$ -stacking and aliphatic interactions, rotamer's probabilities and more.

Handwritten notes: *large protein molecule* (with arrow pointing to 'receptor'), *smaller protein molecule* (with arrow pointing to 'ligand').

Pralay Mitra



Now, apart from the ZRANK, two other software's exist one is the PatchDock and other is the FireDock. So, nowadays people are mostly using this FireDock technique. Because you can consider that FireDock has extended feature with respect to the PatchDock. PatchDock is the basic one which is developed mostly based upon the geometric hashing technique.

So, both are available with Tel Aviv University and the software's web servers are available. You can upload your protein structure and get the prediction. So, what is the advantage of the FireDock, is that, as of now, when we are discussing say protein interaction and then we are considering the bound and unbound cases then you remember that I mentioned about the fitness correct fitness at the interface, and which is there in case of bound complex, but which is absent in case of unbound complex.

Why it is absent? As I mentioned that, if I allow this to form a complex and then crystallize then they are fitting specifically on the say side chain atom will be something. I am not telling that will be perfect, it will be something and which will be different if I allow them to crystallize separately. So, if it will crystallize separately it will crystallize separately then there is no guarantee that the gap between two fingers will be same, if they crystallize together or if they co-crystallize that is true, that you can understand.

Now, if they crystallize separately and you are going for the docking that is the more realistic situation which is called as the unbound docking then at the interface when say mostly you identified that interface mostly like this then one situation maybe so, after identify probable interfaces not the correct interfaces you allow the side chains to say move it little bit to rotate a little bit.

When you will allow them to move or rotate a little bit then you will find what is the best fit because of that movement. That moment is triggered by one situation, which is that if say the simplest one, say, I am considering that alanine it is CH<sub>3</sub>. Now, this can be one CH<sub>3</sub> this can be one, this is simplest one, please note it down, because a lot of variations will be there.

So, you see that in this case, this side is completely empty. So, this part is completely empty, in this case. So as such, that emptiness is not observed under this line. In this case also it is mostly empty. So, this is one orientation, this is one orientation, this is another orientation. So, these three are three different orientations, and I am calling them as rotamer.

So, this rotamer is in the context of the sidechain, which says that, how many orientations of the side chains are allowed. So, that is called as a rotamer. So, there are several software's which after identifying the protein folds can fit the correct rotamer so that the structure is optimized. And also, if you do a quick search in the protein databank to check different amino acids and then for their side chain. If you compute that what is the say, what is the, if you compute that what is the angle bond angle, what is the torsal angle etc, at the side chain, you will find a variety of presence of the side chains. So, each of them can be considered as one rotamer.

So, this FireDock actually makes use of that concept the rotamer. And it actually use that concept and allows some to rotate and try to make a best fit from that. So, this PatchDock is a molecular docking algorithm based upon safe complementarity principle and FireDock includes three main steps features. So, what are those features? Let us look, sidechain optimizer song.

So, the sidechain flexibility of the receptor and the ligand. So, here receptor and ligand indicates the large protein and ligand indicates smaller, both are protein. Do not confuse that ligand means it is a small one. Both are protein of size greater than 25, of course, but when, two protein molecules are docking with each other and it is not homomer then one will be larger another will be smaller.

You remember that concept, I mentioned, that if one is larger another the smaller it is a good idea that if you allow the smaller to translate, rotate or transform and keep the larger one as rigid or consider that as a receptor, so that, your computation time will be minimized. So, that is why it is receptor and ligand is modeled by a rotamer library. So, what is that rotamer library? Just now I mentioned.

So, corresponding to each amino acid what are the possible say orientation of the side chains considering variety of angle and so, angles and the torsional angles. Because the bond length will be same, bond angle and the torsional angle actually. So, you will have several possibilities. Now, from those several possibilities you have to pick one and fit that one. So, that is my side chain optimization. And that way when say for bound complex because of the co-crystallization of the two protein structures, as A and B it is a best or perfect matching, but when they will be separately crystallized, then these kind of spaces may not be present.

So, it can be like this, like this, like this, like this, several possibilities may happen. Now, when I am doing the docking then after reaching up to this position, so, what I can do so, if I give you the projection. So, I can change this one and for each change, I can see that what is the best fitting, best fitting in terms of some score function. So, that is what is the side chain optimization, that I will do.

The optimal combination or rotamers for the interface residues is found by solving an integer linear programming problem. So, this LP is a linear programming problem. Now, the rigid body minimization, this minimization stage is passed on by a Monte Carlo MC technique that attempts to optimize an approximate binding energy by refining the orientation of the ligand structure. And the third stage is the scoring and ranking. This final ranking stage attempts to identify the near-native refined solutions.

The ranking is performed according to a binding energy function that includes a variety of energy terms like desolvation energy, van der Waals interaction, partial electrostatic energy, hydrogen bonding, disulfide bonding, pi stacking, this pi stacking indicates pi stacking, and aliphatic interacts, rotamers probabilities and more.

So, you see that a number of features are being utilized for protein-protein docking, but at the core, of course, you have to incorporate van der Waals interaction, some electrostatic interactions and this hydrogen bond may be separately considered or maybe included in the electrostatic part. Disulfide bond you may consider, but the situation is, as I mentioned earlier also, that the disulfide bond will occur between the cysteine residues.

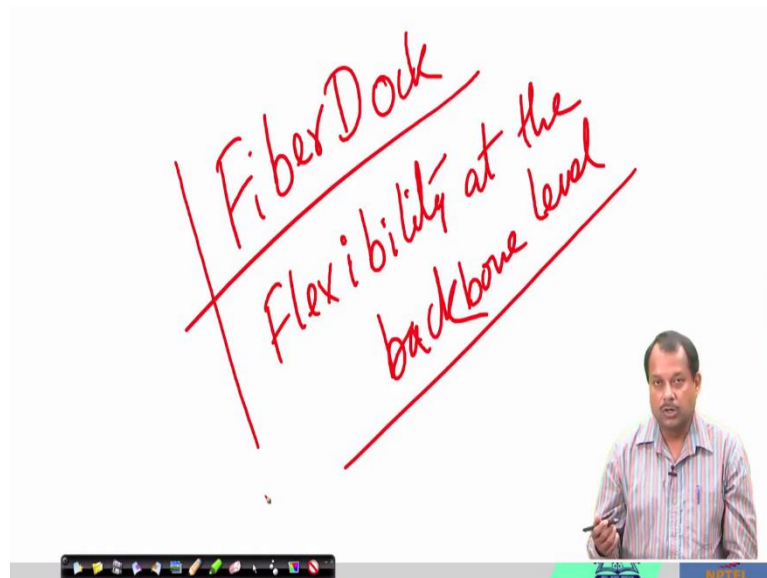
I mean, the side chain of the cysteine residues contains one sulfur. Between two sulfites when two cysteine molecules two cysteine residues are close enough, so between two sulfur molecule sorry two sulfur atom basically one disulfide bond will be formed. So, it is indeed the disulfide bond is a very strong one. And if there is any disulfide bond, it mostly



determines or dictates some structural changes or structural stability, but at the same time, the occurrences of the cysteine is not frequent in the protein molecule.

That is why co-occurrences have two cysteine so that they will form some disulfide bond is also not very much frequent. So, separately you can keep that one or you may not keep that one. I mean, in your score function you can include the disulfide bond or you can have one checking that if there is any disulfide bond then you give some priority and use that information, if not, then in you use some score function which does not include any disulfide bonding. Apart from that pi stacking, aliphatic interaction, so, then rotamer probabilities and more will also be considered as part of the FireDock.

(Refer Slide Time: 20:25)



Now, recently they also published apart from this FireDock, one more paper or algorithm that is called as the FiberDock. This includes the flexibility at the backbone level. As of now, except this FiberDock all the algorithms we have discussed they do not consider flexibility at the backbone level. Even when I considered the unbound docking, they considered that okay the side chain will be flexible and the interaction of the sidechain will be there. So, during that interaction, so, I have to adjust some of the sidechain orientations, but basic fold or the backbone will be same.

But if it is required that you need to adjust the backbone also then that particular topic is taking more and more complex because changing the backbone means changing a lot of things. So, incorporating flexibility in the backbone. But they did the FiberDock well, the success of this FiberDock is debated. And also, few people are trying so, here, I am giving you another open idea that few people are trying that whether molecular dynamics

simulations can be exploited to go for the backbone for flexible backbone protein-protein docking or not. So, people are also working on exploring that concept.

(Refer Slide Time: 22:16)

**Docking measures**

- **LRMSD** ligand RMSD ✓
- **iRMSD**: interface Root Mean Square Deviation (iRMSD)
- **Fnat**: the number of correct residue-residue contacts in the docked prediction divided by the number of contacts in the target complex.

*Handwritten notes:*  
 RMSD (B)  $\rightarrow$  T & Mat(B)  
 align(A<sub>r</sub>, A<sub>b</sub>)  $\rightarrow$  T & Mat

Pralay Mitra

Well, we discussed about few docking algorithms, but what is the measure to know the goodness of your docking algorithms. So, RMSD is one that we understand. Now, in the RMSD also there are two category, in the context of protein docking. One is called as the LRMS T or ligand RMSD, another is called as the iRMS T or Interface Root Mean Squared Deviation or Interface RMSD.

So, what is the basic difference between these two? So, when I am considering say, computing the RMSD. So, given two structures so, this is say one doc structure and say, please assume both are same actually with two different orientations. In this case you can consider this just a translation.

So, this is A, B and this is A, B. Now, in the context of protein-protein docking, what we need to do in this case A and B both are same, as per the sequence as well as the structure. So, only different orientations are coming, and we need to capture that orientation. So, for that what we need to do, first, we need to align red A with blue A. Now, so, red A red will be aligned with A blue. Of course, it will be perfect alignment if it is a bound complex because AB has been extracted so AE aligning with A, so it is not a problem, but in case of unbound complex, so there may be some mismatch.

And if say at the sequence level there are some changes you know how to fix that one. You go for sequence alignment, from there you identify the correspondence and align AR and AB.

After this alignment, you will get transformers and metrics. Now, you apply this transformation metrics on blue B. So, this is basically blue, and this is also blue. So, on blue B, you apply this rotation. So, once you will apply then if both the complexes AB red and AB blue are same, then they will align perfectly, so RMSD is going to be 0 between this one and actual B.

So, if I compute the RMSD then it is going to be 0. If not then I will have one RMSD value report that RMSD value. That is going to be your LRMSD. Of course, this L indicates that A is going to be your receptor and B is going to be your ligand. So, receptor means, the larger partner in your docking. If A and B are homomer, then actually no problem, but if it is not then the larger partner is going to be your A, and smaller partner is going to be your B. So, LRMSD is clear now, I believe?

(Refer Slide Time: 26:03)

**Docking measures**

- **LRMSD**: ligand RMSD
- **iRMSD**: interface Root Mean Square Deviation (iRMSD)
- **Fnat**: the number of correct residue-residue contacts in the docked prediction divided by the number of contacts in the target complex.

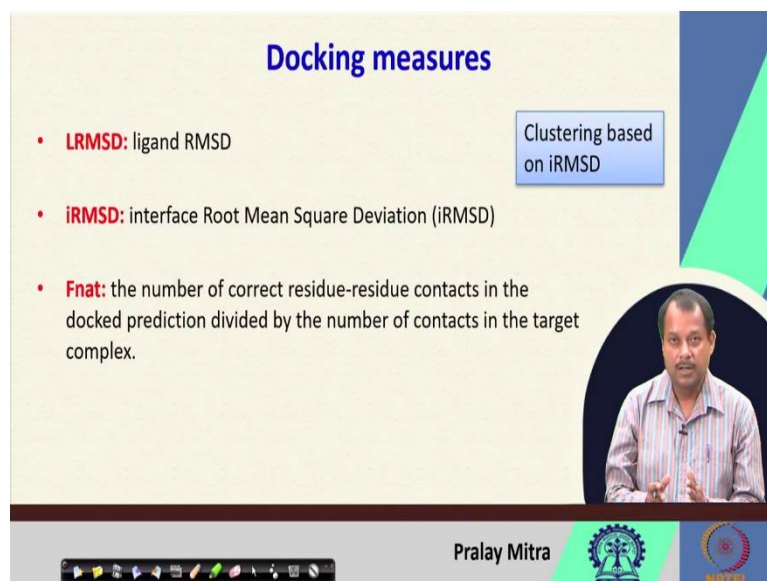
Pralay Mitra

IIT Bombay NPTEL

## Docking measures

- **LRMSD**: ligand RMSD
- **iRMSD**: interface Root Mean Square Deviation (iRMSD)
- **Fnat**: the number of correct residue-residue contacts in the docked prediction divided by the number of contacts in the target complex.

Clustering based on iRMSD



Now, if I take a subset of what So, I identified the interface. Now, instead of the whole protein complex, if I compute the same RMSD, the same technique that I just mentioned to you for only the interface residues then what I will get is called as the IRMSD or interface RMSD. Now, you see, because of the computation of the RMSD, LRMSD maybe, little high, but it may possible that IRMSD is very small. Because interface is aligned properly rest of the part is not much aligned, it can be opposite also. So, overall alignment is done but at the interface level when I see that they are alignment is not good.

So, in order to have a correct docking result you should have LRMSD and IRMSD both within some limit. Usually the limit is say for LRMST at most 10 angstrom for IMRSD at most 5 angstrom, better if LRMSD is within 5 angstrom and IRMSD is within 2 angstrom or 1 angstrom.

Fnat is another measure. The number of correct residue-residue contacts in the dock itself divided by the number of contacts in the target complex. So, that way you can ensure that not only at global orientation level, but at the finer or interface level also the alignment I mean the orientation is correct. So, that information you can gather. So, these three features or sorry the measures are actually considered by the world community. And if you are developing some docking technique then you have to also analyze based upon these three features.

Sometimes the clustering is done based upon this IRMSD. So, you heard one clustering on the last lecture that when we have a lot of decoy complexes instead of treating them separately you cluster in order to reduce the number and that clustering you can done at the IRMSD level.

(Refer Slide Time: 28:23)

Predictor	Affiliation	Software	Algorithm
Abagyan	Scripps	ICM	Force Field
Camacho/Vajda	Boston	CHARMM	Force Field Refinement
Gardiner	Sheffield	GAPDOCK	Shape+Area GA
Sternberg/Smith	Imperial	FTDOCK	FFT
Bates/Fitzjohn	ICRF	Guided Docking	Force Field
Ten Eyck/Mitchell	SDSC	DOT	FFT
Vakser/Tovchigrechko	SUNY/MUSC	GRAMM	FFT
Olson	Scripps	Harmony	Spherical Harmonics
Weng/Chen	Boston	ZDOCK	FFT
Eisenstein	Weizmann	MolFit	FFT
Wolfson/Nussinov	Tel Aviv	BUDDA/PPD/FireDock	Geometric Hashing
Iwadata	Kitasato	TSCF	Force Field+Solvent
Ritchie/Mustard	Aberdeen	Hex	Spherical Polar Fourier
Palma	Lisbon	BIGGER	Geometric+Electrostatic
Gray/Baker	Washington/IHU	RosettaDock	Monte Carlo+Flexibility
Mitra and Pal	IISc, Bangalore	PROBE/PRUNE	FFT

This context one fact that there is one international body, which is called as a critical assessment or prediction of interaction in short CAPRI. So, the CAPRI actually provides some unbound cases, and they ask to predict and they have some data also. So, using that one you predict and based upon that they release that which method is doing good.

So, several predictor methods. So here the list of the authors and their affiliations, the name of their software, and the algorithm they have used is mentioned. You will see a varied number of algorithms. People are using their applications or also across the globe and the name of the software. Some of them may be known to you if you are using this one. So that is it for this protein-protein docking. So, thank you very much.