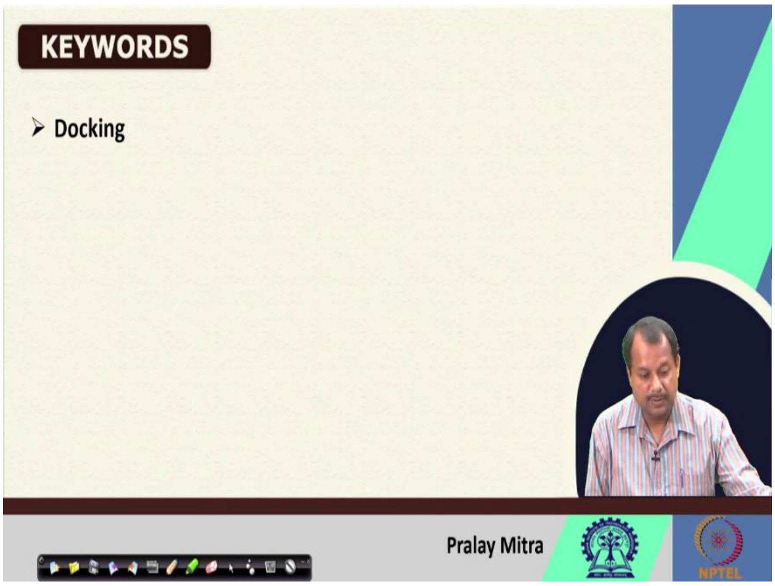
Algorithms For Protein Modelling and Engineering
Professor Pralay Mitra
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur
Lecture 39
Some Protein Docking Methods

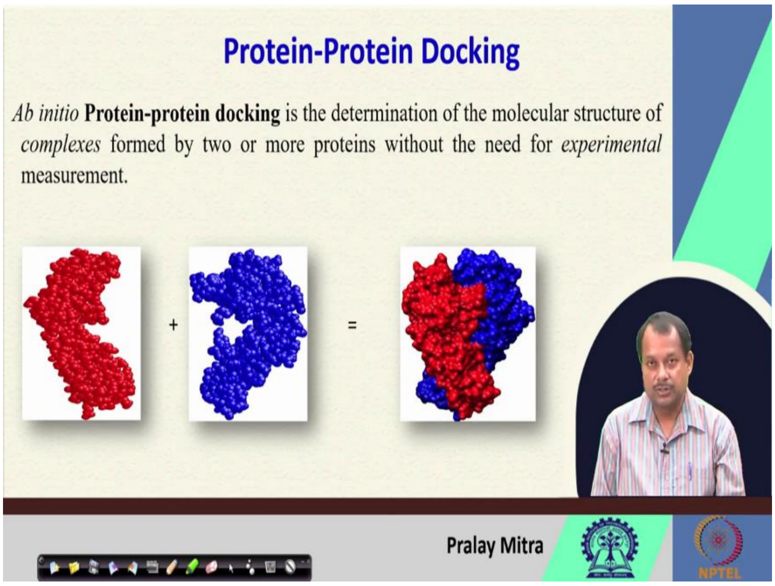
Welcome back. So, let us continue with some our protein docking methods. So, on the last lecture, we extended the concept of symmetry in SymDock, but there are some limitations. So, limitations indicates that when say it is not mostly although we go for the docking of dimers, I mean, given two protein molecule you need to generate the dimers. So, for the dimer if you think about the C2 symmetry then a lot of possibilities out there.

So, in that way, so we do not get much advantage. It will be advantageous when it is homomer and also, we are going for multimeric docking. I mean, definitely, two inputs will be there, but in one side there are not monomer it is dimer or trimer etc, and the side maybe monomer or something. But if it is not then actually you have to go for the traditional docking techniques. So, now we will cover quickly, two other docking techniques. And also wrap up with several docking techniques what are their pros and cons in some tabular method. So, let us start so, some docking techniques.

(Refer Slide Time: 01:24)


KEYWORDS

- Docking



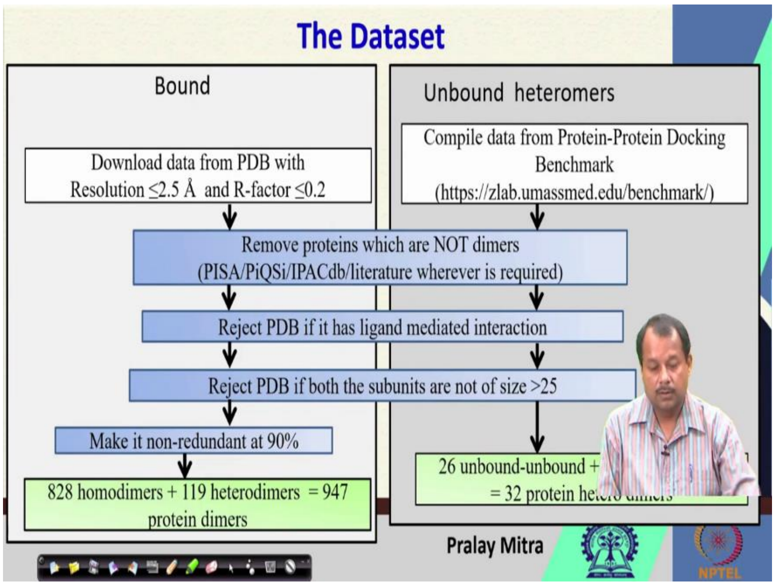
Protein-Protein Docking

Ab initio **Protein-protein docking** is the determination of the molecular structure of *complexes* formed by two or more proteins without the need for *experimental* measurement.

So, the concept will be covered as Protein-Protein docking so keyword is also just a docking, I mentioned this one. Now, start with the definition again. So, what is the protein-protein docking and ab initio protein-protein docking is the determination of the molecular structure of complexes formed by two or more proteins without the need for experimental measurements.

So, experiments are not required given two protein molecules something like red and blue and then you have to generate the red, blue structure without the need for experimental verification or the measurement. That is the definition of that protein-protein docking.

(Refer Slide Time: 02:07)



The Dataset

Bound

Download data from PDB with Resolution ≤2.5 Å and R-factor ≤0.2

Unbound heteromers

Compile data from Protein-Protein Docking Benchmark (https://zlab.umassmed.edu/benchmark/)

Remove proteins which are NOT dimers (PISA/PiQSi/IPACdb/literature wherever is required)

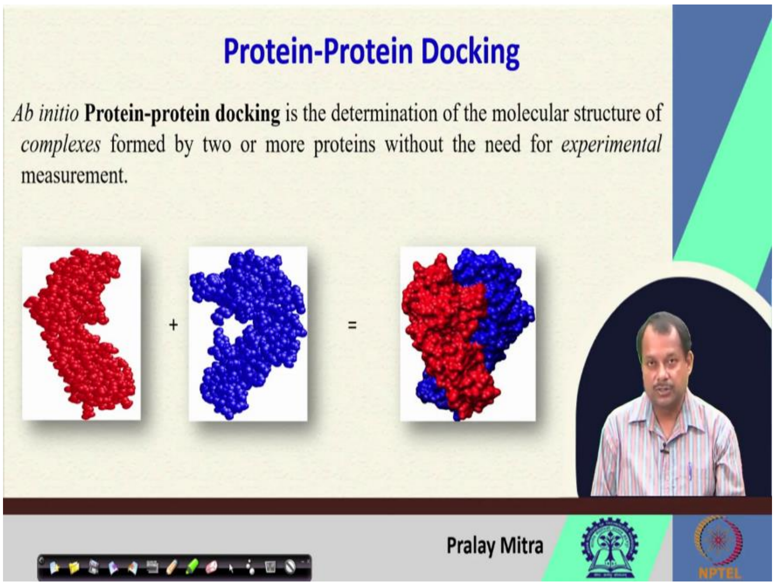
Reject PDB if it has ligand mediated interaction

Reject PDB if both the subunits are not of size >25

Make it non-redundant at 90%

828 homodimers + 119 heterodimers = 947 protein dimers

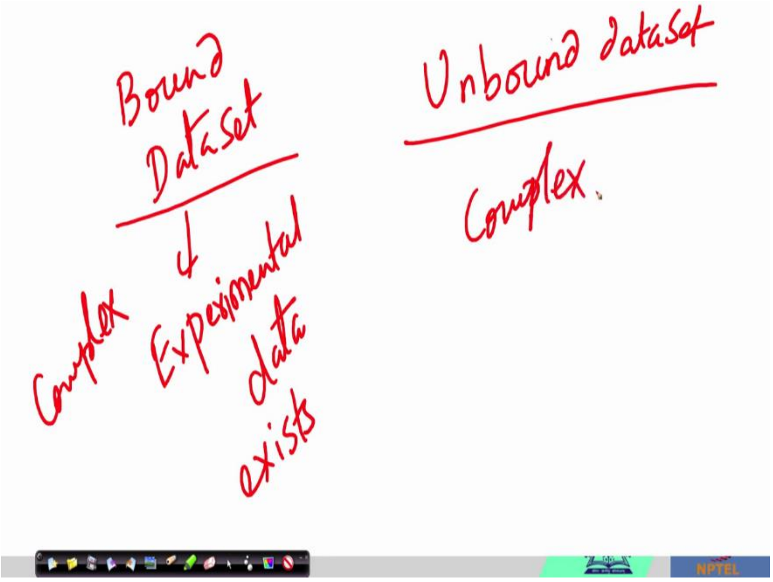
26 unbound-unbound + = 32 protein hetero dimers



Protein-Protein Docking

Ab initio **Protein-protein docking** is the determination of the molecular structure of *complexes* formed by two or more proteins without the need for *experimental* measurement.

Pralay Mitra

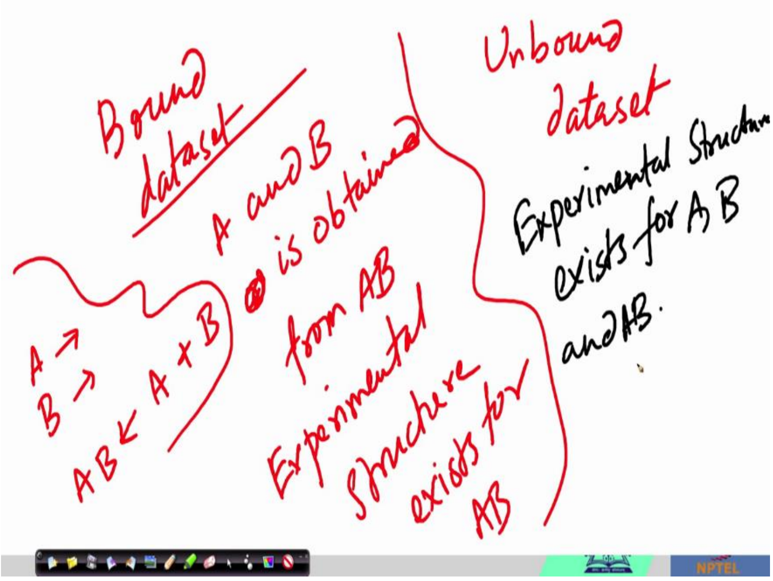


Bound Dataset

Complex ↓ Experimental data exists

Unbound dataset

Complex.



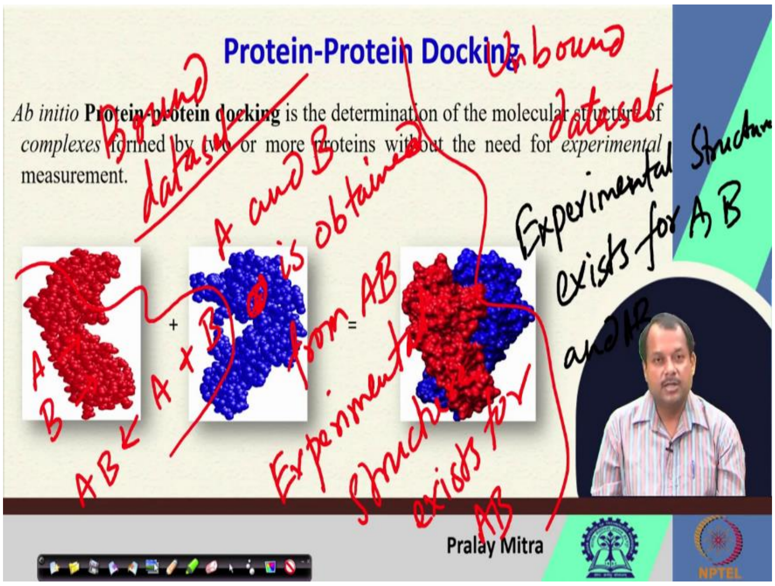
Bound dataset

A →
B →
AB ← A + B

A and B is obtained from AB
Experimental Structure exists for AB

Unbound dataset

Experimental Structure exists for A, B and AB.



Protein-Protein Docking

Ab initio **Protein-protein docking** is the determination of the molecular structure of *complexes* formed by two or more proteins without the need for *experimental* measurement.

Bound dataset
A and B is obtained from AB
Experimental structure exists for AB
Unbound dataset
Experimental Structure exists for A B and AB
A
B
AB ← A × B
+
=
Pralay Mitra
NPTEL

When we are doing the protein-protein docking again from the structure also I can explain that grossly there are two different kinds of docking techniques, not in terms of the way it is doing in terms of the data set they are using. So, one is called as the bound data set and another is called as the unbound data set. So, what is the difference between these two, bound and unbound data set?

Bound bounded asset indicates that it comes from the experimental data or the complex experimental data exists. In this case for unbounded data set or I should give not like this complex experimental data exists for bound as well as for the unbound. So, let me phrase it different way. So, for the docking I need three proteins actually. One is say A, B and AB which I will get when A and B will dock. So, if I define this way then for the bound data set A and B is obtained from AB, experimental structure exists for AB. For unbound data set, experimental structure exists for A B and AB. Now, what will be the difference? You try to understand.

One situation this is the complex, this complex is given to you. Now, what you do that, this is A this is B it is a dimer. So, you take this out you take this out so A and B. Now, you give some random or arbitrary transformation so syou try to understand what I am trying to say. So, this these two protein structures are given to you.

Now, one situation is that, you start with a complex. So, one complex is your input. Now, when the complex is your input then you take this out, you take this out so, then these two will be your new input. What I am trying to say that when one complex from one complex you extract A and B two subunits and then you arbitrarily you provide some arbitrary transformation between A and B after that one if you allow them to dock then it is perhaps easy to identify AB.

Why? Because its structure is something like this, its structure is something like this. And when I take it out then the best fitting will be between this. Best fitting will be when they will be like this. So, it is easy to find the compactness, complementarity and the score if during the generation phase if I generate this particular orientation. If I do not generate then that is a different issue, but if I generate then I will get the best result. So, that is the bound.

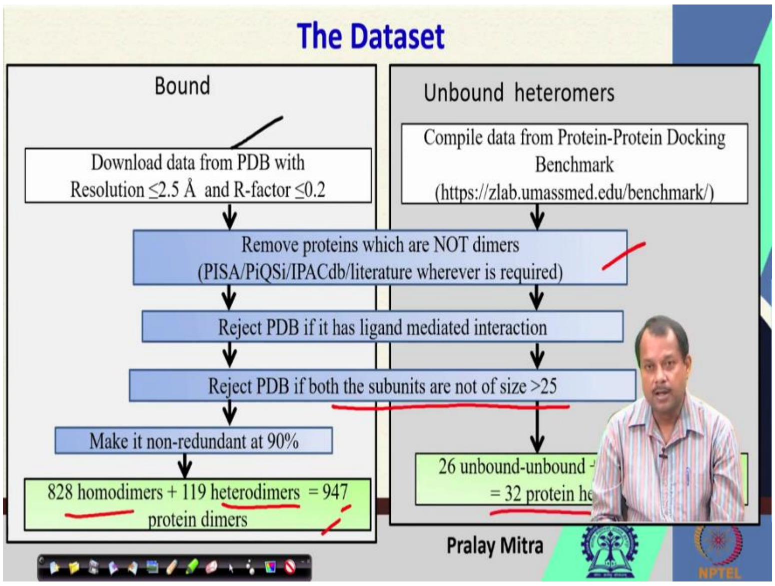
When, I have the AB or the complex as an experimental structure or I can consider say PDB in PDB that diametric complex exists what I am trying to do, I separate out that dimeric complex A and B and then give some arbitrary rotation and translation or transformation into A or B or on both. Then I allow them to dock, I mean, then docking will take A and B as an input and will generate AB, but AB also exist, that is called as a bound docking. When the dock structure exist and the docking partners I am generating from the dock structure.

In case of unbound docking A is separately crystallized B separately crystallized AB is separately crystallized. Now, when A and B are separately crystallized then they will optimize their structure during their crystallization. So, even these two structures are there because of that crystallization separate crystallization, it may be not with this kind of shape, which is going to match. It may be, the fingers will be like this then matching may be different.

Because initially if the fingers was inside then only there is a possibility that in between two fingers there will be some gap, but if it is not there then there is no restriction as if during the crystallization it can close down or it can squeeze a bit so that new finger do not go or enter there. So, unbound docking is going to be challenging. But unbound docking is also going to be that reality, when A and B are two complexes two protein molecules separately obtained and you are given, it is given to you, and you need to generate a complex out of that one, but bound docking is not reality.

Because in bound docking you know the complex structure, and you from that complex structure you are generating A and B. Whereas, for the unbound A and B are separately crystallized. However, you have to do docking on both, bound and unbound because bound is easy and it will also tell how good is your algorithm or your method then you test it on the unbound cases.

(Refer Slide Time: 09:45)



The Dataset

Bound

Download data from PDB with Resolution ≤2.5 Å and R-factor ≤0.2

Unbound heteromers

Compile data from Protein-Protein Docking Benchmark (https://zlab.umassmed.edu/benchmark/)

Remove proteins which are NOT dimers (PISA/PiQSi/IPACdb/literature wherever is required)

Reject PDB if it has ligand mediated interaction

Reject PDB if both the subunits are not of size >25

Make it non-redundant at 90%

828 homodimers + 119 heterodimers = 947 protein dimers

26 unbound-unbound -
= 32 protein he

Pralay Mitra

So, we are that is why differentiating between two different data sets bound as well as unbound. Now, in case of bound, so as you see that on the left-hand side, download data from PDB with resolution less than 2.5 angstrom and R factor 0.2. So, this thing you can mention and what is the advantage of this we have discussed in detail. Then for unbound compile data from protein-protein docking benchmark it is available with Zipping Winks lab as they are which is called as the Zed Lab, this is also the house of Zed doc and Zed Rank software at the University of Massachusetts.

Then you download that one from there. After that few screenings you have to do that, you have to do on both, bound as well as unbound data set. So, first you have to remove proteins which are not dimers, as per say PISA, PiQSi, IPACdb, literature wherever is required because you have to go for experimentally dimer structure.

So, after doing that one reject PDB it has ligand mediated interactions. So, if there is some ligand. Say for example, my two hands are there, so in between if this pen is acting as one pen one pen is acting, then the problem will be when I will take the pain out then there will be some void region there because of that void the packaging the interaction that will not be perfect and that will not be as for the experimental information. So, that is not possible.

And also, when it is mediated means, A is interacting with some ligand, ligand is interacting with B. So, there is no direct interaction between A and B. So, since there is no direct interaction between A and B, so it creates some problem. So, that is why my suggestion is that if there are some mediated interaction like ligand or water molecule etc better remove those cases, because it may bias and you may not get the correct result that may lead to some problem. Then reject PDB if both the subunits are not of size greater than 25. This is required, because during the discussion of the scope, we also discussed that minimum amount of amino acids are required to have one particular fold.

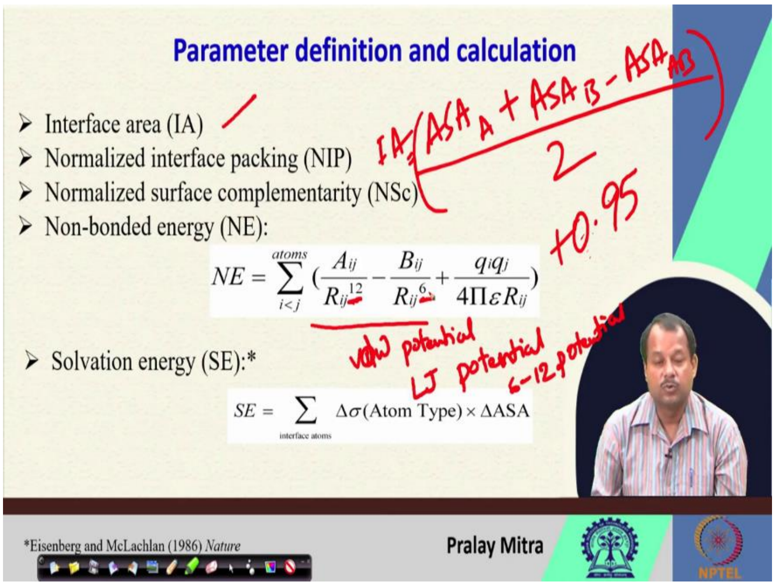
If it will not take a particular fold then it is problem. So, also you can understand say if a small amount say this pen, is a small amount present and there is not such fold or pen may not be a good idea. Say, if a small amount of peptide is present and there is not much fold. If there is not much fold than everything is exposed, which means, I cannot differentiate between the hydrophobicity and the hydrophilicity.

So, the general concepts on is that, and it is true, that protein coat is tend to be more hydrophobic compared to the protein surface. And when I am doing the docking to some extent protein interfaces are little more hydrophobic compared to the protein surface to some extent, not always but sometimes, but those kinds of checking we cannot do if it will not take any particular form.

Similar to this, lot of problems are there. So, it is a good idea to restrict that the size of the protein is going to be more than 25. But then question will come that what if there is one protein and other small molecule or ligand? So, for that we have to come up with some other algorithm which is called as the protein ligand docking or protein small molecule docking. So, prime example is the auto dock technique. That we are not going to discuss right now. So, several variations are there. But if it is protein-protein docking, so special treatment is required. So, we will consider the length of the protein is at least 25.

Then make it non-redundant, since the bound docking is relatively easy, so let us make it non-redundant that kind of checking is not required for the unbound heteromers. That way you will get say, 828 homodimers or 119 heterodimers in total 947 protein dimers that is huge. But considering the situation that bound docking is going to be very easy, so that is not much a problem. Whereas, for the unknown docking, so 32 heterodimers has been compiled, but it will definitely increase now, so that you can check from the recent website, the link I have provided here.

(Refer Slide Time: 14:39)



Parameter definition and calculation

- Interface area (IA)
- Normalized interface packing (NIP)
- Normalized surface complementarity (NSc)
- Non-bonded energy (NE):

$$NE = \sum_{i<j}^{atoms} \left(\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{4\Pi \varepsilon R_{ij}}\right)$$

- Solvation energy (SE):*

$$SE = \sum_{\text{interface atoms}} \Delta\sigma(\text{Atom Type}) \times \Delta ASA$$

$$IA = \left(\frac{ASA_A + ASA_B - ASA_{AB}}{2}\right)$$

+0.95

vdw potential

LJ potential

6-12 potential

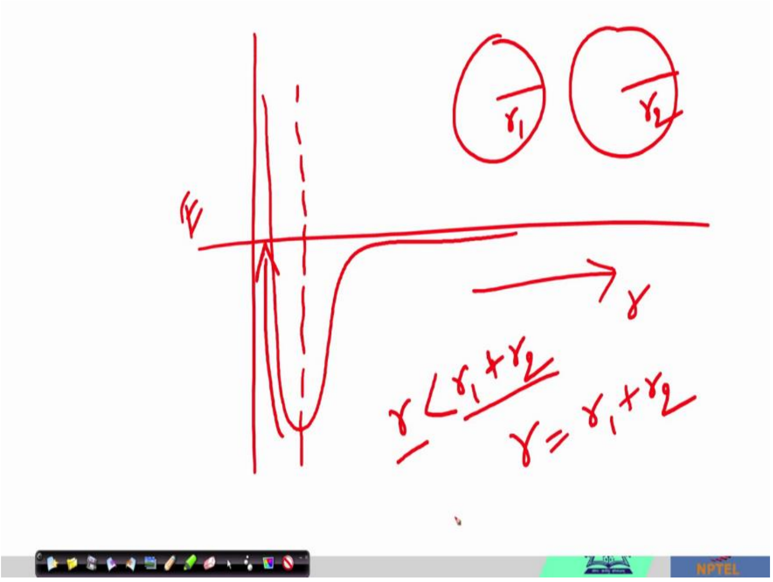
*Eisenberg and McLachlan (1986) *Nature*

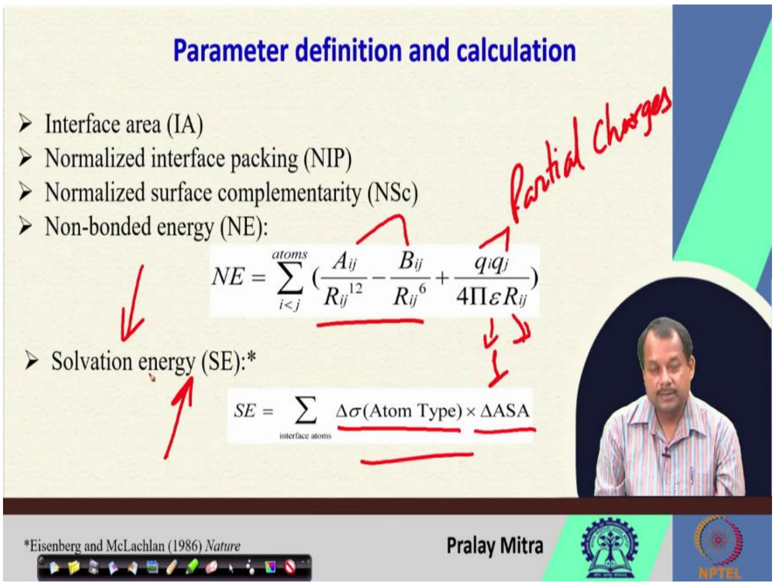
Pralay Mitra

Next you need to define some score function. So, the algorithm we have defined, we have generated a number of decoy complexes now we need to score them. After scoring that then we need to rank. So, for the scoring purposes, we are considering five features here. So, who are they? Interface area. I know how to calculate that one, use N access and for each protein decoys, so you extract A and B separately run N access on A separately run N access on B run N access on A and B. And that way if is ASA of A plus ASA of B minus ASA of AB whole divided by 2 is going to be your IA.

So, that IA is defined like this. Now, extensively we discussed the normalized interface packing and normalized surface complementarity. What is the correlation among those two that also we discussed? So, between NIP and NSc there is 0 point sorry. That is correlation exists. Next is the non-bounded energy NE which consists of two parts. So, this is called as the VDW, Van der Waals potential or it is called as the LJ Lennard Jones potential or this is also called as the 612 potential. Why do it is 612? You see that 12 and power 6 both are present. So, the origin of this forcefield is very simple.

(Refer Slide Time: 16:47)





Parameter definition and calculation

- Interface area (IA)
- Normalized interface packing (NIP)
- Normalized surface complementarity (NSc)
- Non-bonded energy (NE):

$$NE = \sum_{i<j}^{atoms} \left(\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{4\Pi \varepsilon R_{ij}} \right)$$

Partial Charges

- Solvation energy (SE):*

$$SE = \sum_{\text{interface atoms}} \underline{\Delta\sigma(\text{Atom Type})} \times \underline{\Delta \text{ASA}}$$

*Eisenberg and McLachlan (1986) *Nature*

Pralay Mitra

So, it is something like this where this is the potential and this is the distance. So, if you wish to take two atoms close to each other. And if I assume that their radius is say R1 and R2 then when I am computing the R and if R equals to R1 plus R2 then their attraction is most and repulsion is less, that is my R.

But if R is less than R1 plus R2, which means, R is less. Now, so R1 plus R2 indicates when they are touching with each other. When they are touching with each other. Now, if R1 plus R2 is more than R then what will happen if R1 plus R2 is more than R then what will happen, they are trying to penetrate with each other, which is not allowed, that is why, there is a sudden increase in the forcefield and that is a repulsive term R to the power 12. R to the power 6 is the attraction, attractive term.

On the other way if R is greater than R1 plus R2 it is fine so repulsion forces are not there, but attraction will also keep on reducing. That way these are R to the power 12 and R to the power 6, A and B are two constants that you can derive from some so, that is atom-specific and that you can derive from the parameter file of any molecule or dynamic software.

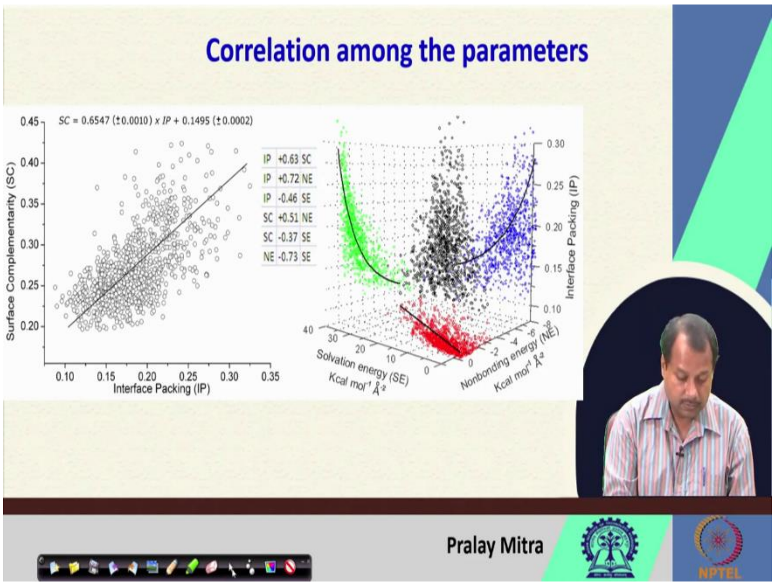
Now, this Qi and Qj are the partial charges. So, this is 4 pi epsilon you can consider as 1, Rij is the actual distance between two atoms that you can compute that will be used here also. These partial charges also you can get from the parameter file of the molecular dynamics simulation.

Finally, the fifth term is the solvation energy SE, and this SE we computed from the Isenberg and McCullough Hans paper a which is a published long back 1986 nature. So, as of now, more than 2000 citations are there because of the technique. This technique we also discussed as a feature when we are discriminating the biological interfaces and the crystal artifacts.

Specifically, there are five atom types. So, carbon is one, charged nitrogen, charged oxygen is another two and a non-charged nitrogen and non-charged oxygen is another. So, internal five categories are there, sulfur will be part of that one or you can keep them separate. So, for that one constant value is there, and multiplied with the amount of loss of accessible surface area that you can consider that will contribute to your solvation energy.

Now, you in order to make a stable protein complex then you need to increase this solvation energy and decrease these non-bonded interactions. So, SE will increase and NE will decrease. Now, this increase and decrease this will be utilized in an efficient way in order to design the score function. How? You will see this.

(Refer Slide Time: 20:29)



Correlation among the parameters

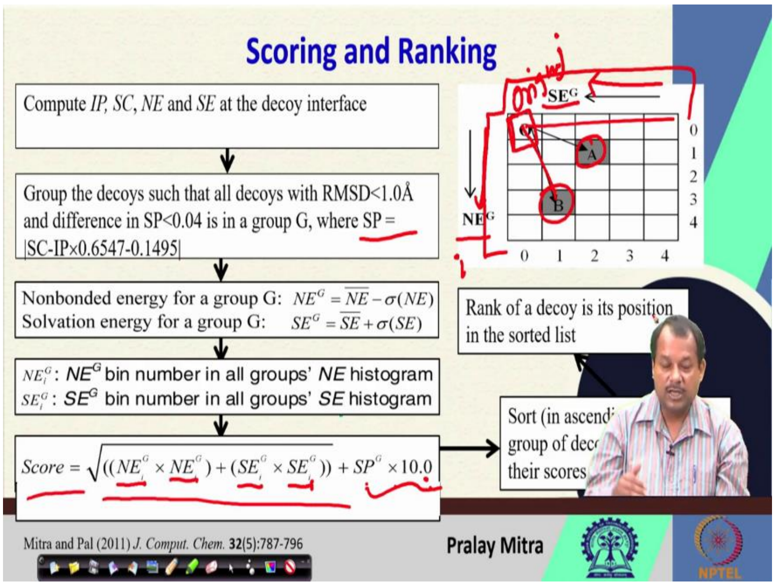
IP	+0.63	SC
IP	+0.72	NE
IP	-0.46	SE
SC	+0.51	NE
SC	-0.37	SE
NE	-0.73	SE

There exist correlation among the features that I computed. So, IP, interface packing, SC, surface complementarity, the definition is known to you from the previous lectures. Then NE is the nonbonded energy that we just demonstrated, SC is the solvation energy that we demonstrated.

Now, if I make a plot on the left-hand side interface packing, and sorry, interface packing and surface complementarity is here. So, it is not NIP and NSc that is why the correlation is little less, it is 0.63 between IP and SC. Now, on the right-hand side solvation, nonbonding energy and interface packing is plotted, and among them what are the correlations that is demonstrated, but not watch correlation.

Only IP and NE is correlated, and then non-bonded energy and solvation energy is negatively correlated. As I mentioned non-bonded energy is going to decrease and solvation energy is going to increase for a stable complex, that is why it is negative that you please note it down. And IP and SC is to some extent positively correlated apart from that one, there is not much correlation among other features.

(Refer Slide Time: 21:55)



Scoring and Ranking

Compute *IP*, *SC*, *NE* and *SE* at the decoy interface

Group the decoys such that all decoys with RMSD<1.0Å and difference in SP<0.04 is in a group G, where SP = |SC-IP×0.6547-0.1495|

Nonbonded energy for a group G: $NE^G = \overline{NE} - \sigma(NE)$
Solvation energy for a group G: $SE^G = \overline{SE} + \sigma(SE)$

NE_i^G: NE^G bin number in all groups' *NE* histogram
SE_i^G: SE^G bin number in all groups' *SE* histogram

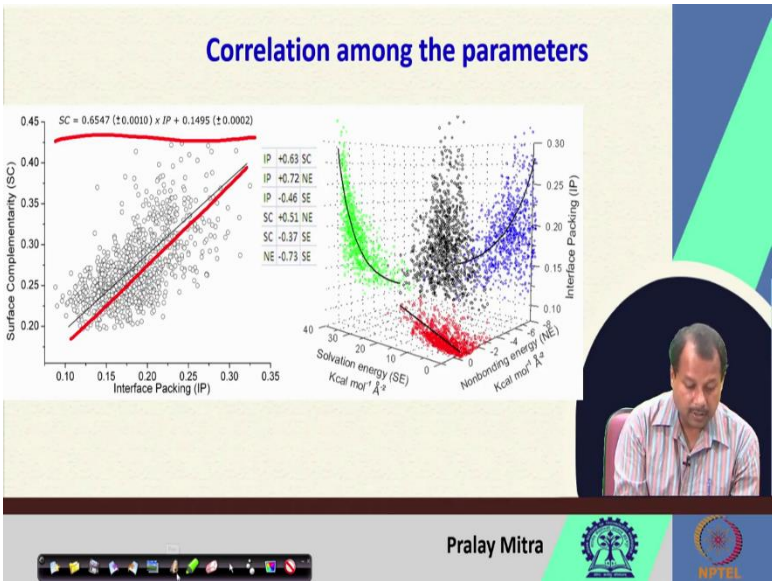
$$Score = \sqrt{((NE_i^G \times NE_i^G) + (SE_i^G \times SE_i^G))} + SP^G \times 10.0$$

Sort (in ascendi
group of deco
their scores

Rank of a decoy is its position in the sorted list

Mitra and Pal (2011) *J. Comput. Chem.* **32**(5):787-796

Pralay Mitra



Correlation among the parameters

IP	+0.63	SC
IP	+0.72	NE
IP	-0.46	SE
SC	+0.51	NE
SC	-0.37	SE
NE	-0.73	SE

Pralay Mitra

Now, to tell you the algorithm, it is very simple. So, the input is two protein structures, output is a ranked complex, that is the decoy. But here, we are focusing only on the scoring functions, we are not dealing with the geneticin phase. So, as you remember that in protein complex in protein-protein docking, so first step is generating the decoys, next is scoring that decoys and ranking the decoys.

So, geneticin phase extensively we have discussed. One topic of the generation phase was a brute force technique followed by fast furrier technique then geometry hashing and then SymDock based generation. Now, once it is generated, then we are passing it for the scoring and ranking.

So, first we computed five features IP, SC, NE and SC at the decoy interface, for all the decoys. Then group the decoys such that all decoys with RMSD less than 1 angstrom and differences in SP less than 0.04 is in group G, where SP is defined by SC minus IP multiplied with 0.6547 minus 0.1495. So, this equation actually has come from my previous slide.

Here you see that, this is the straight line. So, the equation is this one and from here, if I compute SP then I will get basically the deviation of a point from the theoretical or ideal value, that I will get as SP. Next, we will see, that how this SP will be utilized. Non-bonded energy for a group is considered as mean minus standard deviation since I wish to minimize the non-bonded energy. And since I wish to increase the solvation energy, so, for the solvation energy of a group, it is the mean plus one standard deviation. So, what I did, in summary again, that I got a number of decoys. Now, from those requests, I grouped them based upon their RMSD.

Because the similar RMSD or similar decoys is of not much importance because it is not giving much information. So, I group them, after grouping them corresponding to each group I am computing the SP and I am computing the non-bonded energy and salvation energy and, of course, corresponding to each group one representation will only be there or one representative decoys, will only be there.

So, not all the decoys will be utilized, so that way, the number of decoys will be reduced by this grouping technique if similar structures are available with me. Then, NEGI, NG is the bin number of all groups in the histogram. So basically, what I did that, I make one 5 cross 5 matrix and in these 5 crosses 5 matrix starting with 0 then it is going, so this way and this way.

Now, solvation energy is going this way and non-bonded energy is going this way. What I will do that NEG is the bin number of all groups NE histogram and SCG is the bin number of all groups SE histogram. Then what I will plot that, I will compute so all the values of the SC and NE I will plot within these bins.

If I plot that one, then based upon NE, I will have say one I, based upon SE I will have one j so its position will be in some Sij which is either 0 to 4 this way or 0 to 4 this way. For example, A is here B is here. Now, the distance from the origin where this I am considering as origin to A and B will be computed that is my Euclidean distance.

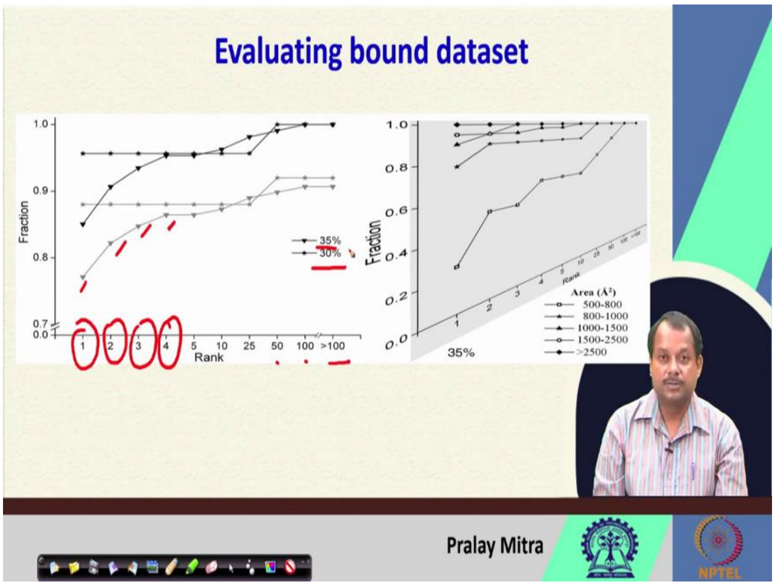
How I am computing, NEGi, NEGi, SEGi, SEGi, so these are my bin numbers. I computed the Euclidean distance after computing the Euclidean distance this will get an offset of SP multiplied with 10. Since this will be a very small number compared to the Euclidean distance of this one. So, in order to give one in order to give one comparable numbers, so what I am going to do is, I am basically give a multiplier that is multiplied with 10.

So, after multiplying with 10 then I am adding with this Euclidean distance and I am getting the score function. So, that is going to be the score function corresponding to each decoys. Now, each decoy will now will get one score function, now this decoy if belongs to some group then actually it is the score function of that group.

Because I started with the number of decoys and I grouped them. Since I have grouped them then based upon each group I computed the score function, and inside the group individual decoy's existence are not there, only one representative will be there. So, I have the score function, I will rank, I will sort that score function and in ascending order on the group of decoys based upon their score and rank of the decoys in its position will be telling me that what is that rank of that solution.

That way, I will not provide only one solution, but the algorithm or the flow diagram will give you a number of solutions and that is also a good idea to provide. Because sometimes if you give only one solution then that may be a problem. So, instead you can say that top 5 top 10 and you are expected to get one correct solution within top 5 or top 10 so that way, you provide at least 5 or 10 solutions.

(Refer Slide Time: 28:29)



Evaluating bound dataset

Pralay Mitra

Now, it’s time to check that whether my solutions are correct or not. So, as I mentioned instead of one rank what I would do that I will give you say, multiple lengths. So, 1, 2, 3, 4, 5, 10, 25, 50, 100 greater than 100 it is given here. And I am telling that out of say n number of cases what fraction of cases are giving me say rank 1, so that is here, and rank 2 that is here, rank 3 that is here, rank 4 that is here.

Now, the difference here between these is that so 30 percent or 35 percent sequence identity I am using based upon that one some differences are there, but this is evaluating on the bound data set. So, it has not much importance. What is important is, when I am going to basically evaluate on the unbounded asset. So, that I will continue to my next lecture. Thank you very much.