

**Algorithms for Protein Modelling and Engineering**  
**Professor Doctor Pralay Mitra**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology, Kharagpur**  
**Lecture 34**

**Discriminating Biological Protein Interfaces from Crystal Artifacts**

Welcome back. So, today I am going to start one new topic that is discriminating biological protein interfaces from crystal artifacts. So, it will take couple of lectures. Let us start today on this topic.

(Refer Slide Time: 00:30)

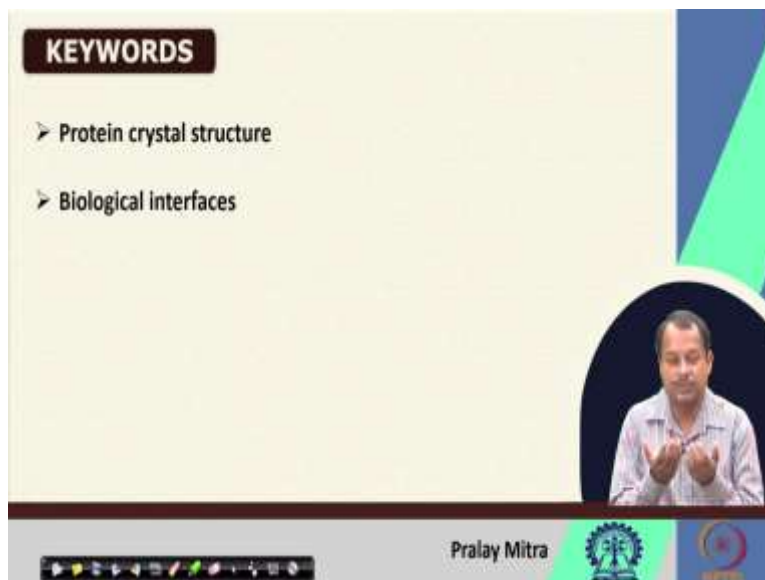


**CONCEPTS COVERED**

- Protein crystal structure
- Biological interfaces

Navigation icons: back, forward, search, etc.

Pralay Mitra






**KEYWORDS**

- Protein crystal structure
- Biological interfaces

Navigation icons: back, forward, search, etc.

Pralay Mitra

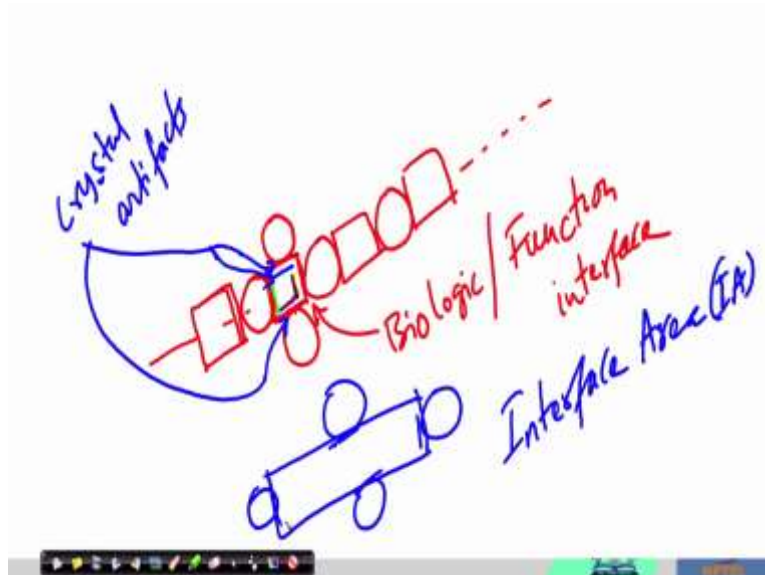


So, the concepts that I wish to cover is protein crystal structure and biological interfaces. Keywords accordingly is bit same to that, that is protein crystal structure and biological interfaces. Now, let us start.

So, let me tell you one thing that I also summarily mentioned in the last lecture that when it is the crystal structure then I cannot say with certainty that whether the interface or the contact between two chains of a protein which is deposited at the Protein Data Bank is the correct biological interface or not. That I cannot say with certainty. It is a known fact.

The primary reason is that unlike to NMR, so in case of protein crystallography, so protein structures are determined from the protein crystals. So, although the solution is required, actually protein will be dissolved and then it will take some crystal structure, but during the crystallization process if it happens that it will, it will misinterpret its interface then it may create one problem. That is not much severe.

(Refer Slide Time: 01:58)



But rather severe thing is that when it is the crystal lattice then what will happen, that it will keep on repeating. So, instead of, say protein if I, say draw one simple rectangle, so if it is one rectangle it is fine. If in solution, along with that one rectangle, say there is one circle in solution, it is fine. We can say that the interface of the rectangle or, say let us consider as square so that I have 4 different possibilities and they are equal.

So, if it is a square and with that square the circle is forming one interface then what you can see? That if it is a solution then I can say and I understand that what is the interface. Now, if I mark those interfaces, say with different colors; this is one interface, say although it is square but I wish to mention that, sorry, they will be different. They are not identical. I think yellow is not visible much. Let us replace. So, four different colors are indicating four different sites.

Now, if it is a crystal structure then what will happen? That the same thing will keep on repeating in a lattice form, so which means, after this one there will be another square here, then another circle, then another square and it will go on, right here there will be one circle, then here there will be one square. Now, it will be extended here also, this side also. Now, it is into 2D. But in 3D if you consider then it will have several opportunities, different opportunities.

Now, since I am considering only the square, in 2D also so I can demonstrate. Now, you see that with this circle there are four interfaces if I consider the lattice. If I consider the solution then only one interface between, say this, this gray which is here. So, this gray kind of color, so it is the correct one. Now, if it is the lattice form then all four are of equal probability. How do I discriminate that whether it is blue side, pink side, green side or gray side. Which will interact with the circle? That is the main problem.

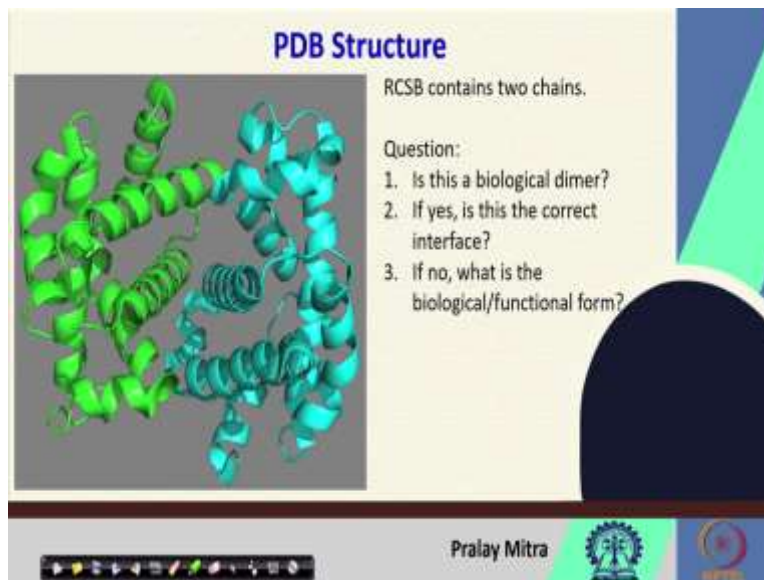
So, when I will say that this is my biological or functional interface but all this, this one, this one and this one, so all these are my crystal artifacts because these interfaces are created because of the crystal lattice formation. If there is no crystal lattice formation then this kind of interfaces will not appear. This is one of the major problem.

That is why we will see that when you open it in PDB and you look at the detail remarks section of the Protein Data Bank then you will find that what is the biological organization, corresponding to that one, one transformation matrix is also provided so that you can get that biological orientation.

So, definitely that is through some software, and based upon some logic. So, initially, so in 90s or early 2000 also it was believed that mostly it is the interface area that plays a crucial role. Which means if instead of it is square if that is a rectangle and there are, say four circles like this, then whose interface area is higher will be biological, more biological.

It is true that if the interface area is large enough then the interaction size or interaction site increases. But that necessarily does not mean that it will be the biological or functional form. So, those things we need to deal with, and accordingly we need some method in place which will take care of that.

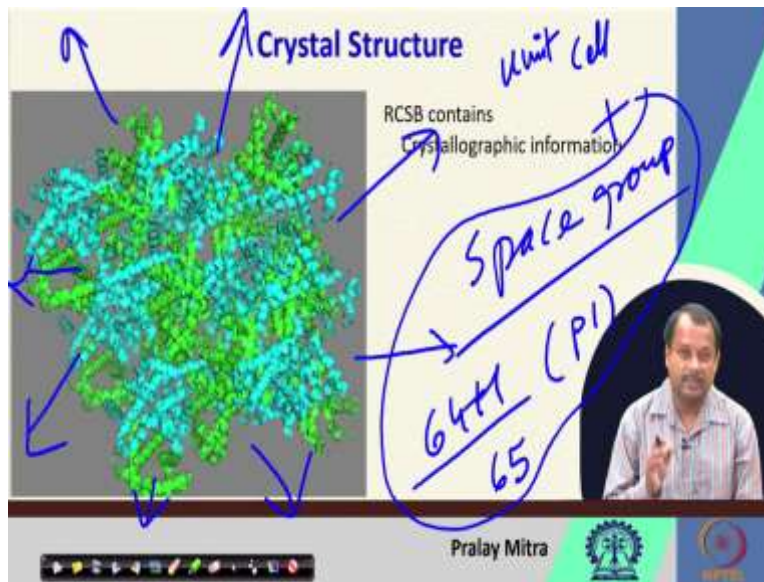
(Refer Slide Time: 06:36)



So, this is one PDB structure. So, RCSB contains two chains. So, the question is this a biological dimer? So, on top of what I have discussed. So, first I discussed that, so one interface is given to you and you wish to check whether that interface is biological or not biological. Next question is that is this a biological dimer? So, I am asking that whether, what is the functional form? After that one I will look for the interface area, or it can be vice versa also. I cannot that way separate it out.

So, basically I have two connected questions here. One is that what is the functional form, and what is the organization. When I say what will be the organization then I will ask basically to report me what are the interface area. So, that thing last time I went mentioned that earlier belief was that if the interaction site increases then largest interaction site is going to be the biological interaction site. But that may not be necessary true. So, two pertinent questions, what is the biological dimer? Is this a biological dimer? If yes, then is the correct interface? If not, what is the biological or functional form? So, this question we wish to answer.

(Refer Slide Time: 08:05)



The reason why I raise this question is clear from here. So, what I did, in pyMOL actually with that structure, that I shown you on the last slide, I just go for create symmetry and I gave some boundary, say 20 Angstrom or, say 40 Angstrom. Within that you create one symmetry operation, one symmetry, and that symmetry is related to that lattice. So, if it is lattice, so in all the directions it will keep on growing. In all the directions it will keep on growing.

So, this symmetry is basically generated by the pyMOL not using the symmetry that we have discussed, that cyclic or dihedral symmetry. This is a separate symmetry which is related to the crystal lattice and that is called as the, based upon the space group. It was a nice work by one professor from the University of California at Los Angeles who identified that there are several space groups but in case of protein only 64 plus 1, so this 1 is basically the identical, so P1. So, in total 65 space groups exist. And if you know how many such space groups are there then basically you can generate the crystal.

So, we are not going into that detail, that space group etc part, this is just for your information. But the point is, given one crystal structure and what is the space group of the crystal, along with that one, that a, b, c; alpha, beta, gamma I mean that unit cell information, unit cell and the space group information, if were given to you then you can generate the crystal lattice. How? That we will not discuss in this course. It is possible.

(Refer Slide Time: 10:23)

Crystal structure suggests multiple interfaces

Which one is functional/biologically relevant?

Pralay Mitra

I actually use the pyMOL software right now. So, using that one I am generating that lattice. So, once I generate the lattice, so it will be looking like that. And drawing the same analogy of the square and the circle that I demonstrate you, if it is a situation then this red color indicates one dimer. So, here there are two chains in the red color. So, if you figure, if you look at it carefully then you can able to figure it out.

Now, if I consider only one, if I consider only one, say chain, so this chain if I consider then along with this, this interface present in crystal but here, one interface is coming from here, one interface is coming from here, one interface is coming. So, multiple interfaces are coming.

So, which is going to be the relevant one or which is going to be the functional or biological one? That is the question we need to ask. And we need to come up with some algorithm which will take one PDB structure as an input, the PDB structure contains atomic level information, its unit cell information as well as its space group information from which it will generate the total crystal lattice, it will identify all possible interfaces from that crystal lattice and then it will discriminate which one is the biological and which one is the crystal artifacts.

(Refer Slide Time: 11:42)

### Introduction

- Existing methods are not adequate to automatically infer the quaternary structure from atomic and lattice information.
- Existing methods:
  - PQS
  - ASA and Pair Score
  - ACV
  - 3D Complex
  - PISA@EBI - Chemical thermodynamics
- Manually curated database:
  - PiQSi@MRC

Pralay Mitra

So, long back people have started to work on that one. So, existing methods are there but those existing methods like 3D complex, PQS protein quadratic structure, ASA and pair score, say few of them are not basically adequate enough except this PiQSi which is manually curated and this PISA which is hosted at EBI. 3D complex is fine but its support is not there much.

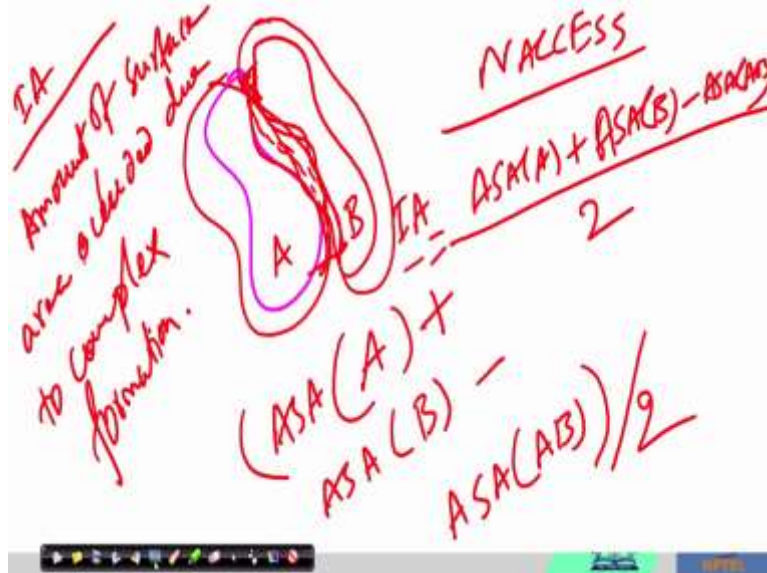
(Refer Slide Time: 12:18)

### Features at protein-protein interface

- interface area (IA),
- normalized interface packing (NIP),
- normalized surface complementarity (NSc),
- normalized surface complementarity and interface packing paired metric (NSP),
- accessible surface area variation (asaV),  
$$asaV = (IA_{1,2} - IA_{1,3}) / IA_{1,4}$$
- interface packing gradient (IP<sub>g</sub>),
- patch ratio (P<sub>r</sub>),
- normalized solvation energy capacity (NSE),  
$$NSE = \frac{\sum \Delta \sigma(\text{Atom Type}) = \Delta ASA}{\text{Interface area}}$$
- hydrophobicity at interface (HPO<sub>i</sub>),
- hydrophobicity at the surface (HPO<sub>s</sub>).

*Handwritten notes:*  
NACCESS  
ASA  
Solvation Accessible Surface Area

Pralay Mitra



So, what we are trying to do is to check if we can have one algorithm in place which can take care of this or not. So, for this we first, we will take the help of machine learning technique. So, before that machine learning technique so we will identify some features which are relevant at the biological protein interfaces and which are unique for the biological protein interfaces rather than the crystal artifacts.

So, first thing you have to understand that when it is the biological interfaces then the fitting is perfect. So, when, say this is one protein molecule, this is another protein molecule and they interact, I am assuming that they are interacting in water molecule then this interaction is perfect.

But in the crystal what is generated is like this. So, which means if I look at the complementarity or if I look at the packing of the interfaces, if I look along with that, at the amount of interactions, if I look at some energy functions like, say solvent energy, if I look at the hydrophobicity at the interface that will definitely be higher compared to any crystal artifacts, because for the artifacts, so because of the crystallization this interface has been created, but actually it does not exist.

So, there is no binding for this interface to be very compact, to be hydrophobic in nature, to support the solvent energy function or to be complementary with each other. So, that is why we picked, so ten features.

So, first one is, of course, my interface area. We discussed in detail that how to compute the interface area. Here we use the NACCESS software which exists with the University College,



London. So, this is very fast, so this will give you solvent accessible surface area. So, sometimes it is called as the SASA, solvent accessible surface area. Sometimes it is called as the ASA, accessible surface area because by default we are assuming that it is solvent in nature.

Next, it is the normalized interface packing. Normalized surface complementarity we will discuss in detail between these two, about these two along with this NSP. Then accessible surface area variation. So, this accessible surface area, ASA, so this interface area we know how to calculate. So, interface area means that first we have to compute the, so accessible area of one protein molecule and the another protein molecule and then we have to subtract that one.

So, let me explain this 5 and 1, so accessible surface area variation and interface area. So, by interface area, so what I mean that, say you have one protein like this, and another protein, say like this. Now, I am interested about the area occluded due to complex formation. So, for that what I need to do, that if I use NACCESS software or any other software, first, say this is my chain A. This is my chain B. I will extract chain A and I will compute ASA for chain A when B is not present. I will compute ASA of B when A is not present. So, when B is not present I will get the ASA of this whole part. When A is not present I will get the ASA accessible surface area of this whole part.

Now, what I will do? I will compute the ASA of AB as complex. Now, if I now add ASA A plus ASA B and then subtract ASA AB then what I will get? I will get this region and this region. So, what I am going to do? With this I am dividing this 2. Or in short,  $ASA A + ASA B - ASA AB$  whole divided by 2, that is going to be my interface area.

As per the definition, so IA, interface area is the amount of surface area occluded due to, due to complex formation. So, that is the amount of surface area occluded due to complex formation. That is my interface area. And how I can compute that one? I gave that equation to you. So, I believe that you can compute that.

(Refer Slide Time: 17:49)

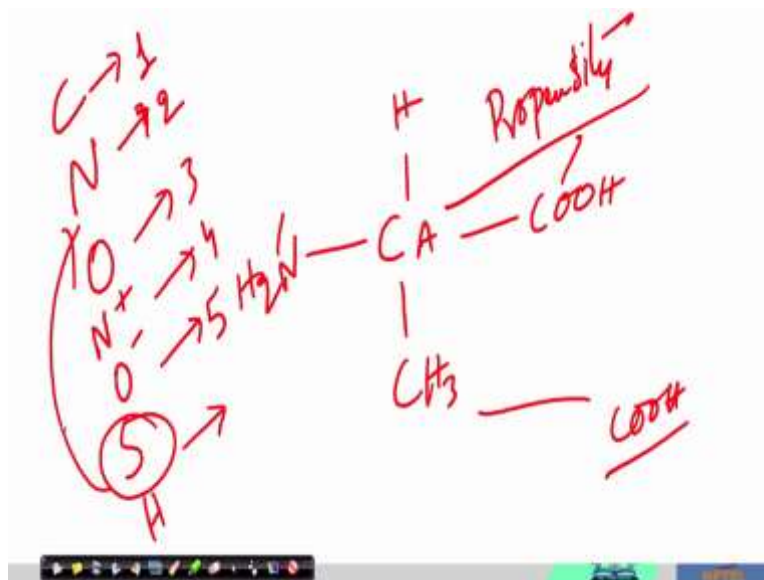
### Features at protein-protein interface

1. ✓ interface area (IA),
2. normalized interface packing (NIP),
3. normalized surface complementarity (NSc),
4. normalized surface complementarity and interface packing paired metric (NSP),
5. ✓ accessible surface area variation (asov),  

$$asov = \frac{(IA_{1.0} - IA_{2.0})}{IA_{1.0}}$$
6. interface packing gradient (IP<sub>g</sub>),
7. patch ratio (P<sub>r</sub>),
8. ✓ normalized solvation energy capacity (NSE),  

$$NSE = \frac{\sum \Delta \sigma(\text{Atom Type}) \times \Delta ASA}{\text{Interface\_area}}$$
9. hydrophobicity at interface (HPO<sub>i</sub>),
10. hydrophobicity at the surface (HPO<sub>s</sub>).

1.4



Now, if I go back and see the next, this point 5th, accessible surface area variation. Now, you remember that rolling the probe concept. And using the rolling the probe concept basically we calculated the accessible surface area. And during that time I mentioned that if the probe size reduces then it can go deeper. If it increases then it will not go deeper.

Now, when I am trying to see the compactness of the two surface; this is one surface, and, say this is another surface, this is another surface. These two are surfaces. If I am trying to measure that then what I am doing? That I am computing the interface area, same definition that I have used here, so using that one I am computing the interface area. When my probe size is 2.0 that

probe size variation I can do in NACCESS. And when my probe size is 1.8, that is little higher compared to the probe size of the water molecule which is 1.4. 1.4 is the probe size of water molecule, and that is the default in NACCESS.

Now, here what I am doing? I am increasing the probe size. One is 1.8. Another is 2.0. And then I am computing what is the difference. Then I am normalizing that one by the actual interface area. Now, my point is that if there is a very good fitting, then because of that one this variability will be less. But if the fitting is not very good, so this is one situation, this is another.

So, on this region, this region, here, on this region you see that here the packing or the fitting is better compared to here. If it is then if I change the probe radius here then what will happen? So, 1.4 will go deeper, 1.8 little less, 2.0 more less. That variation will be there. But here 1.4 definitely will go deeper but the variation of 2.0 and 1.8 with respect to 1.4 will not be much. So, that is the point and we wish to exploit that one.

Another is interface packing gradient and the patch ratio. Then 8th point is very interesting, that is the normalized solvation energy capacity. So, here we are borrowing the concept taken by the Eisenberg long back where he classified the atom type for one protein molecule into five classes. Now, if you look at the amino acid composition, so what we have? Say for, I am considering, say alanine. So, in this case you have C, you have N, you have O. Occasionally for methionine and cysteine you have S. And you have H. So, these are the different atoms.

So, what that Eisenberg has done long back? So, carbon is one group. For this nitrogen he has divided into two parts, one when nitrogen is being part of this, say main chain, here, oxygen part of this main chain here. Another, when this nitrogen and oxygen will be part of the side chain so when? For, say aspartic acid, glutamic acid, so acid. So, here one C double O H will come for aspartic acid and glutamic acid. Along with asparagine one OH will go there. Theanine, one OH will go there. For OH, so one tyrosine will go there.

So, his point is that when N and O belongs as a charged one, as part of the charged residue, aspartic acid, glutamic acid or lysine and arginine, then their behavior will be different compared to the other existence of N, O. When it is main chain or is being part of, say tyrosine, phenylalanine, theanine, etc. oxygen.

So, based upon that one he picked five categories, say 2, 3, 4, 5. Since this sulphide occurs very less, only for the cysteine and methionine; and if you look at the propensity, propensity of the, or the frequency of the occurrence of the sulphur, sorry the cysteine and methionine, sulphur is part of that one, then that is less compared to other residues. So, for this S you may have optionally another category or you may put it into the category of, say charged or uncharged nitrogen or oxygen. That way, five different category classes has been created.

(Refer Slide Time: 22:57)

**Features at protein-protein interface**

1. interface area ( $IA$ ),
2. normalized interface packing ( $NIP$ ),
3. normalized surface complementarity ( $NSc$ ),
4. normalized surface complementarity and interface packing paired metric ( $NSP$ ),
5. accessible surface area variation ( $asaV$ ),  
 $asaV = (IA_{1,2} - IA_{1,3}) / IA_{1,4}$
6. interface packing gradient ( $IP_g$ ),
7. patch ratio ( $P_r$ ),
8. normalized solvation energy capacity ( $NSE$ )  
 $NSE = \frac{\sum \Delta \sigma(\text{Atom Type}) \cdot \Delta ASA}{\text{Interface area}}$
9. hydrophobicity at interface ( $HPO_i$ ),
10. hydrophobicity at the surface ( $HPO_s$ ).

ask the interface  
 Contact area  
 Solvation energy  
 NO FP data

Pralay Mitra

So, that is my atom type here. That is my atom type here. So, along with that one, the change of the ASA due to complex formation is being multiplied and those are summed up for all such occurrences divided by the interface area is giving me the normalized solvation energy.

Basically this is the equation when I am not dividing by interface area so the numerator part is the basically the definition as per the Eisenberg of solvation energy. I am normalizing by interface area. That is why I am calling that as a normalized solvation energy.

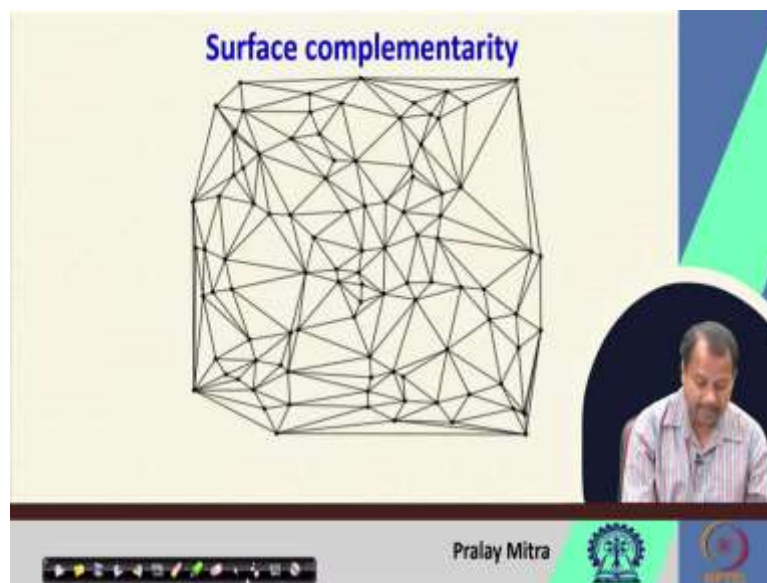
Next ninth and tenth are my hydrophobicity at the interface and hydrophobicity at the surface. Here what I would do? That based upon one, one hydrophobic scale, so there are several hydrophobic scales. So, where the amino acids are ranked based upon their hydrophobicity. So, I will pick one hydrophobic scale.

So, it can be Kyte Doolittle, KD or it can be Fauchere and Pliska, FP, so, or it can be Janin, Chothia. So, few are based upon, say experiment. Few are statistical. So, if I pick one then based upon that one I can have the value. Actually based upon that scale, I will have one index, one value about their hydrophobicity. So, that hydrophobicity information when I will multiply with the residues again.

So, all these features I am calculating at the interface. I am not telling that whether this interface is going to be biological or not but whatever the contact area, or better I should use the term, instead of the interface I should better use the term contact area because two protein molecules are in contact with each other. I can reserve this word interface for the purpose when the contact area is going to be biologically relevant or they are interacting one.

Now, these, all these features I have computed at the contact area. After computing that one, so I will go for the machine learning technique. But as I promised that the second and third, interface packing and surface complementarity I will explain in detail about their algorithm and then I will go back for the actual algorithm for that classification.

(Refer Slide Time: 25:45)

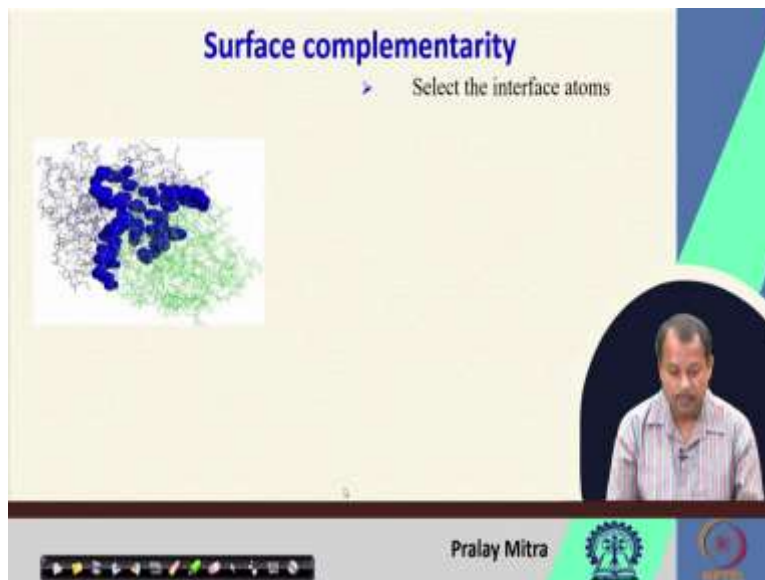


So, in order to calculate the surface complementarity what I will do? First I will identify the interface atoms. After identifying the interface items then I will give one transformation. That transformation we have discussed during the rotation about an arbitrary axis. And after the transformation I will consider each atom as one point and considering that one I will go for the

Delaunay triangulation. So, this Delaunay, sorry I will go for the Delaunay triangulation of the atoms on the surface. That will give me a number of triangles which are looking like this here. So, it will be done in 3D. That is why one 3D example is given here.

Now, we are getting triangles. Then what we would do? We will check the complementarity between those two triangle, one triangulation will come from one protein molecule, another will come from the another protein molecule. So, we are considering the contact area. So, for A and B, for A one set of atoms from the contact area, for B another set of atoms from the contact area I got. And that contact area I can get using that NACCESS atom which will declare that whether it is at the interface or not.

(Refer Slide Time: 27:26)



After getting that information, so this is our algorithm. What it will do? That select the interface atom first and you can see blue and green lines indicating one protein complex or say, one decoy where blue colors are the atoms at the interface which belongs to blue line protein. Similar to that green line protein there will be atoms, but in order to make it simple the diagram, so I kept only blue and not showing you the green. But green is also there.

(Refer Slide Time: 27:55)

The slide is titled "Surface complementarity" in blue text. On the left, there is a 3D ball-and-stick model of a protein surface, colored yellow, with a coordinate system (x, y, z) overlaid. To the right of the model, a blue arrow points to the text "Apply transformation". In the bottom right corner, there is a circular inset showing a man in a light-colored shirt speaking. At the bottom of the slide, there is a navigation bar with icons and the name "Pralay Mitra" next to a logo.

So, after identifying that one, so I will apply some transformation with respect to a set of points so that irrespective of the initial orientation of the atom it will always assume one particular orientation during the Delaunay triangularization so that, irrespective of the initial orientation, every time, corresponding to one particular, one particular decoys, docking decoys actually, I will have or, say corresponding to one particular contact area or orientation I will have one unique value for the surface complementarity.

(Refer Slide Time: 28:35)

The slide is titled "Surface complementarity" in blue text. On the left, there is a 3D ball-and-stick model of a protein surface, colored grey and yellow, with a Delaunay triangulation overlaid. To the right of the model, a blue arrow points to the text "Delineate interface atoms of each subunits by Delaunay triangulation". In the bottom right corner, there is a circular inset showing a man in a light-colored shirt speaking. At the bottom of the slide, there is a navigation bar with icons and the name "Pralay Mitra" next to a logo.

Next I will delineate the interface atom of each subunit by Delaunay triangulation that I have mentioned. Now, I am showing you two interfaces. One is with the gray, another with the yellow. From here also we can see that that triangles are now facing to each other.

(Refer Slide Time: 28:54)

**Surface complementarity**

Compute the complementarity between the triangles of the different subunits.

Surface complementarity =  $\frac{\text{Complemented Area}}{\text{Total Triangle area}}$

Surface complementarity is divided by interface area to get **Normalized surface complementarity (NSc)**

Pralay Mitra

What I will do next is I will compute the surface complementarity. In order to do that one here is the schematic diagram where it indicates that this has come from one protein molecule, and this has come from another protein molecule. So, this red is from one protein molecule and, say blue is from another protein molecule.

Now, our threshold value is, if the centroid between two triangles are within 6 Angstrom and if the plane containing the two triangles are within 25 degree then we declare those two triangles are complemented with each other. So, surface complementarity is the complemented area divided by total triangle area. So, that way we will compute whether this triangle is complemented with this or not, whether this is complemented with this one or not. And we will identify how many such triangles are complemented.

So, we will compute the area of those triangles. It is very easy to compute the area of the triangles. Then we will normalize that one by the total triangle area. And that will give me the surface complementarity value. And if I divide further it by the interface area similar to NAc calculation then I will get normalized surface complementarity. So, that is it for today's lecture.



(Refer Slide Time: 30:35)



Now, we will continue actually to other features, interface packing and then we will integrate to some machine learning technique. Thank you very much.