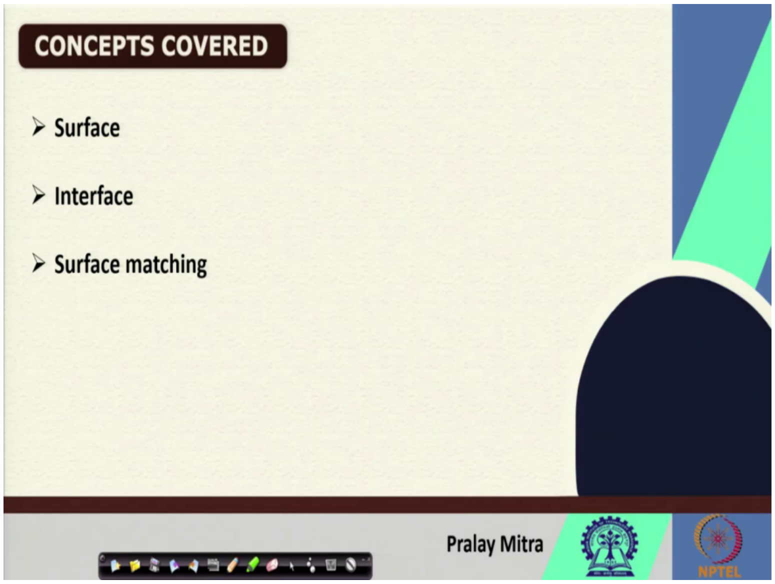
Algorithms for Protein Modelling and Engineering
Professor Pralay Mitra
Department of Computer Science and Engineering
Indian Institute of Technology Kharagpur
Lecture 07
Digitization of a Molecule

Welcome back. We are in the process of discussing the digitization of a molecule. I promised that I will show you what is the advantage of digitization of a molecule? Before that let us look into the detail of this digitization of a molecule.

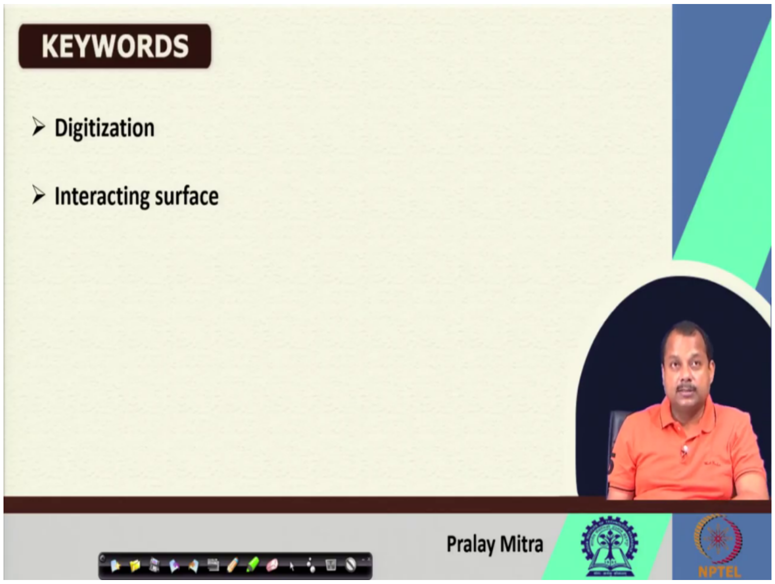
(Refer Slide Time: 00:42)



CONCEPTS COVERED

- Surface
- Interface
- Surface matching

Pralay Mitra



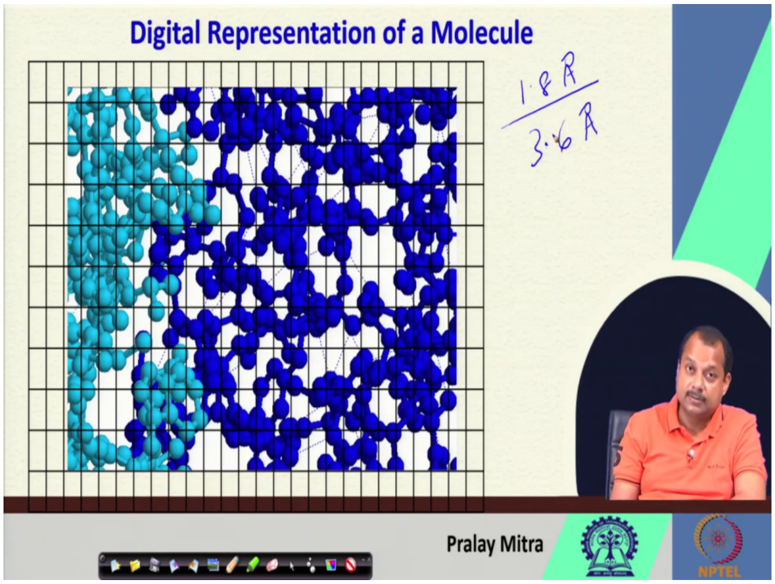
KEYWORDS

- Digitization
- Interacting surface

Pralay Mitra

Today we are planning to cover the discussion of the surface of a molecule, interface of a molecule and surface matching. That’s why I picked the keyword interface, interacting surface and digitization.

(Refer Slide Time: 01:00)

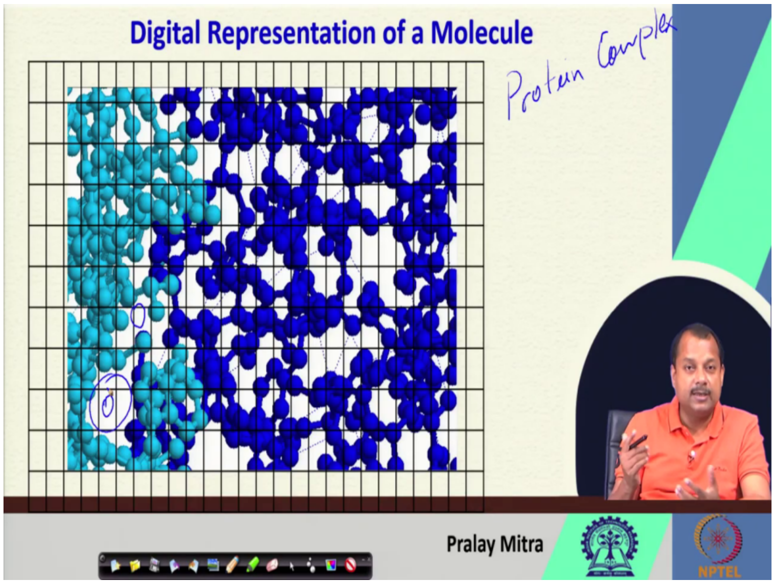


Digital Representation of a Molecule

1·8 Å

3·6 Å

Pralay Mitra



In the last lecture, I mentioned putting the atoms in a grid and before that one I discussed if I put the protein chain or the protein molecule inside the grid, then what will happen? For that first I need to decide the grid size which includes calculating the left extreme, right extreme, means X_{min}, X_{max}, Y_{min}, Y_{max}, Z_{min}, Z_{max}. After deciding that I can add some offset so that the extreme points are not touching the boundary of this rectangle or say in three-dimensional space.

If I add that extra space, then I can consider as if in that container the protein molecule is placed, the size of the container is decided by the X_{min}, X_{max}, Y_{min}, Y_{max}, Z_{min}, Z_{max}. After that, you need to divide that container along the X-axis, along the Y-axis, and the Z-axis.

When you are dividing into grids then you have to decide what will be the grid step. If it is in angstrom (Å) then the grid size will also be of the order of angstrom. Also, keep in your mind the van der Waal radii of those atoms. For example, if I consider sulphur that is with van der Waal radii 1.8Å, hence the diameter is 3.6Å. Now, if you decide grid step that is more than 3.6Å then you can accommodate multiple atoms inside one grid cell position. If you decide small grid cell, say 0.5Å then you cannot accommodate any atom completely inside one grid cell, rather you have to split it over the different cells.

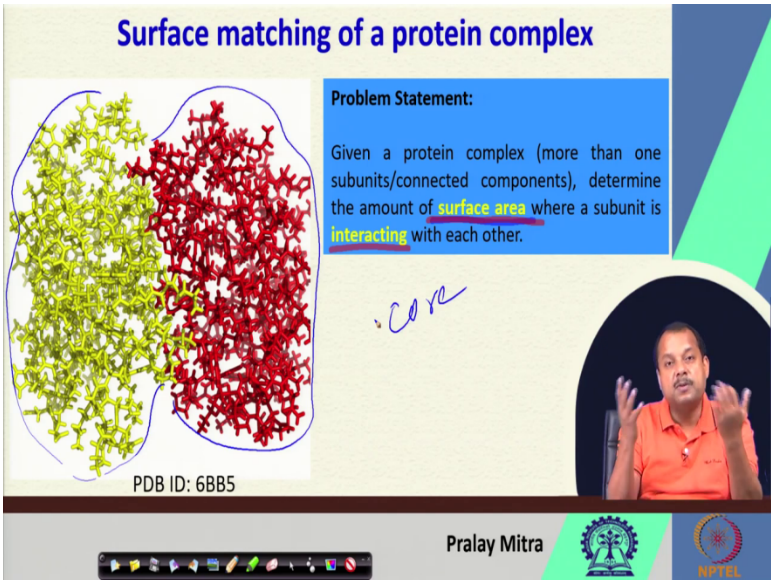
Which one is the best? I do not know. It is based upon the application but those are two things you have to keep in your mind when you are deciding on the grid step. In this context, I am talking about putting one protein molecule in a grid and then deciding on the grid size, grid step.

I did not mention whether it is only one protein molecule or multiple protein molecules. I did not mention whether it is a one connected component or subunit or chain (all are used for the same purpose), or multiple subunits, multiple chains. In this example, you can see two colours are indicating two different subunits or chains, which means it is a protein complex. That indicates either there are two protein molecules, or one protein molecule another DNA, or one protein molecule another RNA - in general, one protein molecule along with another biomolecule. I am putting them inside this grid box. After that, if I use the same concept that I discussed in the last lecture that using some thresholding operation or percentage of occupancy of an atom inside a cell, I shall mark the cell by 1 or 0. The cell will be marked 1 if contains any atom, 0 otherwise.

Likewise, I can say whether the particular cell contains some molecule or not. You may be thinking that there may be some empty spaces inside the protein molecule. Since no atoms are present so these cells will have the value 0. That is true.

In reality, hardly it occurs. Considering the ball-and-stick model of the protein molecule where atoms are not representing their radius, you may be feeling that there are empty or white spaces inside a molecule. Barely there is any white space inside the protein molecule. I said barely. There may be. During the crystallization process sometimes some water molecules or heteroatoms are co-crystallized with the biomolecule. If you remove those from your structure then you might see some void that is inside the protein molecule. That way, at times you cannot differentiate between a cell position containing 0 whether it is inside the protein molecule or outside. Nonetheless the concept we are going to discuss you will see hardly matters. So, we are stick on to our definition that if one grid cell contains some atom then it will be 1, 0 otherwise.

(Refer Slide Time: 07:49)



Surface matching of a protein complex

Problem Statement:

Given a protein complex (more than one subunits/connected components), determine the amount of surface area where a subunit is interacting with each other.

·core

PDB ID: 6BB5

Pralay Mitra

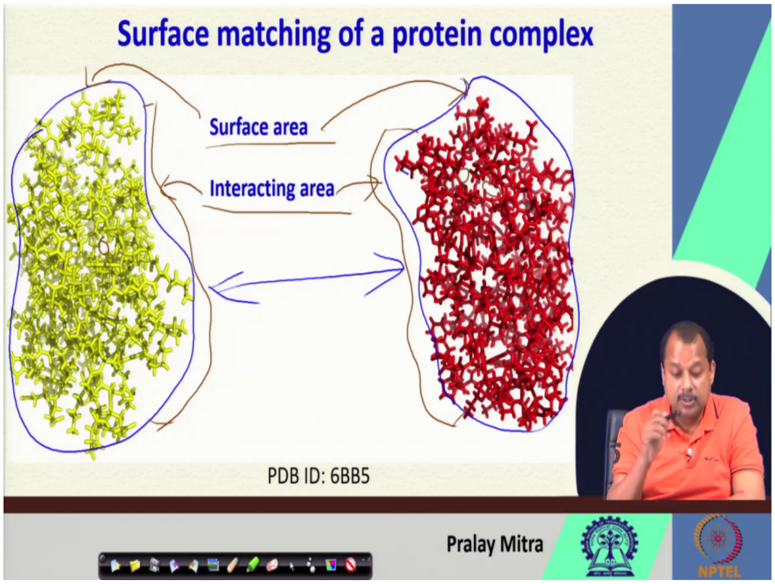
Here is the problem statement - given a protein complex (more than one subunit or connected component are present) determine the amount of surface area where a subunit is interacting with each other. So, two new things you see here one is surface area and another is interactions. Now, if I look at the structure again - this is the PDBID: 6BB5 that we discussed in our first week.

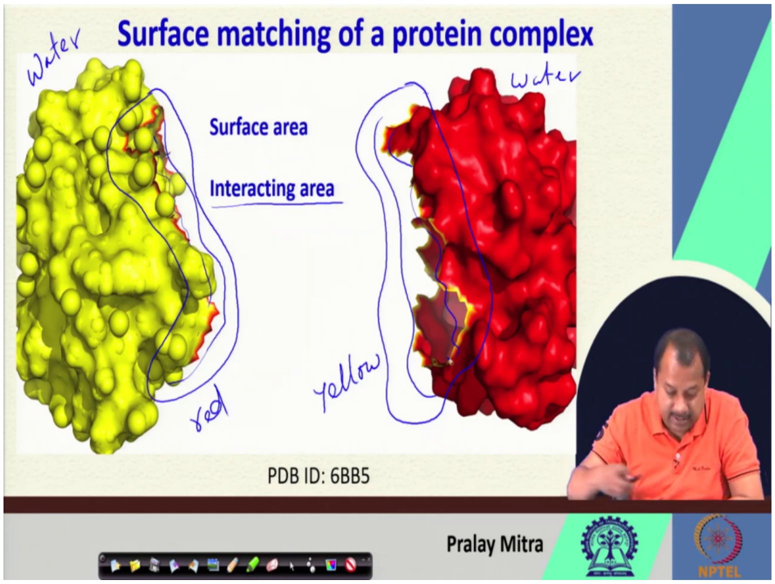
You see this is the complex - yellow colour indicates one subunit, red colour indicates another subunit. In the complex, you can assume that this is the surface and this is the surface. The ***surface*** is defined by the atoms which interact with the water molecule. Typically a protein molecule is always in some solvent and one of the most well-known solvents you can consider is water. Consider that one protein molecule is deep inside in water. Hence, some atoms will be in contact with the water molecule and some atoms will not be in contact with the water molecule. If I consider that this is one protein molecule. Now, if I put this into some water then I shall see that these regions will be wet and if there is no leakage, then after taking it out, if I open you will see that this flat of the hand area is not wet. Since folding doesn't allow water molecules inside.

Those regions where water molecule has access when the protein molecule is inside the solvent water is considered as the surface and the atoms which interact or which are in contact with the water molecule is called the ***surface atoms***. Incorporating thresholding on the percentage of the atoms of a residue acting as the surface atoms gives me the information of whether a residue is on the surface or not. If it is not on the surface then I can call that it is inside the ***core***. Similarly, I can able to identify the atoms or the residues which are on the surface. There are several techniques or algorithms to decide on that. One simplest one we shall discuss here and using that we shall identify the percentage of overlap at the interacting surface.

Now, I noted that it is the surface. Similarly this protein molecule also, I have another surface. If these two-protein molecules interact with each other then only the surface atoms or the surface amino acids will be interacting. Core (inside the protein molecule) will not be interacting with each other. Interaction means that when two protein molecules are talking together or forming a complex then those two protein molecules are in contact and some residue or atom of that particular residue is making some bonds, mostly non-covalent bonds with each other, like hydrogen bonds. So, for this one, it is the interacting region. You can also understand from this example that it is the surface atoms or surface amino acids which will be part of the interface regions, not the core. Why? Because here you can see that in this region, the interacting region now, who will interact only those who are on the surface.

(Refer Slide Time: 13:28)


Surface matching of a protein complex



Surface matching of a protein complex

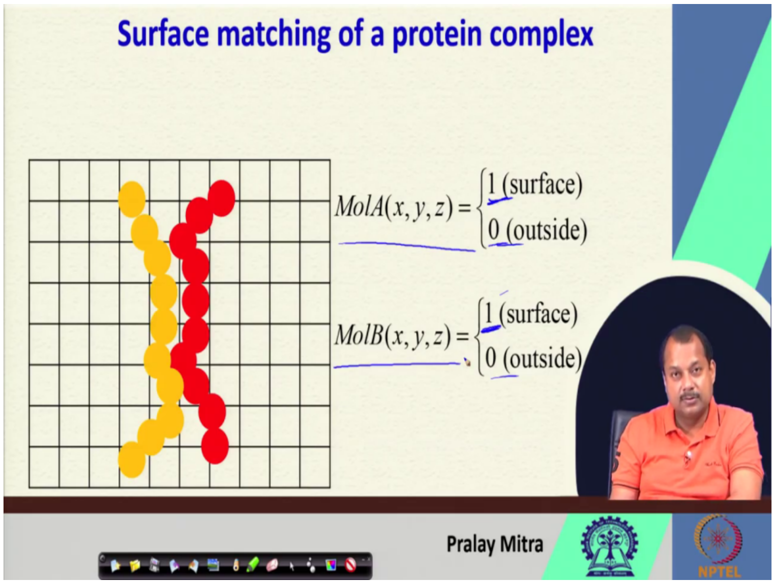
The more detailed diagram is visible when I am taking those two protein molecules apart from each other - yellow part and red part - chain A and chain B. You see these are surfaces that were interacting before I have taken them out. Considering they are coming closer to each other then – for this side let me give some other colours which will be visible - this one this region. The surface from this region will contribute to the interacting area and this blue is going to be the surface area. This is my surface area, this is my surface area, this is my interacting area. Again, you may feel that okay so there is some space inside which may go black etcetera. Don't think that way because if I go for this representation, which is the surface fill then you will see that there is no space. But may some spaces as I mentioned earlier filled with some water molecule or with some heteroatoms. And if I take those water molecules or heteroatoms out during my calculation then there will be some space.

I am not considering those spaces right now. But from this diagram, I believe it is more or less clear to you what does the surface area mean. How can I calculate that is a different issue. There are several algorithms for that. We shall discuss the simplest one for this calculation. But you know what is the surface area.

Now, when these two protein molecules exist separately then this is the surface area. But when they will interact then there is some interaction area. Probably you can see it here. Let me pick a colour, say blue. You can see the red region and here you can see some yellow region. The yellow region is interacting with this red region. That's why this colour has come.

These two are interacting areas - this area and this area. Rest is the surface. Truly, the interacting area is the subset of the surface area before the complex formation. Now, probably, I can give you another definition of the interacting area. An interacting area indicates the surface area which becomes occluded because of the complex formation. That indicates individually if I put it into the water on the left-hand side and right-hand side, these regions will get wet. I mean those atoms will be in contact with the water molecules, but when I allow them to form a complex then these inside regions and these inside regions will be interacting and they will be occluded from the water molecules. The situation is, this is a surface area, this is another surface area when I allow interactions and because of that interaction the water molecule cannot go inside. So, they were not in contact with the water molecule. Thus the definition of the interacting area is the subset of the surface area which will become occluded due to complex formation. Now, combining this definition and the grid representation, let us see what we can do.

(Refer Slide Time: 18:17)



Surface matching of a protein complex

$$MolA(x,y,z)=\begin{cases}1\text{ (surface)}\\0\text{ (outside)}\end{cases}$$

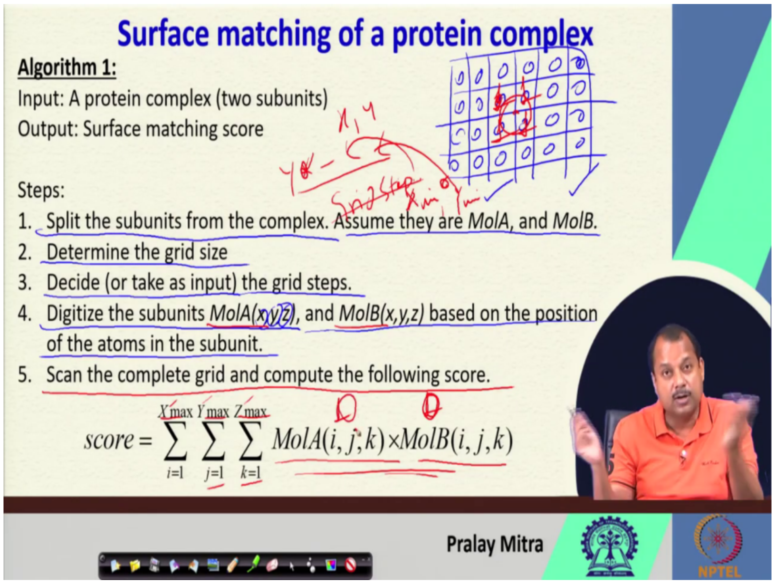
$$MolB(x,y,z)=\begin{cases}1\text{ (surface)}\\0\text{ (outside)}\end{cases}$$

Pralay Mitra

I decided about grid size and grid step. I put the complex inside this grid. You remember the first blue and light blue complex figure at the beginning of this lecture - here it is red and yellow. Atoms are in some cells. I decided that for *MolA* and *MolB*, 1 is the surface, 0 is the outside. So, if I multiply *MolA* and *MolB* for each cell position, then this 1 and this 1 sorry, this 1, this 1 and this 1, these two will be multiplied and that will give you one contributor indicating one cell (x,y,z) where yellow and red is overlapping. Overlapping means they are interacting, they come in close contact with each other.

But if one is 0 and the other is 1, then it will cancel out. If one is 0 and the other is 0 that will also cancel out. To count in how many cells they come into close contact with each other, you can design one algorithm which will scan this total grid, it will multiply *MolA*(x,y,z) and *MolB*(x,y,z) for each x, y, z and then take the sum of those multiplications to know in how many grid cells this red and yellow is becoming in close contact with each other.

(Refer Slide Time: 20:30)



Surface matching of a protein complex

Algorithm 1:

Input: A protein complex (two subunits)

Output: Surface matching score

Steps:

1. Split the subunits from the complex. Assume they are *MolA*, and *MolB*.
2. Determine the grid size
3. Decide (or take as input) the grid steps.
4. Digitize the subunits *MolA(x,y,z)*, and *MolB(x,y,z)* based on the position of the atoms in the subunit.
5. Scan the complete grid and compute the following score.

$$score = \sum_{i=1}^{X\max} \sum_{j=1}^{Y\max} \sum_{k=1}^{Z\max} MolA(i,j,k) \times MolB(i,j,k)$$

Pralay Mitra

I am putting them together here where the input is the protein complex of two subunits under consideration. Truly, you can extend it for others, for more than two subunits and output will be the surface matching score. So, what you need to do first is split the subunits from the complex assume they are *MolA* and *MolB*. Explicit splitting may not be required, if you can manage your algorithm to indicate which atoms are part of *MolA* and which atoms are part of *MolB* then together you can run that one also. Determine the grid size based upon the protein complex and use this *min* and *max*. Next, decide or take user input on the grid steps. Finally, digitizing the subunits *MolA*(x,y,z) and *MolB*(x,y,z) based upon the position of the atoms in the subunit.

Doing it is very simple. You need to compute the index position of grid cell i, *j*, *k*, and after that place 1 when an atom is present. Also, you initialize your three-dimensional matrix by all zeros so that the default will be 0.

What I say is in two dimensions you initialise all with 0 first after deciding the grid size and grid step. Assume that there is one atom that is placed here using some threshold you decided. It will be part of all these four cells. Now, this particular atom has some coordinate *x*, *y* and *z*. Now, it is in 2D if I remove this Z part. Please note that the current discussion can also be extended for three dimensions. You have *x* and *y*. You computed X_{min} and Y_{min}. You take the difference from X_{max} and X_{min} to know the grid step. Now, it is easy to determine the cell index *i* when you know *x*, X_{min}, and grid steps. Likewise, you replace it with *y*, Y_{min} and the grid step to index *j*. If you decide that along the X-axis grid steps will be something different than the Y-axis grid step and/or Z-axis grid step then you need to mention that. But if you decide on the same grid step, then it will be easy for you and I also feel that deciding on the same grid step along the X, Y and Z direction will be convenient. Then you can identify these four cells and once you will identify those four cells, then you give 1 1 1 1 here, corresponding to the red subunit. Iterate the process for the yellow subunit.

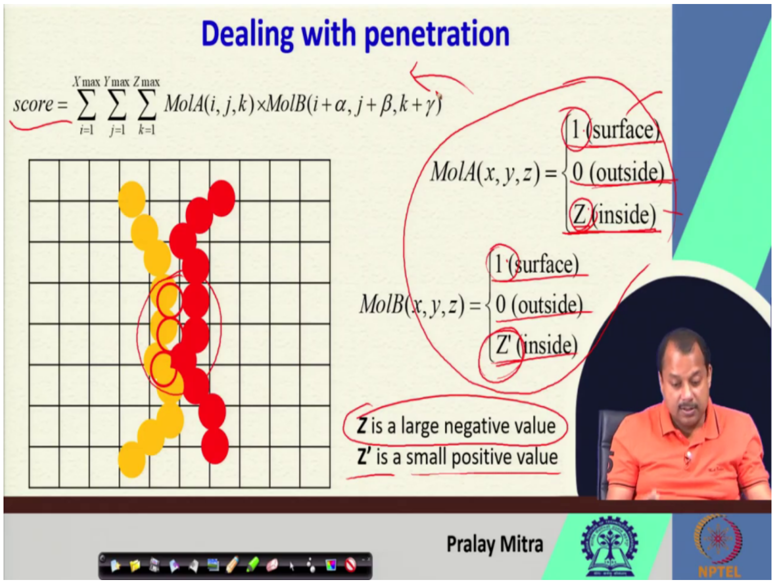
Nonetheless, you need to store those in two different arrays. It should not be on the same array. Else you lose the individual information that I need here for the multiplication. So, this *MolA* and *MolB* information must be there.

Once you have that one then you need to scan the complete grid to compute the following score. Here, *i* varies from 1 to X_{max}, *j* varies from 1 to Y_{max}, k varies from 1 to Z_{max}. These are the grid max. Multiply $MolA(i, j, k)$ and $MolB(i, j, k)$, if both are 1 then they will contribute 1 grid cell if one is 1 (on the surface) another is 0 (outside the molecule) then no contribution. Also, there will not be any contribution if both are 0 (outside) the molecule.

I understand you have a confusion. The confusion is if it is on the molecule, then it is 1, outside 0. What if it is inside the molecule, then also it is 1 if it is on the surface, then also it is 1. How do I discriminate between surface and inside when both are 1? How do I ensure whether 1 multiplied with 1 is coming from the surface or the core?

Remember, I started with protein complex, which means, there is no possibility that one protein molecule will go at the core of another protein molecule. This is called penetration. Penetration is not allowed in the protein complex. But if that kind of situation happens, then what do I need to do?

(Refer Slide Time: 27:20)



Dealing with penetration

$$score = \sum_{i=1}^{X\max}\sum_{j=1}^{Y\max}\sum_{k=1}^{Z\max} MolA(i,j,k)\times MolB(i+\alpha, j+\beta, k+\gamma)$$

$$MolA(x,y,z) = \begin{cases} 1 \text{ (surface)} \\ 0 \text{ (outside)} \\ Z \text{ (inside)} \end{cases}$$

$$MolB(x,y,z) = \begin{cases} 1 \text{ (surface)} \\ 0 \text{ (outside)} \\ Z' \text{ (inside)} \end{cases}$$

Z is a large negative value
Z' is a small positive value

Pralay Mitra

Let us discuss this example. In this case, the red circle, which is not filled, is going inside the yellow. Truly, it is not an experimentally valid structure since penetration is not allowed. Nevertheless, in computational modelling, it may be possible. Thus, you need to redefine the function *MolA* and *MolB*, where 1 is on the surface 0, is outside, Z is inside, and Z prime is inside. Therefore, 0 remains the same, 1 is divided into two parts one is on the surface and another is inside.

When it is on the surface, then I am assigning 1. When it is inside I am assigning some value Z and Z prime. What are Z and Z prime? I shall come to that. If I have this one, then probably this 1 multiplied with 1 will retain and you will not object with that one. Now, what is Z? Z is a large negative value and what is Z prime that is a small positive value. With this, if you go by the modified (one small modification I have done alpha beta gamma I incorporated because there is a penetration) previous equation.

How far the penetration can go that I have to mention and that is why I included that alpha, beta and gamma in one molecule and because of that incorporation when I multiply this 1 with this Z prime. This is a small positive value you can choose a fractional small value that will not make a significant change in the score value. On the other hand, this Z is a large negative value and if this Z is multiplied with this one of *MolB* then what will happen is a huge subtraction because Z inside this one is a large negative value.

If I consider Z= -100, you see 100 grid cells of overlapping is erased out due to one small penetration. Therefore, you consider it as a penalty because this penetration is not allowed. With this, you can deal with both surface outside and core. To decide on whether one atom is on the surface or at the core, you have to check what is its neighbour. If it is on the surface then at least one neighbour on the grid cell the value is 0, if it is not then all the neighbours themselves will be 1. Using this we can decide or discriminate between the surface and inside and also the outside.

Now, we can see one real-life application of this simple score value which says that computationally if I model two protein molecules and check their different fittings, say if this is one protein molecule, this is another protein molecule. So, whether this is a fitting, this is a fitting, this is a fitting, this a fitting and all other possible fittings - I need to calculate the score to select the best fitting. For that surface matching score will be simplest to calculate and useful. During this modelling, I cannot rule out the situation where there will be a penetration – that surface matching score definition also supports it.

During algorithm design, I can use this function along with this one to penalize the penetration and to allow only the valid orientations or valid complex formation. Hence, it can be utilized for protein docking algorithms where given two protein molecules you need to decide which orientations are feasible. We shall extend that in our next lecture. Thank you.