

Algorithms for Protein Modelling and Engineering
Professor Pralay Mitra
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur
Lecture: 25
Structure Alignment Measures

Welcome back. So, today in this lecture, I am planning to introduce to you one measure to check whether two structures are same or not. So, when say we are discussing say protein folding or say we are talking about protein complex modeling. Then the question will come specifically in case of protein folding that.

So, one protein was there structure and sequence and you designed one protein folding technique. So, using that one you have another structure. How do I know whether these two structures are same or not what is the measure? So, if I go by this then first thing is coming that given two structure first you need to align and after aligning those two structure.

Then you have to check that whether they are same or not. But today in this lecture first I will discuss that one simple measure for checking whether two structures are same or not. Then, in the next week in the next lecture, I will discuss the structure alignment and other structure alignment measure techniques.

(Refer Slide Time: 01:34)



CONCEPTS COVERED

- Structure alignment
- Root-Mean-Square-Deviation (RMSD)

Pralay Mitra

KEYWORDS

- > RMSD
- > TM-score

Pralay Mitra

The simplest one that is the root mean square deviation measure that we will discuss today.

(Refer Slide Time: 1:44)

Structure Alignment

Pralay Mitra

Given two structures, first you need to do that alignment, why there is a need of alignment because I need to know whether these two structures are same or not or what portion of that structure is known, same or not. Why it is simple reason is that we know that there is a direct relation between structure and function.

So, the simplest application you can think and very very useful application you can think there is one structure which is deposited in say protein databank. It is annotated, a lot of experiment has

done or say a lot of inferences regarding its function has been taken and there is a database, it is published, everything is available regarding the function of that particular protein structure.

Experimentally you design some protein structure or computationally using a protein folding technique you design one protein structure. Then the question will come that what is the function of the structure? If say it is experimental, then you have to go for extra experiment, but if it is a computational, completely. So, you have no clue that first of all whether it will be stable or not. In vivo or in reality it is going to be stable or not.

If it is stable then what will be its function. So, you are completely clueless. In order to know that one the simplest way you can do is that you align with some known structure. Known structure means the structure is deposited in the protein databank, its function is known, it is annotated and then from there you infer.

So, alignment then if they are same, then you infer. That alignment I mentioned that I will discuss on the next week. But, if I assume the existence of one alignment technique, then given to protein structure. I align one with another and then after that alignment process, I will compute whether these two structures are same or not.

In this case you can see green is one structure green color, red color is another structure and they are almost identical, there is a little variation that you can see in some places, but almost they are identical. Now, that almost identical I am telling by looking at that structure, how to quantify that one because when you will be given with that one, you remember that PDB structure, file format. So, you will be given with a PDB file format.

(Refer Slide Time: 4:35)

Why do we need structure alignment?

Frame of reference is different.

Pralay Mitra

Why do we need structure alignment?

Frame of reference is different.

Pralay Mitra

And it can be that one structure is like this and another structure is like this. So, by looking at this you can understand there is a little rotation of the structure on the right hand side. If I rotated it back then perhaps it will be same and yes it is same. I agree. It is same.

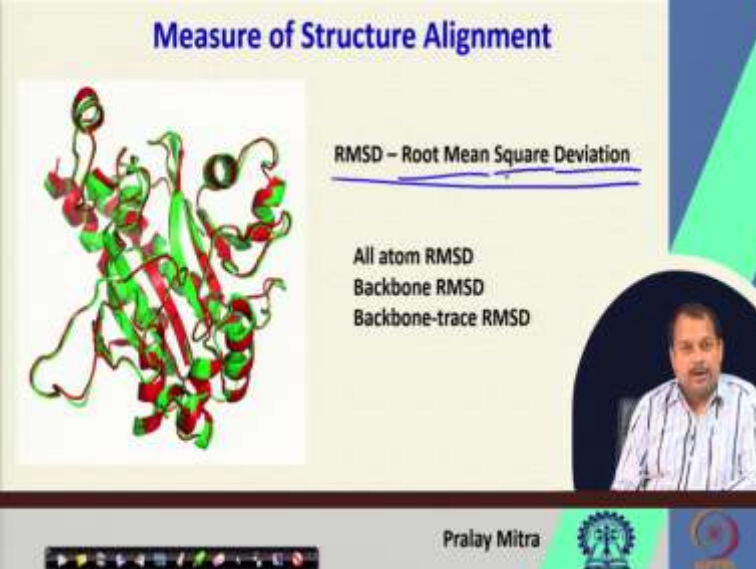
But when at the I am giving you the coordinates since it is rotated the coordinate will be different by looking at the coordinate it is not possible for me to say that it is just rotated or whether they

are same or not. So, for that you need alignment it may be easy and those two structures are same. These two structures are not very easy to infer that whether they are same or not.

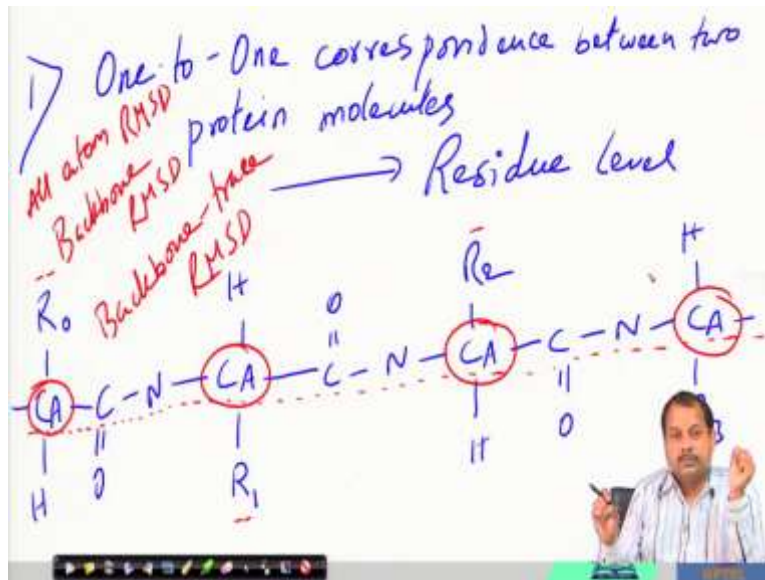
You have to go for alignment and also alignment is required for another purpose that purpose says that when two different sequences are there, but they are structured is same or sequences are identical, but they are structured is different. So, how to deal with that one? So, for that alignment is required.

So, that we will discuss later, but now we are assuming that structure alignment algorithm is in place, we got that aligned structure, we need to compute the measure, we need to compute how much similarity is there between these two structures.

(Refer Slide Time: 06:09)



The slide is titled "Measure of Structure Alignment" in blue text at the top. On the left, there is a 3D ribbon diagram of a protein structure, colored in shades of green and red. To the right of the diagram, the text "RMSD – Root Mean Square Deviation" is underlined. Below this, three types of RMSD are listed: "All atom RMSD", "Backbone RMSD", and "Backbone-trace RMSD". In the bottom right corner, there is a circular inset video of a man with a beard, wearing a light blue striped shirt, speaking. At the bottom of the slide, there is a navigation bar with several icons and the name "Pralay Mitra" next to a logo.



So, for that, we are going to define the root mean squared deviation that is the measure that we are going to define first. So, when we are defining root mean square deviation. So, there is a deviation. So, which means deviation of one point with respect to another point. Then square, taking the square, so you can very much guess that it is going to be an equilibrium distance deviation, then root mean that is known to you.

So, that is why the full name root mean square deviation. Now, when we are looking for this root mean square deviation, then even after the alignment few things are required first of all you need one to one correspondence between two protein molecules of course, at the residue level if that correspondence is not there then how do you compute that whether they are deviating or not.

So, one to one correspondence is required that is the first requirement. Next requirement is mentioned here all atom RMSD, backbone RMSD or backbone trace RMSD what are the differences? Let us start with the protein amino acid represent itself. So, we know CA at the core, H then say side chain I am writing for the timing as R₁.

Next this and then C, O, N, CA then C, O, N, CA that way I am going or say better instead of correcting this one you can erase this and you can write this N, C, O, CA that way it is going. So, now, what is backbone here? Backbone indicates this C alpha C N, C alpha C N, C alpha C N, C alpha. This is backbone.

Which one is side chain? This R1, R0, R2, R3 that is side chain. So, when I say that one to one correspondence between these two protein structures and known to you then that this C alpha corresponds to which C alpha in another structure this C alpha corresponds to which C alpha in another structure that way C alpha to C alpha correspondence is there.

Then that from this C alpha to that C alpha what is the deviation that you compute after the structure alignment not before the structure alignment clear. So, when you are doing now you see that lot of atoms are there based upon these three different variations are their first one is all atom.

So, when you have correspondence between C alpha this C alpha say C alpha apart 0 to C alpha of another protein say R0, then the deviation of C alpha to C alpha H to H and in R0 whoever the atoms are they are corresponding to each atom I will have the correspondence. So, once you have, my point is that once you have C alpha to C alpha correspondence then rest of the side chain and other atom correspondence you can also establish.


So, that way I will establish only the C alpha to C alpha correspondence and that is nothing but the amino acid to amino acid correspondence. So, once amino acid to amino acid correspondence is there then accordingly the atom level correspondence will also be done by you. You can do that one. Now, if you consider the deviation of all the atoms then that is going to be the all atom RMSD. Now, you remember that while doing the protein folding problem I mentioned that primarily what is important to us is only the fold or only the structure of the backbone.

Backbone means this dotted red line if you follow that one that is important to us that side chain hydrogen that we can add later if you are also with me and like-minded, then you may think. Instead of all atom RMSD where there might be a little more deviation possible, what I can go for is that backbone RMSD that is following the dotted line whereas all atom RMSD is calculating the deviation of all the atoms, even sidechain, hydrogen, et cetera whoever is present.

Backbone tells only C alpha C N, C alpha C N, C alpha C N not trace. So, that is the backbone RMSD. Third variation will be that backbone trace which is considering all the C alpha atoms. So, three variations based upon how many atoms I am considering whether I am considering all the atoms whether I am considering atom C alpha C N, C alpha C N that is the backbone atoms or I am considering only the C alpha atoms.

(Refer Slide Time: 13:20)


Measure of Structure Alignment




RMSD – Root Mean Square Deviation

- All atom RMSD
- Backbone RMSD
- Backbone-trace RMSD

Pralay Mitra




Measure of Structure Alignment




RMSD – Root Mean Square Deviation

$$RMSD = \sqrt{\frac{1}{N} \sum_{i=1}^N \delta_i^2}$$

- All atom RMSD
- Backbone RMSD
- Backbone-trace RMSD




Pralay Mitra



Structure Alignment

Identify the residue/atom level correspondence:

CA (x_{1i}, y_{1i}, z_{1i})



$$\delta_i^2 = (x_{1i} - x_{2i})^2 + (y_{1i} - y_{2i})^2 + (z_{1i} - z_{2i})^2$$

Pralay Mitra

So, based upon that one three different variations will occur and those are listed here all atom RMSD, backbone RMSD, backbone trace RMSD. Now, the actual equation for this RMSD; RMSD equals to square root of 1 by N summation over i equals to 1 through N delta i square. So, that is the deviation. Now, how much deviation is how the deviation is calculated?

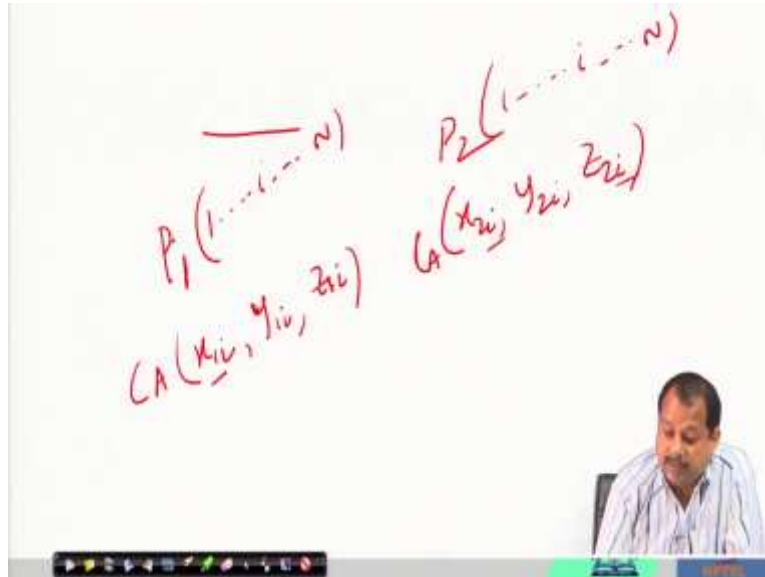
So, by taking the Euclidean distance between two points. So, before that one again I have to identify the residue or atom level correspondence. So, what I am suggesting that residue level is important not that atom level because atom level will come automatically. If you assume that. So, I am aligning two protein structures and at residue level it is a kind of say leucine is aligning with isoleucine.

So, at the sidechain level definitely some atoms mismatch will be there. In that case actually you cannot go for atom level calculation then you have to focus on backbone level RMSD or backbone trace RMSD. So, identify the residue level correspondence, atomic level correspondence will come automatically if two structures are same or homo if they are not then I have to go for backbone level then also the residue level is enough.

So, atom level correspondence is not required residue level correspondence is enough for me. And this is the equation. So, delta square is x1 i minus x2 i whole square plus y1 i minus y2 i whole square plus z1 i minus z2 i whole square. Where x1 i is the ith residue of the first protein, x2 i is the ith residue of the second protein, y1 i ith residue of the first protein, y2 i ith residue of

the second protein, x_1 i th residue of the first protein, x_2 i th residue of the second protein. So, this x_1 so, you can assume that C alpha is x_1 i .


(Refer Slide Time: 16:22)



Structure Alignment

Identify the residue/atom level correspondence:

$CA(x_{1i}, y_{1i}, z_{1i})$



$$\delta_i^2 = (x_{1i} - x_{2i})^2 + (y_{1i} - y_{2i})^2 + (z_{1i} - z_{2i})^2$$

Pralay Mitra

So, this says that two proteins are there say P1, P2 and C alpha is x_1 i , y_1 i , z_1 i this is C alpha x_2 i , y_2 i , z_2 i . So, for P1 there are say 1 i N and 1 i N, N number of residues are there. So, in both cases and there is a correspondence. So, i th residues, i th residues, i th residues all first protein i th residues of the second protein that way I am computing.

Now this i you can consider that it is not ith residues, so it is the ith item, it is fine. So, only thing is that you need to have the correspondence and that correspondence accordingly you have to calculate this one.

(Refer Slide Time: 17:30)

Measure of Structure Alignment


Algorithm 9:

Input: Two protein structure whose alignment and residue-level correspondence is known.

Output: Their RMSD measure


Steps:

1. $\text{error} = 0.0;$
2. For each pair of residues (one from each protein molecule)
 - a. Compute their δ^2
 - b. $\text{error} = \text{error} + \delta^2$
3. Compute the RMSD



Pralay Mitra


Measure of Structure Alignment



RMSD – Root Mean Square Deviation

$$RMSD = \sqrt{\frac{1}{N} \sum_{i=1}^N \delta_i^2}$$

All atom RMSD
Backbone RMSD
Backbone-trace RMSD



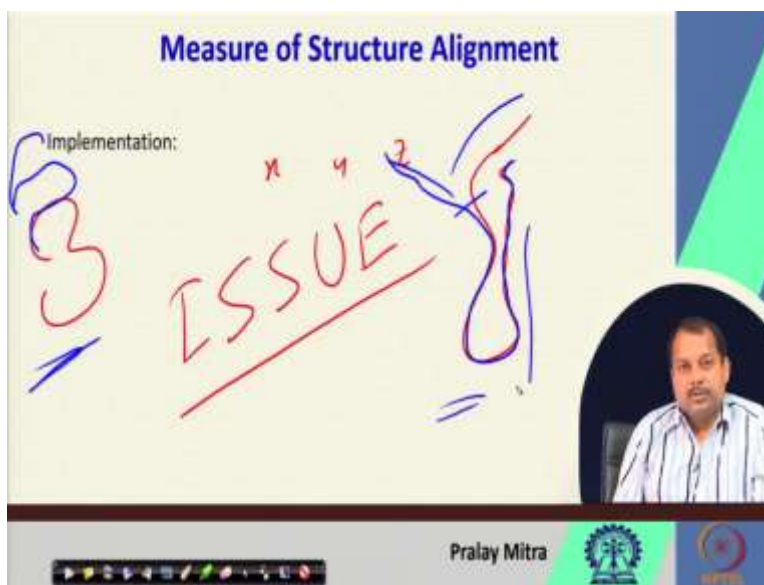
Pralay Mitra

So, the algorithm is very simple for computing the RMSD of two structures. So, input is two protein structures whose alignment and residue level correspondence is known to you. Output they are RMSD measure. So, step one so first so, there is a type two here. So, you may ignore that this one. So, error you can assume first as 0. Now, for each pair of residues one from each

protein molecule compute their delta square then error equals to error plus delta square the deviation square you are computing.

And they at will give you the, that will give you this summation part, this part you will get this part. So, after that one what do you have to do once you will come out then you divide this error by N and take the square root you will get the RMSD values. So, that is the simplest algorithm. But it is very important in order to know or check whether two in order to check that whether two protein molecules are structurally same or not.

(Refer Slide Time: 19:03)



Now, if I go a little ahead then how do you implement this one very simple PDB you have atomic coordinate information and the residue information. So, x, y, z that information is there. In an introductory class also I mentioned the data structure where you can store this information atom level and residue level information as the structure used to that one.

Only thing you will need to write that you have to take that Euclidean distance between two points, the correspondence is also known and that is given as an input. So, Euclidean distance between these two points take the summation over that one divided by the number of points you computed and after that one take the square root it is very simple.

However, there is some issue with this one you recall the RMSD formula. In the formula, there is a summation of the deviation. So, (())(20:11) summation you are taking and dividing by the

number of atoms. Now, you assume a situation where say one structure is something like this and another structure is although they are placed side by side.

But you can assume that this is alignment or so, I can write again let us assume this is the alignment of two structures. Now, in this case, you see that two structures are almost same except in these two regions and in that region also red portion is very small, but blue portion is large. And if I assume that, so, there is as of now, we are considering only a good correspondence.

So, both are same. Now, you see that only a fraction or part of the region which are deviating too much, because of that one the RMSD value will be very high. So, it will suppress the information that the two structures are very similar in nature except in some tail region and that it is true that during the protein folding process or the nature of the two terminal region N terminal or C terminus is little bit floppy compared to the overall core of the protein.

So, because of that the total RMSD value will be very high in nature and also from the equation it is not that much easy to infer that whether definitely when an RMSD high indicates that there is not much similarity between two structures. But how much the similarity is there? Or what is the good value for RMSD? Again zero is correct. But if it is not zero, it is nonzero whether less than one less than two, which one is correct?

So, it is very difficult to say. Because the RMSD value is not normalized, normalized in the sense that its value does not vary between one range and higher RMSD value, what is the structure? But how what is? Say when it is one situation say this two blue and red is one situation and say another situation I am drawing here.

So, these two structures clearly you see on the left hand side what I have drawn there is no structural similarity and this structure you got after that alignment you assume and that is correspondence, the end the correspondence also say there are both are with the same length say 100 and 100. So, that is why I got a correspondence 1 to 1, 2 to 2, 3 to 3 here and in this case on the right hand side say the length of the protein structure is say 1200, 1200 amino acids are there out of which say 800 is aligned and 400 is out of range.

So, here 400 is out completely out here and 800 is here on the other end here 100 and 100 lengths and you see that only 10 percent is kind of aligned, but rest is not. So, if you compute the RMSD whether the RMSD value has a capability to tell that this structure is bad, this is bad, but this is not that much bad. So, is it possible? So, it is not. So, that is one of the measure bottleneck for this measure of structure alignment, specifically when we are dealing with different four level or two structures are not same.

So, that is one problem. So, that problem we will discuss on the next week, but here why we are discussing this RMSD because you understand that algorithm is very simple, implementation is very simple and when we are dealing with the same protein then this is the best. Why should I go for say complicated structures? So, same protein means that, one protein structure exists I took the sequence, I said developed on protein folding problem that protein folding problem gives me some structure then I am aligning and computing the RMSD.

If I am doing that one then RMSD is good enough for me, when two structures are same, but when two structures are varying largely or partly matching, partly varying and they are from two different structures, well again. So, another limitation or issue with the RMSD competition is that when two sequences are different, then you have to go for sequence alignment first.

And after that sequence alignment, you have to go for structural alignment, after that structural alignment, if you compute the RMSD then that when you are doing the sequence alignment, so, you are optimizing that alignment score again that we will discuss in our next class. We are optimizing the alignment score and when we are optimizing the alignment score, then there are multiple possibilities and accordingly there will be multiple RMSD values.

If that is there, then what will be the conclusion? So, as such although it is very simple algorithm very easy to implement and there are a lot of applications also specifically when you are dealing with only one protein structure only one protein structure mean that you are say comparing with the known structure and both are same. So, I mean the model structure and the benchmark structure both are same, if it is the case, then RMSD is good enough and you can work with that one.

So, in that case, if you have say millions of data, though, it is not possible, but if I assume that there is no limit on the number of data, but if you are comparing the benchmark data and the

model or the structure that you have predicted from the benchmark data, if that is the case, then RMSD is good enough and you can use that one. But if it is not the case, if two structures at the sequence level may vary, but still you wish to know whether they are structurally same or not why?

As I mentioned at the beginning of this lecture that I got one structure whose function is not known, I wish to annotate that one. Now, the best way is that you check that which structure it is aligning or it is fully same as which structure. If you identify that one then it will be easy for you to infer that structure that structures function and mention that this is also the function of this particular new structure. So, for that you need to go for some other technique that we will discuss on the next week.

(Refer Slide Time: 28:15)



So, what is remaining now is we have to give one technique which will give me the correspondence between two sequences, when two sequences are not known. Then we have to come up with some alignment technique. And next, the issues or limitations of the structural measurements score like RMSD, which does not tell anything about the similarity, four level similarity specifically of two structures, whether we can address that one or not. So, those things we will discuss in that next week. Thank you very much.