

Algorithms for Protein Modelling and Engineering
Professor Pralay Mitra
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur
Lecture: 43
Protein Design Energy Function

Welcome back. So, we are continuing with protein design and in this lecture we will discuss the energy function of the protein design.

(Refer Slide Time: 00:24)

The image shows a screenshot of a presentation slide. The slide is divided into two main sections: 'CONCEPTS COVERED' and 'KEYWORDS'. The 'CONCEPTS COVERED' section lists four items: Protein Design, Evolutionary energy, Physics-based energy, and Combined energy. The 'KEYWORDS' section lists two items: Protein Design and Energy function. The slide also features a video feed of Professor Pralay Mitra in the bottom right corner. The slide includes a navigation bar at the top and bottom with the name 'Pralay Mitra' and logos for IIT Kharagpur and NPTEL.

CONCEPTS COVERED

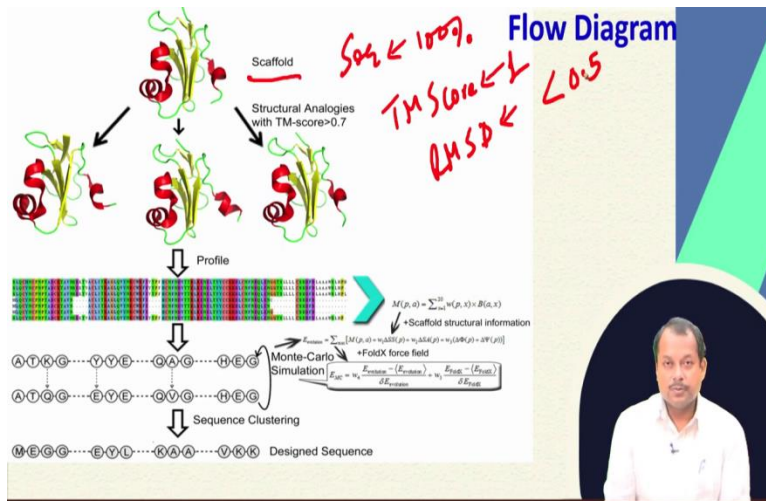
- Protein Design
- Evolutionary energy
- Physics-based energy
- Combined energy

KEYWORDS

- Protein Design
- Energy function

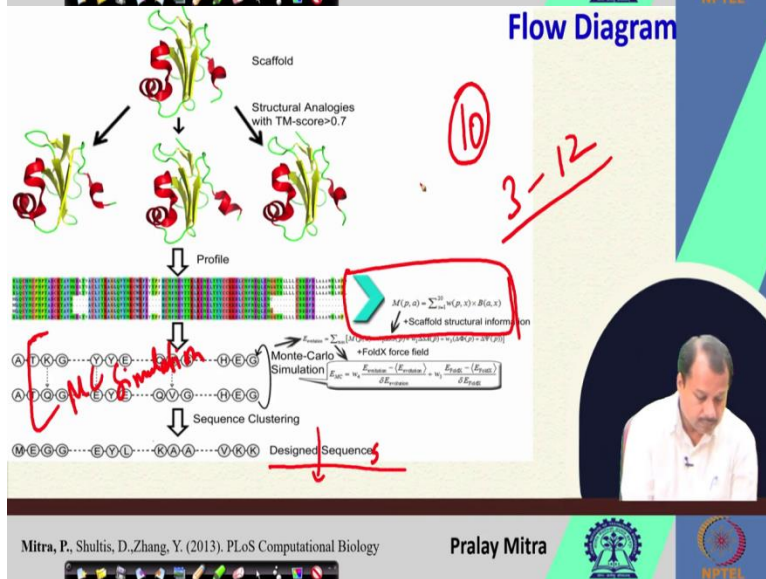
Pralay Mitra

Pralay Mitra



Mitra, P., Shultis, D., Zhang, Y. (2013). PLoS Computational Biology

Pralay Mitra



Mitra, P., Shultis, D., Zhang, Y. (2013). PLoS Computational Biology

Pralay Mitra



So, the concept we will be covering protein design and in the protein design the evolutionary energy Physics-based energy and the combined energy those we will discuss and what is the effect of this energy on the accuracy of the result or on the effect of the new design sequences that we will also discuss.

So, the key word is protein design and the energy function. So, this is the overall flow diagram of the protein design algorithm that I was discussing on the last lecture. So, input is the protein structure which I am calling as a scaffold then what we are doing that we are looking for homologous structures from the protein databank and we identified few, but while we are doing this in order to keep our design without bias.

So, what we will do that if we get some structure whose identity or say a sequence level identity is 100 percent and structurally they are perfectly aligned say TM score is 1 and say RMSD is less than 0.5 or something, then we will exclude those structures. Because you see that we are doing multiple sequence alignment from this multiple sequence alignment.

We are computing the position specific scoring matrix in order to guide our simulation and also in protein design algorithm our main intention is to design a new protein sequence so, which should not be biased by the input protein sequence so, that is why we will remove those cases in order to avoid the bias apart from that one whatever the structure we will get.

So, that we will use for the multiple sequence alignment now, from that structure here on the right hand side you can see that this profile we computed after computing the profile so, we also computed the energy function the detail of that energy function we will discuss shortly, but just it is keeping that energy function detailed right now.

So, then what we did we started from one random sequences and then we go for some Monte Carlo simulations technique, this is my Monte Carlo simulation. So, in each step randomly we picked some random positions and we mutate by some random amino acids, but while mutated by the random amino acids, we will try to exploit the profile information that we created by aligning the homologous structures.

So, after that aligning alignment of the homologous structures, we got the multiple sequence alignment, but please note it down that when we got these multiple sequence alignment, so, this sequence alignment is not the sequence alignment of the homologous structures, it can be very much thus after the structural alignment, what is the best sequence alignment you got it can be that one also.

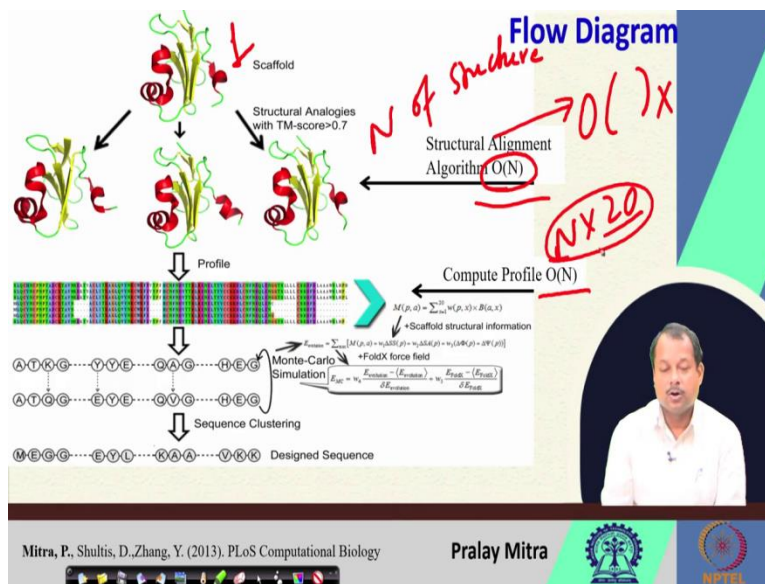
So, both way it is possible. But if you go with the structural alignment and from there you got the sequence alignment then that is more towards the convergence of the protein design algorithm rather if you just look for the sequences, because when I will align the sequences of even for the homologous structures.

Then it will align or it will optimize the sequence alignment whereas if I align based upon the structure and then pick the sequences directly from there, then it will align the structures. Now,

during this Monte Carlo simulation process a number of sequences has been accepted. Those sequences will be analyzed based upon some clustering technique that we have discussed on the last lecture. And finally, we will output the design sequence. Now, this design sequence may not be one so I can very much add one (04:56) here to make it plural.

So, that I will output a number of design sequences maybe the all the clusters I will output then based upon the existence of how many clusters will be there, this number of design sequences will also vary it can be a varying from something 3 to say 12 that is my understanding and if it is so, you can also very much keep it say around 10 at most 10 so, living 11 and 12 out so, what you can then say that I will output 10 best design so, where the confidence will be the first rank is with the highest confidence next the second rank that way it will go.

(Refer Slide Time: 05:48)

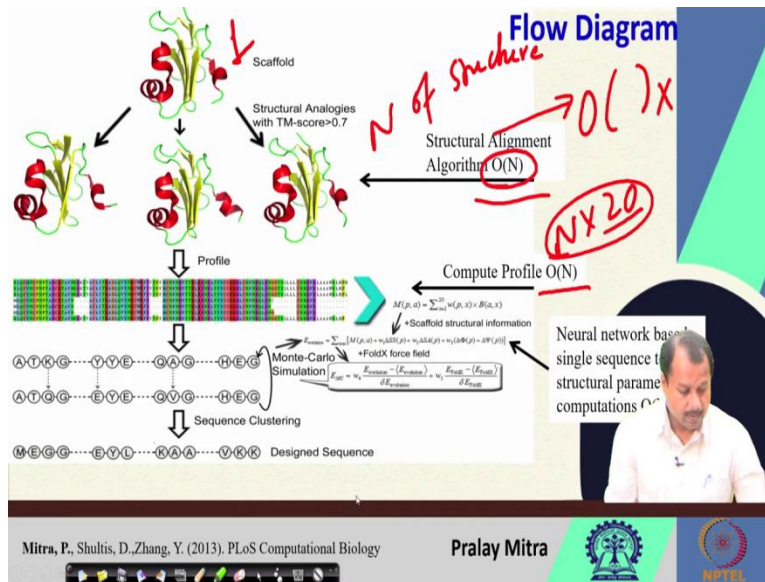


If I look at the computational time requirement corresponding to the structure so, let us analyze the competition complexity first. So, first of all when we are doing the structural alignment that alignment technique is having its own complexity, but here this order of N indicates this order of N indicates that for this 1 and if in the PDB there are, N number of structures then I have to perform a linear operation.

However, each structural alignment based upon which algorithm you are using may have its own complexity which will be multiplied with this order of N. Now, after the alignment is done then the profile computation part will take all the order of N although I am computing a matrix and as

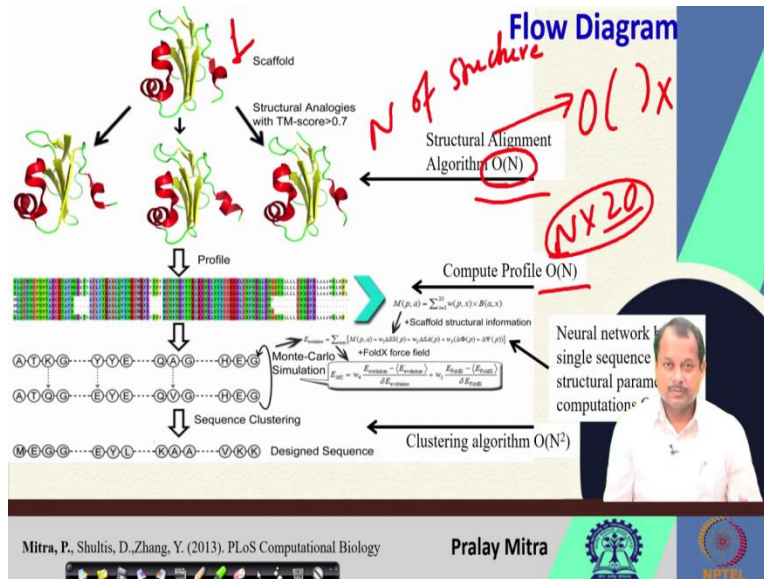
I mentioned that it is N cross 20 but you see this 20 is a constant and since it is a constant so, computing that matrix you can consider that it is going to be order of N and 20 is fixed always because I am considering only 20 amino acids.

(Refer Slide Time: 07:14)



Next neural network based single sequence to structural parameter computations. So, during this energy function we will see a number of features we need to calculate but most of the features are already pre computed or I should not say that pre computed most of the features. So, what that so, I have I can have one neural network technique which will train on the known protein structures and it will have the model next what you have to do that for that particular sequences you can predict all the features just by looking at the model. So, in one scan you can do that one that is what is mentioned here.

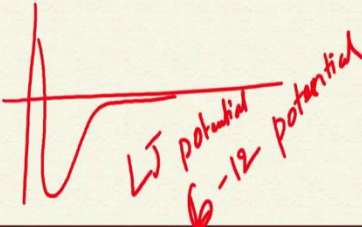
(Refer Slide Time: 08:06)



Next clustering algorithm will take order of N square but here please note it down the, although it is always order of N but this N will vary. So, for say structural alignment or drop N will be the size of the protein databank for compute profile, the N is going to be the length of the protein sequence neural network based sequence through structural parameter that is also going to be the length of the sequence or length of the input protein and for the clustering algorithm this N is going to be the number of sequences which are accepted multiplied with the number of sequences erase order of N square. So, it is the number of sequences which are accepted during the Monte Carlo simulation steps.


(Refer Slide Time: 09:01)

Physics based Energy function

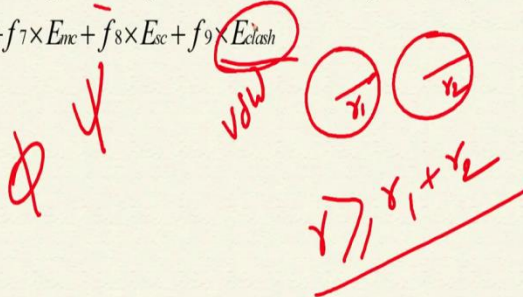
$$E_{phy} = f_1 \times E_{vdw} + f_2 \times E_{solvH} + f_3 \times E_{solvP} + f_4 \times E_{wb} + f_5 \times E_{hb} + f_6 \times E_{el} + f_7 \times E_{mc} + f_8 \times E_{sc} + f_9 \times E_{clash}$$


LJ potential
6-12 potential

Pralay Mitra




Physics based Energy function

$$E_{phy} = f_1 \times E_{vdw} + f_2 \times E_{solvH} + f_3 \times E_{solvP} + f_4 \times E_{wb} + f_5 \times E_{hb} + f_6 \times E_{el} + f_7 \times E_{mc} + f_8 \times E_{sc} + f_9 \times E_{clash}$$


ϕ ✓
vow
 $r > r_1 + r_2$

Pralay Mitra



Physics based Energy function

$$E_{phy} = f_1 \times E_{vdw} + f_2 \times E_{solvH} + f_3 \times E_{solvP} + f_4 \times E_{wb} + f_5 \times E_{lib} + f_6 \times E_{el} + f_7 \times E_{mc} + f_8 \times E_{sc} + f_9 \times E_{clash}$$

Handwritten notes:
~~NN~~
~~Exponential~~
 $f_1 \leftrightarrow f_9$
 weight factors

So, the first one is the Physics-based energy function which can include a number of terms using linear combinations. So, here if you look at the terms then first one is the vdw or vander waal interaction. So, the basic of the vander waal interaction is that plot that we also discussed in the context of protein docking.

So, it is LJ potential or say 6-12 potential because there are two terms 12 and 6. So, A divided by R to the power 12 minus vd divided by R to the power 6. So, 6 that R to the power 6 is the attractive term and R to the power 12 is the positive term. So, based upon that it is the vander waal interaction.

Then salvation energy as the hydrophobic residues salvation energy for the polar residues, then existence of water molecules hydrogen bonding then electrostatic interactions then here are the main chain interaction then side chain mc is main chain sc is side chain and this clash. So, what is this clash? Let me explain in detail.

So, the clash indicates that during the protein design or say, any sort of modeling when say we are developing algorithm for protein structure modeling in general then there is a possibility that we will generate some positions which are not following the steady clashes. What is that steady clash? So, let us assume that there are two atoms. So, with radius r_1 and say r_2 now, it is true that atoms are not hard sphere but throughout our discussion on in this course, we are assuming that the atoms are hard spheres with its radius are vander waal radial.

So, if it is so, then r_1 and r_2 corresponding to two atoms and they should be r distance apart where r is greater than equals to r_1 plus r_2 . Because I am considering those as a hard sphere and they cannot penetrate with each other. Now, it may possible that during the modeling, we are placing the atoms in such a way that eat violets this one r is not greater than equals to r_1 plus r_2 sometimes r is less than r_1 plus r_2 if that is the situation then we are declaring that as a clash and that clash will give some penalty in my scoring function.

Similar to that although it is not included here, but you remember that as per that Ramachandran plot so, ϕ and ψ angles in the protein main chain is supposed to follow some pattern or some allowed reasons are there in the ϕ and ψ angle. So, since these are angles so, theoretically 360 degree possibilities are there and trust me during the modeling any values starting from 0 to 360 may generate.

But, that Ramachandran plot is considered as one of the say criteria for a good protein structure or a vary protein structure. So, if there is one protein structure you got either experimentally or computationally through some modeling, but it is not following the Ramachandran plot which means that the white region or blank region as for the Ramachandran plot which is the formidable region where some ϕ ψ value may not a should not go but as per your say structure it is going to that position then that is not a good structure.

So, you can also consider that during your energy parameter. Now, since all the parameters here you are considering say starting from vander waal interactions always an energy of hydrophobic or polar residues then hydrogen bond the electrostatic introduction main chain main side chain or clash information all are dealing with some physics.

So, this you can consider as that physics based energy function. Now, you can see that each term definitely as per my definition each term will have some mathematical equation we are not going into details of that one for the time being, but each term is the linear combination. So, each term is linearly combined in order to get my final Ephy or final score function.

So, that is why during this linear combination I am having if f_1 through f_9 different weight factors. So, these weight factors. I can either you can learn from some neural network technique learning process or you can define empirically by looking at that protein structure, but this I will not suggest this equation itself is empirical.

So, one good suggestion could be so, this is not suggested. One good suggestion could be that once you have the mathematical expression for each of the terms starting from EVDW, ESolvH through a clash then corresponding to all valley structures which are present in a protein databank you can compute those terms and then from there you can learn your f1 f2 f3 up to f9 using some learning technique or some neural network technique. So, that is one which is called as a physics-based energy function.

(Refer Slide Time: 15:32)

Evolutionary Energy function

30000

$$M(p, a) = \sum_{x=1}^{20} w(p, x) \times B(a, x)$$

$$E_{evo} = \sum_{\max} [M(p, a) + w_1 \times \Delta SS(p) + w_2 \times \Delta SA(p) + w_3 \times (\Delta \phi(p) + \Delta \varphi(p))]$$

torsional angles.
 ↓
 Solvent-accessibility
 ↓
 Secondary Structure
 ↓
 PSI Prod
 ↓
 PSS Prod
 ↓
 30K X Lin

Pralay Mitra

Neural Network Technique

SS, SA, TA

30K X 1 sec

Pralay Mitra

Next in the evolutionary energy function although I am calling it as evolutionary information because it is including so, four terms which are very much related to evolution. So, that is why I am calling that as evolutionary energy functions which are then one is coming directly from the structural alignment and which structures homologous structures.

So, some sort of evolutionary information is already there. So, that matrix that we computed it will be here after that one we are having SS which means secondary structure, solvent accessibility and torsional angles. Now, this delta indicates the deviation, deviation from what deviation from a true value now, you see the situation.

So, please listen it carefully I am doing protein design which means during my simulation at each intermediate stage or at the i th step what I am generating is a sequence it is not a structure whereas, the terms the secondary structure solvent accessibility torsional angles are very much related to the structure.

\Then the question is when I am generating the sequence then how can I compute the secondary structure solvent accessibility or torsional angles from the sequence. So, for that definitely you have to come up with some sequence based prediction system because it is not advisable that during a Monte Carlo simulation methods one you generate some temporary or intermediate sequences then you go for protein folding algorithm.

In order to model that sequence to a structure then compute the secondary structure or say torsional angle or solvent accessibility fortunately, some techniques used for secondary structure prediction one is called as the PSS pred another is called as PSI pred but, both are computing intensive.

So, they run sequence alignment and a lot of after doing that a lot of sequence alignment through the blust, that blust software they compute one temporary scoring matrix using that scoring matrix that PSS something like PSSM scoring matrix then they compute that what is going to be your secondary structure.

Now, that takes about a minute sometimes based upon the sequence otherwise several seconds I cannot afford that one also because in my simulation if I assume that so, I am running say for example say 30,000 steps. So, 30,000 multiplied with so, if I assume 1 minute for predicting the

secondary structure and say for the solvent accessibility torsional angle, et cetera. So, if that much amount of time it takes then you can see that it is going to be computationally intractable. So, it will take huge amount of time and honestly speaking I cannot afford that one. So, I have to modify this PSS pred also I prayed so that it will take a single sequence.

Maybe it will have some neural network technique through which it has learned and developed one model that model is there now what he will do that during your Monte Carlo simulation. The moment you got one design sequences you take the design sequence use that model in order to get a prediction of secondary structure solvent accessibility and torsional angles very quickly in few seconds if not in 1 second.

So, if it is then also assuming one second you are taking. So, this much time is required so, 30,000 seconds only for this purpose apart from that one other jobs like generating the random number then computing the score function then going for the condition of checking whether to accept that sequence or not then housekeeping work like.

So, you have to accept that sequence you have to store that sequence you have to store that energy function once the simulation is over, then you have to go for clustering before the similar you have to pre process that structure in order to generate that position specific scoring matrix combining those it will take huge amount of time. So, which is not advisable, so, you have to be very careful so, that this 1 second will be the maximum or should not be more than 2 or 3 seconds. So, computation time is also very crucial for our purposes, it is not only computational complexity.

(Refer Slide Time: 21:44)

The slide is titled "Combined Energy function" in blue text. It features handwritten red equations and annotations. The top equation is $E_{MC} = W_4 \times E_{evo} + W_5 \times E_{phy}$, with E_{MC} circled and a checkmark next to it. Below this is a more complex equation: $E_{MC} = W_4 \times \frac{E_{evo} - \langle E_{evo} \rangle}{\delta E_{evo}} + W_5 \times \frac{E_{phy} - \langle E_{phy} \rangle}{\delta E_{phy}}$. A red arrow points from the top equation to this one. At the bottom left, the text "i-th stage" is written in red and underlined. To its right are the symbols " \angle " and " γ ". In the bottom right corner, there is a video inset of a man in a white shirt. The slide footer includes the name "Pralay Mitra" and logos for a university and NPTEL.

What you can do that you can combine these two scoring function physics based and evolutionary information based in order to generate your final scoring function and while you are combining what you can do that this is your score function this is your score function. So, one simple way to combine in case of a simple way of doing this is E MC.

So, MC stands for your energy function for Monte Carlo simulation. So, you can do say $E W_4$ prime multiplied with $E E_{evo}$ plus W_5 prime E_{phy} . So, this is simple linear combination again similar to the way we combined the physics based energy function, we combined the evolutionary based energy function.

We use this similar to the way we did it for our protein folding problem also in protein folding problem if you remember that the energy function was the linear combination of a number of features and that is what is the first one and that way I can think that it is kind of a positive design. Which means that always I am trying to optimize this because, I am computing evolutionary energy physics based energy I am taking a linear combination of them and always I am trying to optimize or say in this case I can consider that I am trying to minimize this energy function.

Another situation which can give you a quick or better convergence is if you combine along with this a kind of G score concept. So, what I am doing now is that as of now say I am at i th stage. So, i th stage means that the first situation says that if I for the timing just ignore the Metropolis

criteria, then i indicates that this energy is going to be or I am expecting that this energy is lowest among i minus one step or so, it is the lowest compared to first to i minus one step and that is then only I am going to accept if I am assuming that Metropolis criteria is not in place.

Now, I am assuming that one and based upon that I am accepting that is my first case. Next, in the second case what I am doing so, for all the accepted cases, I computed the mean so, that less than greater than indicates the mean and the delta indicates the standard deviation. So, for all of them I computed the energy function, I am taking the mean and standard deviation of them.

Then I am combining this way so that the energy is further reduced by the mean of all the accepted one and scaled by the standard deviation of the all the accepted one then I am combining that one and finally I am getting EMC. So, some studies indicate that this the second approach may give you an age when you wish to go for quick convergence.

So, this EMC so that way I will have the combined energy function either the first one where it is just a linear combination of the two energy functions in evolutionary energy based and physics based and the second one, so, this mean and standard deviation of all the accepted state up to this is also being considered during computing the final energy function.

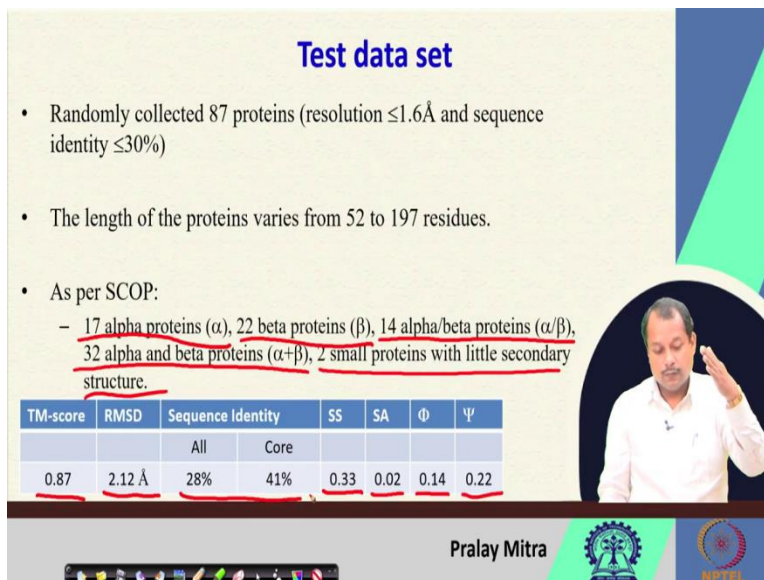
(Refer Slide Time: 25:55)

Test data set

- Randomly collected 87 proteins (resolution $\leq 1.6 \text{ \AA}$ and sequence identity $\leq 30\%$)
- The length of the proteins varies from 52 to 197 residues.
- As per SCOP:
 - 17 alpha proteins (α), 22 beta proteins (β), 14 alpha/beta proteins (α/β), 32 alpha and beta proteins ($\alpha+\beta$), 2 small proteins with little secondary structure.

TM-score	RMSD	Sequence Identity		SS	SA	Φ	Ψ
		All	Core				
0.87	2.12 \AA	28%	41%	0.33	0.02	0.14	0.22

Pralay Mitra



So, once we have one such protein design algorithm for which I also discussed the energy function then we have to test its design capability. So, for that we randomly collected or you

collect randomly 87 proteins the resolution is less than 1.6 angstrom and sequence identity less than 30 percent.

So, from here you can understand that sequence level they are not similar identity is less than 30 percent at resolution is also 1.6 less than 1 point angstrom which means they are really good structures. Now, apart from that one if you find that some structures are with the missing coordinate information then you have to rule them out because missing coordinate means that you have that sequence but do not have that structure which is not good for our protein design.

So, let us exclude them if you do not wish to exclude because in your say application mandatorily you want that particular structure then before giving that as an input to your protein design better to say refine it or model it or say repair it using some say modeling software something like protein folding software, et cetera.

Now the length of the proteins if varies from 52 to 197 although it has the capability to design that particular algorithm to design any protein sequence, but that there are some computational limitations that is why it escaped as 197 but you can extend it whereas, there should be a lower limit for our protein docking you remember 25 was the lower limit because we need at least one fold here in this case it is you can consider as one it is consider as just 50.

Now, as per the scope clash, there are 17 alpha proteins denoted as alpha 22 beta proteins, 14 alpha slash beta proteins and 32 alpha and beta proteins to small proteins with little secondary structure. So, you see, a variety of cases are there. So, it is going to be a real test for the protein design algorithm because different varieties or variations exist and I believe that now, it is clear to you that whether the, what is the scope clash and how to use that?

Now, if I look at the overall summary of the run. So, the TM score, so, when I say TM score, which means what so, you have the design sequences, so, on the 87 proteins, you design some sequences, you took the first design sequences or the sequence of the cluster whose cardinality is high then you use some protein folding software.

Because you got the sequence in order to get the structure because if you have only the design sequences and you give it to the biologists there is a possibility that it may not fold correctly. So,

you do some sort of bioinformatics analysis along with this one. So, for that, so, what I am suggesting that, with that sequence you go for protein folding you will get the structure.

Now, you compare this design's protein structure with the structure of your input protein that way the average you will get TM score is point 87 RMSD 2 point 12 angstrom. Then secondary structure point 33 percent is solvent accessibility phi point 02, point 14 and psi point 22 sequence identity interesting at the core.

The sequence identity is high 41 percent overall it is 28 percent. So, you see started for the random sequences it has the capability to recapitulate 28 percent identical sequences, of which 41 percent is at the core means it can preserve the core very much compared to the overall surface. So, with this let us stop here in this lecture, we will again continue this in the next lecture. Thank you very much