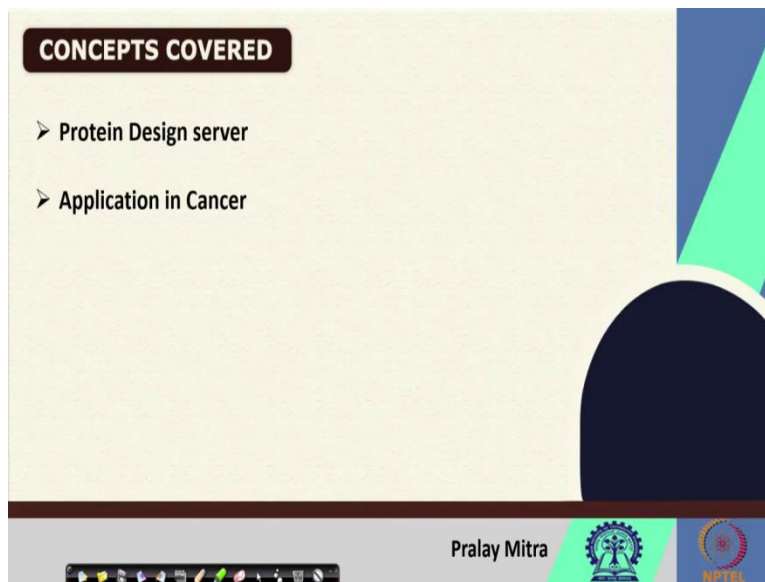


**Algorithms for Protein Modelling and Engineering**  
**Professor Pralay Mitra**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology, Kharagpur**  
**Lecture: 45**  
**Application of Protein Design on Drug Design**

Welcome back. So, in this lecture, we are planning to discuss some application of the protein design algorithm that we started three lectures back. So, we are planning to discuss on this computational protein designs and then we extend this for the application of cancer. So, there are several applications of protein design, but we will demonstrate few applications and then you can draw some analogy to your own problem and you will see that probably in your problem also you can apply this protein design technique.

(Refer Slide Time: 00:58)



The slide features a light beige background with a dark blue header bar containing the text "CONCEPTS COVERED" in white. Below the header, two bullet points are listed: "➤ Protein Design server" and "➤ Application in Cancer". The slide is decorated with a large dark blue semi-circle on the right side and a green and blue geometric shape in the top right corner. At the bottom, there is a dark blue footer bar with the name "Pralay Mitra" on the left, the IIT Kharagpur logo in the center, and the NPTEL logo on the right. A small toolbar with various icons is visible at the very bottom of the slide.

**KEYWORDS**

- Protein Design
- EvoDesign

Pralay Mitra

**Protein design in cancer research  
(Glioblastoma)**

- ✓ Root cause for glioblastoma high infiltrative potential is yet unknown and no clear way to prevent it.
- ✓ Interaction between Heat-shock protein (HSP) and Matrix Metalloproteinase (MMP) plays a crucial role in glioblastoma.

Pralay Mitra

So, the concept we are planning to cover protein design server what is there so, I will mention that one and application in cancer that I will mention. So, the keyword is protein design and the Evo design which is one protein design server. I will also mention that so, let us start with a protein design in cancer research. So, there are several applications.

So, people are working on XIP then BCL2 then apopto trick process in the cancer. But here we are starting with glioblastoma. So, the root cause for glioblastoma high infiltrative potential is yet unknown and no clear way to prevent it. So, interaction between Heat-shock protein or HSP and matrix metalloproteinase MMP plays a crucial role in glioblastoma. So, through some step by step way, we wish to explore that how we can go for designing a new drug for glioblastoma.

(Refer Slide Time: 02:06)

### glioma

Pralay Mitra

### Protein interaction

Complex	Interface Area (in Angstrom <sup>2</sup> )	NSc	NIP	Interaction energy (kcal/mol)
HSP27-MMP19	864	$3.1 \times 10^{-4}$	$2 \times 10^{-4}$	6.15
HSP27-MMP9	760	$3.4 \times 10^{-4}$	$2 \times 10^{-4}$	9.66
HSP27-MMP7	1085	$2.6 \times 10^{-4}$	$2 \times 10^{-4}$	15.73
HSP27-MMP2	1032	$1.5 \times 10^{-4}$	$2 \times 10^{-4}$	17.25
HSP70-MMP2	854	$2.1 \times 10^{-4}$	$3 \times 10^{-4}$	19.89

Pralay Mitra

*HSP27 - MMP*

*Interface Area*

*WACCESS*

*ASA (HSP 27) + ASA (MMP 19) - ASA (27-19)*

*2*

**Glioma**

Interaction site of HSP27

Pralay Mitra

So, first of all, in order to go for applying this technique that we have learned, so, for glioblastoma, you have to identify the proteins which are responsible for the initiation or progression of the glioblastoma. So, it is identified that hits a protein 27 and 70 and matrix metalloprotein 1, 2, 7, 9, 19.

So, among this there could be some interactions and because of that interactions the cancer may progress. So, to do that one, so, first thing you need to do that, whether this HSP and MMP interacts or not. So, you can see that there are two HSP proteins and there are five MMP proteins. So, in total 10 different interactions are possible.

So, who are they? So, which HSP 27 you can interact MMP-1, MMP-2, MMP-7, MMP9, MMP-19 and then with HSP 70 you have MMP-1, MMP-2, MMP-7, MMP9, MMP-19. So, 10 interactions out of these 10 interactions, how many are feasible. So, computationally we need to find that.

So, for that we discussed a number of docking techniques you pick any one of them and you check and after checking only one HSP 70 candidates are present in the complex form with MMP-2 rest are HSP 27 out of say five possibilities of HSP 27. So, one was rejected. So, 2 7 9 and 19. So, complex form, are there.

After doing this you have to go for some computational analysis to check whether all those five are equally probable or what is their probability. So, for that the computational analysis we have done very simple. So, all of them we have discussed so far. So, one thing is that interface area, you remember how to compute that one, if I pick any HSP 27 and MMP19.

So, you have to compute ASA using in NACCESS software ASA HSP 27 plus ASA MMP19 minus ASA complex of these 27 and 19 I am writing not I am not writing HSP and MMP explicitly here divided by 2 usually it is divided by 2 that interface area but, you understand that when I divide by 2 which means average but if I do not divide by 2 then I will get total but here I am using divided by 2 that is why I mentioned that it is divided by 2.

Now, this area is in again there is small typo here area interface area will be in square angstrom or angstrom square. So, better give angstrom square. So, we are having 864, 760, 1085, 1032 and then 854. Now, if you follow the basis of that PQS or PETA web services which was there. So, they relied mostly on the fact that if the interface increases that interaction site will increase and there is a possibility the interaction is going to be biologically valid.

So, it is good, but here if you see then except this guy all are mostly compatible specifically these two are compatible these two are compatible. So, from that point of view, it is not very easy to differentiate except I can say probably this is not going to be the correct interaction because this is 760 and also if you wish to pick the odd man then this is one another odd man. Where HSP 70 is there whereas 20 HSP 27 is interacting with four different MMPs but HSP 70 interacts only one MMP.

So, if it is true then it is very much selective in nature. If not then probably this is not correct. Next the NSc is computed. So, how to compute NSc that algorithm we discussed how to compute NIP normalized. This is normalized interface packing, this is normalized surface complementarity we discussed all of them.

So, normalization is done based by the interface area this interface area which is the average one. Now, these are the values if you look at the values then you will see that at the NIP level. So, all are same except this guy at the NSC level, the complementarity of these guy is very high. Now, you see that there is a DC some problem.

So, one is that 760 with low interface area, but with the high complementarity and as far the NIP also its value is comparable with the others interesting. Now, if you look at the interaction energy kcal per mole, then you will see a variety of the values. So, based upon that one, you may have some guests in some cases, but not always, but do not forget to do some bioinformatics analysis before going for experiment for the success.

Again here as I mentioned that HSP 70 was the odd man. So, let us leave that out. Consider all the HSP 27 interaction and identify the interface residues. So, here are, the interface residue that has been identified. So, the circles are the, this golden circles are the interface residues at the HSP 27 site.

Now, what we are going to do as of now, we are doing protein docking in order to identify the interaction site. Now, we identified the interaction site then what we are going to do that for HSP 27 we will try to design the HSP 27. So, that all the interaction sites will be allowed to muted. So, now I am designing the problem.

So, the problem initially was to understand that interaction of HSP and the MMP cause some cancer progression. In order to do that one using the docking I identified what are that interaction site if experimentally this is available very good you pick that one. Now, you have HSP 27 and what are their interfaces.

So, for all those interfaces, you allow the protein design. So, what I will do in my protein design algorithm that I have discussed. So, I will put restrictions that only the interface residues are allowed to mutate or change rest is not supposed to change because my interest is to mutate or

my interest is to say change the interface residue and then check before changing and after changing what is the overall effect on the interaction.

(Refer Slide Time: 10:48)


**Glioma**

*Molecular Dynamics (MD) Simulation*

Native HSP27	MTERRVPSLLRGPMDFFRDNYPHSRLFDQAFGLPRLPEENSQLGGSSMPGVVRLPFAAIESFAVAAPAYRALSRQLSSGVSEIRHTADRHRVSLVNH FADELTVTKDGVVEITGKHEERQDENGYISRCFTRKTYLPGVDPTQVSSSLSEGETLVEAMPKPLATQSNIEITPVTFESRAQLGGPEAAKSDETAAK
Mutant 1	<del>MTFFDLIKLLAHQW</del> SHDFFRDNYPHSRLFDQAFGLPRLPEENSQLGGSSMPGVVRLPFAAIESFAVAAPAYRALSRQLSSGVSEIRHTADRHRVSLVNH FADELTVTKDGVVEITGKHEERQDENGYISRCFTRKTYLPGVDPTQVSSSLSEGETLVEAMPKPLATQSNIEITPVTFESRAQLGGPEAAKSDETAAK
Mutant 2	MTERRVPSLLRGPMDFFRDNYPHSRLFDQAFGLPRLPEENSQLGGSSMPGVVRLPFAAIESFAVAAPAYRALSRQLSSGVSEIRHTADRHRVSLVNH FADELTVTKDGVVEITGKHEERQDENGYISRCFTRKTYLPGVDPTQVSSSLSEGETLVEAMPKPLATQSNIEITPVTFESRAQLGGPEAAKSDETAAK
Mutant 3	MTERRVPSLLRGPMDFFRDNYPHSRLFDQAFGLPRLPEENSQLGGSSMPGVVRLPFAAIESFAVAAPAYRALSRQLSSGVSEIRHTADRHRVSLVNH FADELTVTKDGVVEITGKHEERQDENGYISRCFTRKTYLPGVDPTQVSSSLSEGETLVEAMPKPLATQSNIEITPVTFESRAQLGGPEAAKSDETAAK


*Fold these sequences using same Protein Folding S/W*

*TM Score  
SS  
SA  
TA } NRE*



Rajesh et. al. (2019). BBA – General Subjects 1863(7):1196-1209


Pralay Mitra



**Glioma**


Native HSP27	MTERRVPSLLRGPMDFFRDNYPHSRLFDQAFGLPRLPEENSQLGGSSMPGVVRLPFAAIESFAVAAPAYRALSRQLSSGVSEIRHTADRHRVSLVNH FADELTVTKDGVVEITGKHEERQDENGYISRCFTRKTYLPGVDPTQVSSSLSEGETLVEAMPKPLATQSNIEITPVTFESRAQLGGPEAAKSDETAAK
Mutant 1	<del>MTFFDLIKLLAHQW</del> SHDFFRDNYPHSRLFDQAFGLPRLPEENSQLGGSSMPGVVRLPFAAIESFAVAAPAYRALSRQLSSGVSEIRHTADRHRVSLVNH FADELTVTKDGVVEITGKHEERQDENGYISRCFTRKTYLPGVDPTQVSSSLSEGETLVEAMPKPLATQSNIEITPVTFESRAQLGGPEAAKSDETAAK
Mutant 2	MTERRVPSLLRGPMDFFRDNYPHSRLFDQAFGLPRLPEENSQLGGSSMPGVVRLPFAAIESFAVAAPAYRALSRQLSSGVSEIRHTADRHRVSLVNH FADELTVTKDGVVEITGKHEERQDENGYISRCFTRKTYLPGVDPTQVSSSLSEGETLVEAMPKPLATQSNIEITPVTFESRAQLGGPEAAKSDETAAK
Mutant 3	MTERRVPSLLRGPMDFFRDNYPHSRLFDQAFGLPRLPEENSQLGGSSMPGVVRLPFAAIESFAVAAPAYRALSRQLSSGVSEIRHTADRHRVSLVNH FADELTVTKDGVVEITGKHEERQDENGYISRCFTRKTYLPGVDPTQVSSSLSEGETLVEAMPKPLATQSNIEITPVTFESRAQLGGPEAAKSDETAAK

Study presented HSP27 promoting EMT-like features in glioblastoma tumor cells through a direct interaction with MMP-2 and MMP-9 at an interface site AA29-40 of HSP27.



Rajesh et. al. (2019). BBA – General Subjects 1863(7):1196-1209

Pralay Mitra



**Glioma**

Precisely identify the target for drug development that inhibits GBM infiltration and migration.

Drug / inhibitor

Pralay Mitra

IIT Bombay NPTEL

So, for that I will identify the interface residues and for the glioma. So, I will do the protein design now, you see what I did so, here native sequence is given at the first so, the first one is the native sequence next three different mutant are given now, three different mutants they are colored red here, next here blue and pink.

So, these three regions are actually allowed to mutate based upon the information that which can be the interacting surface or interface. Now, what I will do that once these mutations are done again on these mutated sequences I will perform some biopolitics analysis. I am not showing you the details of that one but I believe that you can very much to that one what kind of analysis.

So, these are the sequences. So, first thing you will do that fold these sequences using some protein folding software then you compute what do you need to compute TM score, what else then secondary structure solvent accessibility torsional angle normalized relative error of all of this after that one some other analysis so that you can do.

So, another application I have that I will so that you can do molecular dynamics or MD simulation. MD simulation on what MD simulation on the folded structure folded means that the design sequence which was modeled using some protein folding software, so, that model structure you go for molecular dynamics or MD based simulation so after doing that one. So, several conclusions are there. So, that is not part of our course, but just for the, for your interest.



So, study presented HSP 27 promoting EMT-like features in glioblastoma tumor cells through a direct interaction with MMP-2, MMP-9 at an interface site. So, site is also mentioned. So, that is one application of the protein design software. And once you will identify that one that, this interaction is probably responsible for some action.

Now, if you wish to inhibit that action, then what do you need to do, you need to design a small molecule that will go and bind there so that they will not allow their interaction. So, what I am trying to say for example, so this is one protein and say this is another protein, now, you identified that this is my region using docking and then conservation of say, residue.

So, after the docking, so you got the list of interface residues, and then you mutate that interface residues using protein design and after the mutation, you say do the docking and compute the interaction energy computationally, and then you identify that there is a change in the interaction energy previously it was high.

Now, it has reduce which means because of this mutation at the interface, the interaction is losing. So, you can pinpoint it is not the complete interface. So, you see that went two protein protein interactions up there. So, what is the area of their interaction, so, minimum was around 760 and maximum was around 1050.

So, within this range that is a huge range. Now, you cannot design a small molecule which will cover everything. So, what you need to do you need to pinpoint out of that thousand say fifty interface area so, which residues are crucial for the interaction. So, if you cannot find one particular crucial but if you rank them and if you break their interaction by designing some small molecule then there is a probability that their interaction will be broken.

So, for example, if I consider the pink rectangle encloses that interaction area and also say this region in interaction area of red and blue protein then from here we can analyze that this region this pink circle at maybe the very much conserved region. So, what you can plan so, once using this protein docking and protein design you identify and pinpoint this region you design a say drug or inhibitor molecule, which will go and bind at that position now, if that will bind at this position, so, definitely the binding energy will reduce. Now, if the binding energy or the interaction energy reduces below a threshold then the interaction will not happen at all. So, that way you can kill some or you can inhibit some interaction.

(Refer Slide Time: 17:09)

**Protein design in Mycobacterium Tuberculosis (MTb)**

**Binding affinity of association of two molecules**

- *Mycobacterium Tuberculosis* (MTb) proteins are encoded by 4,062 genes.
- Out of 243 distinct proteins 151 (143) designed (target) proteins has confident binding partners of which 51.3% (50%) is enzyme.
- Designed sequence has on average more binding sites (6.2 per protein) than the target proteins (5.5 per protein).

Pralay Mitra

NPTEL

Now, if we apply the protein design technique for mycobacterium tuberculosis or MTb, so, binding affinity of association of two molecules that we are interested to look at. So, this MTb proteins are encoded by 4062 genes. Out of 243 distinct proteins 151 in total. So, 143 designed and 151 designed 143 target proteins as confident binding partners of which 51.3 percent is in enzyme.

Now, designed sequences after the analysis indicates has on average more binding sites then the target protein now up to this we can do the computational analysis and please note it down again that we are limited realistic states because we cannot always provide the final solution but definitely we can screen or we can prune down the solution space to some range within which the experimentalists or biologists or pharmacist can work and that way we can save huge amount of experimental resources which are required say we do not need we are suggesting that you need not have to go for several experiments.

But you go for say 20, 30 experiments you will get that correct it. But initially what you understand that given a protein of length  $n$  20 to the power  $n$  theoretical possibilities exist now, if you considered if you know that it is not the whole protein only a small stretch and that small stretch is also say is a 20. So, again 20 to the power 20 that stretch or that possibilities are also huge in nature.

So, this computational our protein design has a lot of application here in industry also it has lot of applications. So, in order to increase the yield of some product, so you can design your function of a protein so that the interaction increases and that way it increases the yields of product.

(Refer Slide Time: 19:37)

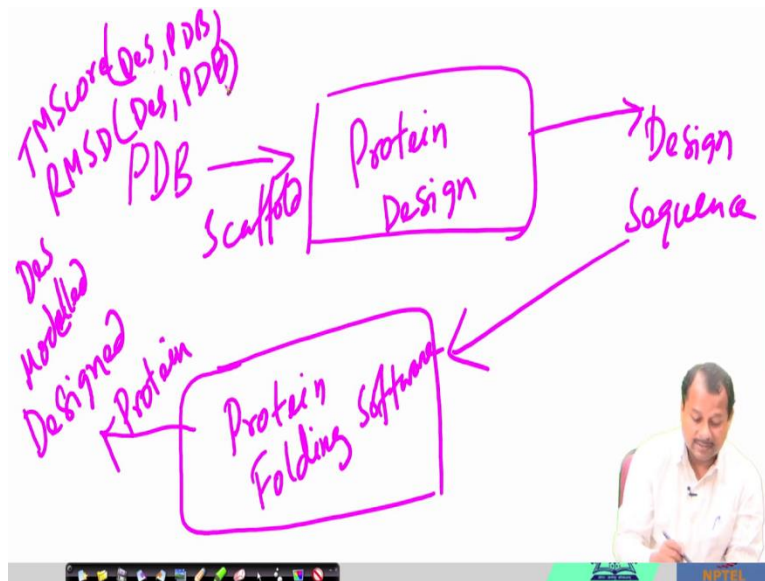
### More applications of protein design

Computational Analysis

Target	PDB ID	Length	SCOP class	RMSD (in Angstrom)
hnRNPk	1ZZK	80	$\alpha\beta$	2.99
thioredoxin	1R26	105	$\alpha\beta$	1.33
CISK-PX	1XTE	116	$\alpha\beta$	2.06
LOV2	2V0U	146	$\alpha\beta$	2.74
TIF1	3I4O	68	$\beta$	1.67

Mitra, P., Shultis, D., et al. (2013). *PLoS Computational Biology*

Pralay Mitra



So, more applications, our protein design. So, five separate proteins has been considered again for the further detailed analysis. So, one is the hnRnPK it is PDB ID is 1ZZK, thireodoxin it is PDB ID is 1R2K then CISK-PX it is PDB ID is 1XTE and they LOV2 it is PDB ID is 2V0U then TIF1 it is PDB ID is 3I4O.

So, if I look at the protein then its length varies from 80 to 146. So, and the scope class if I look then all so, except that TIF1 all are alpha beta, so, it (20:33) and beta both I mean helix and sheet both. Now, after performing the protein design, again in this case, we use the combined score function for the better accuracy and result then that design sequences has been modeled using some protein folding software and we got the structure then the design structure.

So, what is design structure? So, from input protein structure, which is taken mostly from the PDB because in this case exclusively it is from PDB because PDB ID I have mentioned. So, it has been taken from the protein databank then I took the sequence and the structure and scaffold I perform some protein design or say if I make it so PDB is giving me the scaffold.

Then I have protein design module or protein design box then I am getting design sequence then I am using protein folding software and I am getting designed protein this is modeled. So, this modeled designed protein I am getting starting from this PDB scaffold. Now, if I say this is my desk and this is say scaffold.

So, SC or say instead of SC I can use this PDB itself then if I compute that TM score between these Des and PDB if I compute RMSD between these Des and PDB then what is the value is being represented here. So, this is my RMSD in angstrom. So, it indicates that so, maximum is for hnRNPk that is 2 point 99 almost 3 and lowest is 1.33 for thioredoxin although the you note it down. So, the length of the hnRNPk is 80 only then also it is RMSD is high whereas these 1.33 I am getting from 105 length thioredoxin. So, this kind of analysis first we have done and since it is almost at the safe side. So, I am assuming probably I can go for further analysis.

(Refer Slide Time: 23:40)


### More applications of protein design

Experimental verification

Target	Expressed?	Soluble?	Secondary Structure?	3D structure?	$\alpha$ -helix%	Strand%
hnRNPK	Yes	Yes	Yes	No	32	16
thioredoxin	Yes	Yes	Yes	Yes	36	21
CISK-PX	Yes	Yes	Yes	Yes	27	28
LOV2	Yes	Yes	Yes	No	31	24
TIF1	Yes	Yes	Yes	Yes	9	37

Mitra, P., Shultis, D., et al. (2013). *PLoS Computational Biology*

Pralay Mitra




### More applications of protein design

Computational Analysis

Target	PDB ID	Length	SCOP class	RMSD (in Angstrom)
hnRNPK	1ZZK	80	$\alpha\beta$	2.99
thioredoxin	1R26	105	$\alpha\beta$	1.33
CISK-PX	1XTE	116	$\alpha\beta$	2.06
LOV2	2V0U	146	$\alpha\beta$	2.74
TIF1	3I4O	68	$\beta$	1.67

Mitra, P., Shultis, D., et al. (2013). *PLoS Computational Biology*

Pralay Mitra



So, what is the further analysis? So, further analysis is the experimental verification. So, for experimental verification what is performed. So, for that particular sequence. So, sequences are designed. So, sequence is known to you. So, you first check whether it will be expressed or not. So, what I get that all the design sequences are expressed Yes, Yes, Yes, Yes, Yes. Then when it is expressed.

So, I can extract that protein from that, say if I assume that in the equally I expressed that one that sequence then from the equally I can extract that sequences and after extract their sequences means I can extract that protein now, after extracting that protein then I check whether it is soluble or not because for further analysis if the protein is not soluble, then we cannot go for the further analysis. So, I have to check the solubility Yes, Yes, Yes, Yes, Yes So, all five are soluble also very good. Then will they demonstrate some secondary structure. So, there are several techniques like CD polarization, et cetera through which you can check that whether they are swing say helix sheet a secondary structure or not.

If yes then yes, it is also in secondary structures then we go for 3D structures. So, using say nuclear magnetic resonance or crystallography you can go for the study structure and when we go for this then so, for hnRNPK it is No for LOV2 it is NO but for thioredoxin, CISK-PX these are Yes and TIF1 is also Yes.

Now, if I go back to my computational prediction or computational analysis, then for this hnRNPK a LOV2 because I did not get 3D structure for them, then I will see that you see that for this and this LOV2, you see that RMSD is very high show they are easily indication during the bioinformatics analysis also that probably it is not going to be very good and also I mentioned that this 2.99 angstrom is for hnRNPK whose length is the smallest that is 80 only.

Whereas, for thioredoxin whose length is 105 it is 1 point 33. Now, you see that 116 it is 2 point 06 it is fine for LOV2. So, 2.74 is not good not bad, because it is length is very high 146. But, later when we discussed with the biologist, then we understand that this LOV2 has some specific nature and that is why it may not be very easy to design or maybe some special treatment or special care should be taken for this purpose.

So, this is the computational analysis and this is the experimental verification similar to this we can apply this protein design technique for several other diseases or for several other applications. So, to summarize, what we discussed in this week started from the new topic on protein design, we felt the need of some computational framework for protein design. So, protein design says given one protein structure or scaffold as an input.

So, the sequence is also known to me, but I have to come up with some alternative sequences which perhaps is not present in nature, but that sequence will fold to this particular structure,

particular structure means to this input structure. So, that way you are getting in new sequence now, we have the sequence after we got the sequence before going for experimental verification.

So, what we are doing that we go a series of analysis on the sequences first we take the sequence and using some protein folding software we folded that sequence and or we modeled that sequence then we compare the TM score and RMSD with the modeled structure modeled design protein and the input structure then we computed at the secondary structure level solvent accessibility level torsional level.

So, what are the variations and based upon that one we are trying to build our confidence that whether we should go for real experimental verification or not. So, that way we see that performances are good specifically when we use the evolutionary energy function as well as the Physics-based energy function and at the homologous structure level we consider the structures which are more than point 7 TM score. So, I mean that we should consider if I consider the structures or if I declared the structures homologous to the input structure whose TM score is more than 0.7 with the input structure. If I consider that one then perhaps my accuracy will increase.

(Refer Slide Time: 29:29)



So, next, this total Evo design housed at University of Michigan at the Professor Yang zhang lab. So, it provides this protein design the algorithm that we discussed as a web service. But modifications are going on and also I discuss some of the modifications which you can do and

based upon that one and based upon your application, you can customize the software or you can develop similar to this some other protein design technique or algorithm and apply in your method. So, thank you very much for your attention.