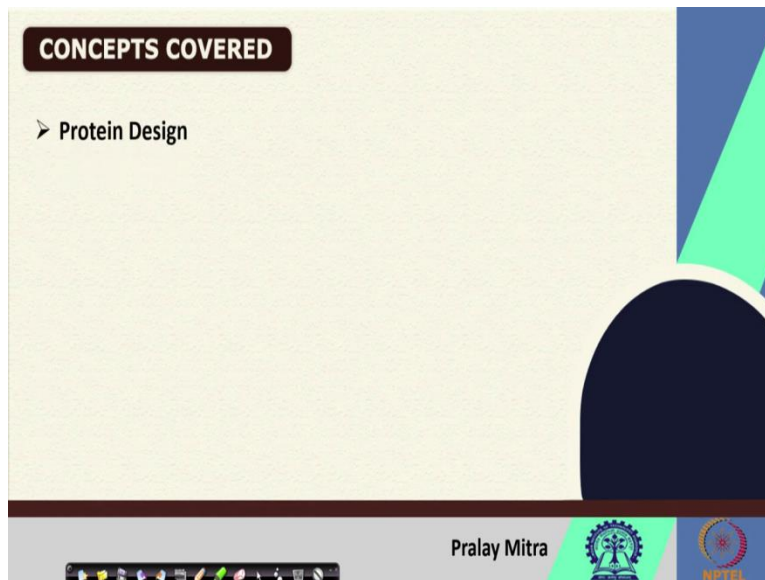


Algorithms for Protein Modelling and Engineering
Professor Pralay Mitra
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur
Lecture: 44
Protein Design Analysis

Welcome back. So, we are continuing with the protein design algorithm. So, the detail algorithm we have discussed part of that the energy or the score function we also discussed and the data set that we used and what is the on an average the result of the data set that we discussed on the last slide of the last lecture.

Now, in this lecture I wish to do a little more analysis and definitely in the next other lectures also we will see that when we are going to design some new sequences, some sort of analysis are required that analysis I am talking is bioinformatics analysis, because, again we are doing computational modeling then in order to gain the confidence of the biologists or in order to give the right product as an output of our protein design algorithm, so, we have to be careful and we have to give the right solution so, that the reproducibility in the experimental laboratory will very high.

(Refer Slide Time: 01:21)



KEYWORDS

> Protein Design

Pralay Mitra

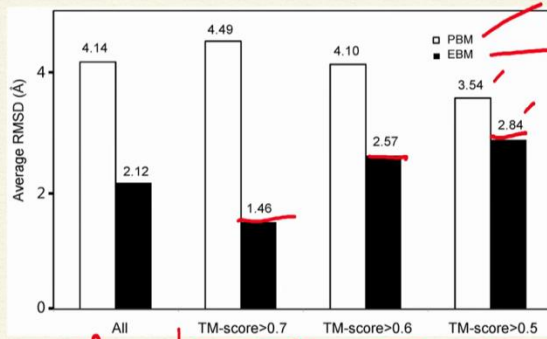


As the TM score increases (above 0.5) seq structural similarity increases and sequence level diversity decrease

lower limit on TM score 0.5



Role of homologous structures in protein design



Physics based / Energy based function

Evolution based function

NX20

11

18

25

Pralay Mitra



So, the concept we will be covering protein design and then different methods I will be talking. So, first of all, let us discuss another interesting one. What is the role of homologous structures in protein design? So, you remember that the first step of our algorithm was that, after getting the scaffold or the input protein structure, first job was to identify the homologous structures from the protein databank.

So, by homologous structure, we wanted to say that, for the inputs with the input structure, you do the structural alignment and after that alignment, if the RMSD or the TM score the measure for checking the goodness of the structural level match is beyond some threshold then you consider I also mentioned that, in order to reduce the bias in this process, you should exclude those structures which are identical.

Because, you should remember one thing that input structure that user is providing to you mostly may come from the protein databank itself. Now, you are again comparing the input structure with the protein databank in order to search for homologous structures. So, there will be a chance that you will get the same structure which is supplied as the input so, you have to definitely exclude that structure from your homologous structures list apart from that one.

If there are multiple copies of that particular sequence of structure in the database, we try to avoid that. But I did not mention anything regarding the other bound. So, that is regarding excluding the identical structures in order to make it without bias but what will be my threshold for RMSD or TM score in this case, I will prefer to use TM score because as I mentioned during the discussion of the TM score.

It is a normalized score it varies from 0 to 1 with 0.5 indicate that is the threshold so greater than 0.5 indicates that at the fold level the structures are same and something less than a 0.3 indicates that is random structure, et cetera. So, while I am looking for homologous structure, first thing I am considering TM score because it is a normalized score, so this kind of say 0.3 threshold level which indicates that below 0.3 it is random structure above 0.5 at the fold level they are similar, this kind of conclusion is not easy to infer for RMSD based calculations.

Now if I move on to TM score the next thing is that as per the published work by the authors that greater than 0.5 indicates that fold level accuracy present so our lower limit on the TM score should not be less than 0.5. So, that way, I am getting one lower limit of TM score 0.5. Now, if I

go above 0.5 then you see that homologous structures are there and since the structures are very much similar with each other, then at the sequence level the diversity will keep on decreasing. So, you can consider that as the TM score increases. Above 0.5 of course structural similarity increases and sequence level diversity decreases.

So, this is a very good observation and based upon that one and based upon your application what you wish to achieve with a diverse sequence is your interest during protein design or you wish to retain most of the sequences and you need the design sequences should be very much or as much identical as possible with the original sequence but with some variation should not be it is 100 percent.

But, it will be something like that if it is the case then accordingly you can tune or you can basically fix your threshold value that is what is the proposal as for the algorithm So, you see that when say TM score is greater than 0.5 then the average RMSD in this plot along the x-axis it is the different threshold value for TM score which is 0.5, 0.6, 0.7 and along the y-axis it is the average RMSD of the predicted or model design protein.

So, you remember on the last lecture of the last slide I mentioned 87 test proteins are there with the 87 test protein we perform the protein design with different TM score level and when we got the design sequences the first rank design sequence has been picked and that particular sequence using some protein folding software I got the structure of that once I will get the structure of that design protein then I align with the input structure and compute that TM score and the RMSD.

So, here that RMSD I am talking about. So, after computing the RMSD of the design protein structure and the input protein structure how did I get the design protein structure protein design algorithm will give you design sequence then from that sequence using some protein folding software then you can predict some protein structure with that structure and the input structure you are computing what is the percentage of alignment definitely it will be 100 percent because this is same then you compute that what is the RMSD and that RMSD over the 87 proteins is being considered here.

Now, that 87 proteins when I average over then you see that what is the average RMSD. So, the average RMSD here is 3.54 and 2.84. Again I am dividing the result into two parts. So, the PBM indicates physics based and EBM indicates evolution based. So, this is regarding the energy

function. So, one is physics based another is evolution based. So, among these you can see that the physics best is always performing little poorer compared to the evolution best. So, that is why we are just focusing only the evolution best and then for TM score greater than 0.5 where the diversity in the structure as well as the sequence level is high is giving you average RMSD 2 0.84 angstrom which will be reduced if I consider TM scores 0.6 then it will be 2.54 if I consider TM score 0.7 then it will be 1.46 you see what is the change?

So, from 1.84 to 1.46 just half. So, this indicates that as the TM score will increase which means you are going for more and more homologous structures then your prediction is going to give you more close to the original structure. But, it is also true that or the question I may ask you that this way if I going to change then definitely the number of protein structure that I will get will also change here of course.

So, if the TM score is 0.5 I am looking for a variety if it is 0.6 then it will I will get a subset 0 0.7 I will get less number of structures. So, from here I believe it is trivial and clear to you that if I put a threshold of 0.7 then that is going to be a subset of the structure that I will get using TM scores 0.6 and that is going to be a subset of the structure for which TM score is greater than 0.5.

If I am getting that, then safely I can say assume one scenario where say for TM score 0.5 I am getting say 25 structures in this case I am getting say 18 structure in this case I am getting say 11 structures. Now, you think about the computing the PSSM. So, for 0.7 I am computing using 11 sequences 18 sequences 25 sequences, not only the number but I am also computing the PSSM based upon the less number of sequences and less diversity sequences that way you may not expect variety of sequences or the low sequence identity during the design.

So, you have to fix it the threshold on the TM score as per your need. One question at this position may be asked that is there any limit on the number of structures I need for calculating the PSSM because very much it may possible that for one input protein structure which is also kind of a novel there is no homologous structures what will happen?

So, if no homologous structures are there, say then you can assume say for 0.5 greater than say 1 or 2 I will get or I will do not get if I do not get then you see that the basic protein design algorithm will boil down to a situation where that N cross 20 PSSA matrix will have the equal

probability in all places. So, algorithm will run but the accuracy will vary and the variation is based upon the availability of the structural information.

So, in that way also if we do not have a sufficient number of homologous structures in order to compute the PSSM then it may bias say for example, if I got two sequences and the sequences are very much similar to each other then hardly in position I will get some mutation otherwise the same mutation I mean the vallin will be replaced by vallin alanine will be replaced by alanine that kind of mutation will keep on happening. So, which is not also expected so, that thing also you have to keep in your mind you cannot run it blank.


(Refer Slide Time: 14:16)

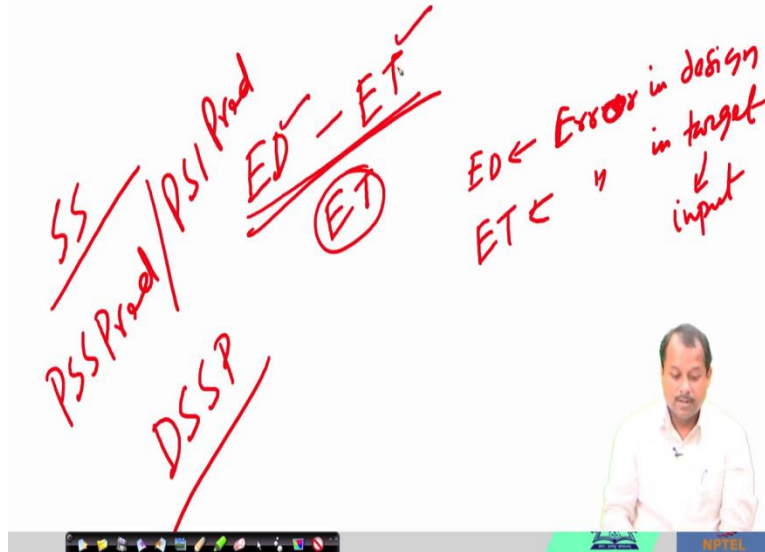
Role of energy functions in protein design

Energy function	TM-score	RMSD (in Angstrom)	Sequence Identity	
			All	Core
Physics-based	0.74	4.14	21%	35%
Evolution-based	0.82	2.82	27%	35%
Combined	0.87	2.12	28%	41%

Energy function	Normalized Relative Error (NRE)			
	SS	SA	Φ	Ψ
Physics-based	2.40	0.43	1.01	0.41
Evolution-based	0.48	0.26	0.30	0.04
Combined	0.33	0.14	0.22	0.02

Pralay Mitra





Next is the role of energy function in protein design. So, on the last slide also we noted that if we use just for physics based force field then at the RMSD level the RMSD value is higher compared to if we use evolution best scoring function. And also last slide we discussed based upon the evolution best only.

Now, it is the time to demonstrate since we have three different variations of energy functions, then which one to use physics based evolution based or the combined and similar to that at the same TM score level or at the structure homology level the way we fixed that how many we should consider in order to generate our PSSM in this case, what is the effect of the energy function in protein design that we are going to analyze.

So, we divided that result based upon this or you can consider on the same data the algorithm is drawn three times one for physics based energy function and another for evolution based energy function another for the combined one. Now, you see the TM score if it is combined then it is a high 0.87 TM score very high RMSD 2.12 sequence identity 28 percent identity core 41 percent and in all aspects physics based force field is failing even with respect to the evolution based.

So, you need to consider some sort of evolutionary based information along with the physics based. If I look at the sequence identity and if your intention is to come up with some sequence which is as lower sequence identity compared to the input protein sequence then also at the core level or at the core you have to keep your sequence identity is very high.

Otherwise, there is a chance that at the protein folding process or experimentally when you are trying to express that protein or say design that you need to reproduce that protein then it may not fold because core is mostly hydrophobic and it is trying to retain its nature of hydrophobicity even after its design please note it down we are considering sequence identity which means that during the sequence alignment here the alignment is the identical because the same structure same sequence is being designed.

So, the input sequence and then design sequence one to one correspondence is also there no gap nothing and we will count that how many cases they are matching how many cases they are not matching if they are matching plus 1 if they are not matching then 0 that way you are computing what is the identity normalized by the length of the sequence of course. Now, next other features normalized relative error we computed on secondary structure SS solvent accessibility SA and phi and psi.

So, what we did for these features that we computed normalized relative error. So, what is that normalized relative error? So, for this we are considering that we are having only the sequence if final design sequence if we are having the final design sequences then so, one is the design another is the predicted or say this may be confusing. ED indicates error in design ET indicates error in target this target is nothing but the input or input.

Now, you compute the error on the design and compute the error on the target and then normalize and based upon that one you will get the normalized relative error. So, this is the relative part and this is the normalized part. Now, when you are computing this then one thing you should keep in your mind that whether you will be computing these using the structure or using the sequence it will be uniform for target as well as that design.

So, target means the input or the scaffold say either you can say predict the secondary structure solvent accessibility torsional angle based upon the design sequence and based upon the sequence of the input structure, then you compute using this equation or you model the design protein from the design sequence using some protein folding software then on that model structure you compute the or you predict the secondary structure which are structure based software using that one you compute that one and you take this difference and you have this computation.

So, if say for example for secondary structure. So, if at the sequence level you are computing you can use say PSS pred or say PSI pred but if you are using at the structural level you can very much use DSSP software but whatever you will use you use it both for both target as well fossa design.

(Refer Slide Time: 20:51)

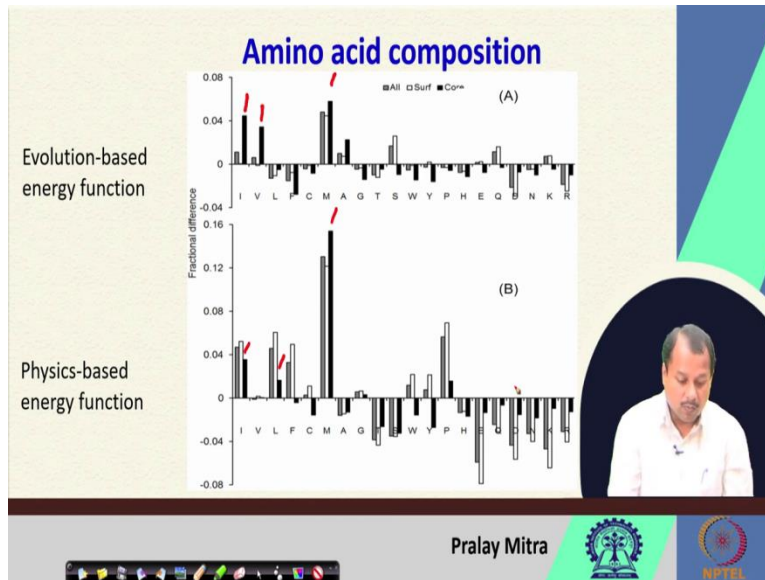
Role of energy functions in protein design

Energy function	TM-score	RMSD (in Angstrom)	Sequence Identity	
			All	Core
Physics-based	0.74	4.14	21%	35%
Evolution-based	0.82	2.82	27%	35%
Combined	0.87	2.12	28%	41%
Energy function	Normalized Relative Error (NRE)			
	SS	SA	Φ	Ψ
Physics-based	2.40	0.43	1.01	0.41
Evolution-based	0.48	0.26	0.30	0.04
Combined	0.33	0.14	0.22	0.02

Pralay Mitra

And if I look at the data then you will see that for the physics based you see that 2.40 is the secondary structure. So, 0.43 is the solvent accessibility phi and psi angles are also very high it is swing very high relative error for combined like the previous one here and here also you see that this is the lowest among all. So, in both cases it justifies that in the energy function, you should combine the evolution based and physics based and if you use that combined then you are going to get better design sequence compared to others.

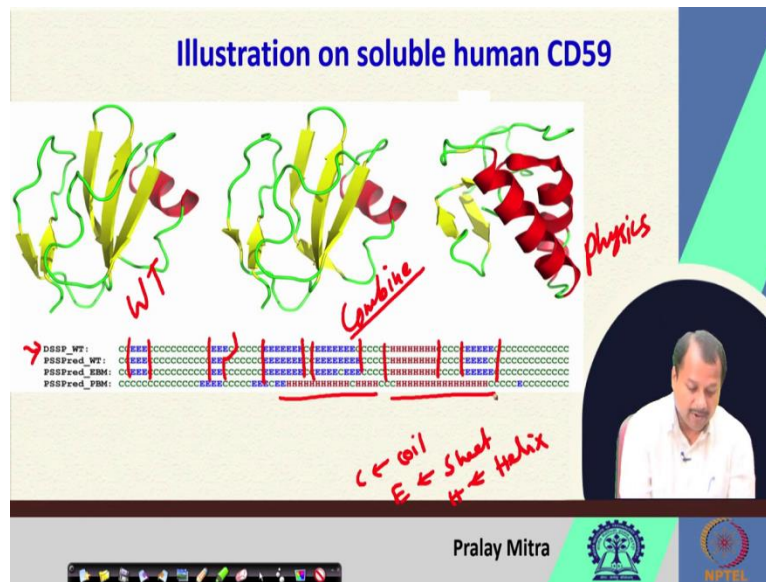
(Refer Slide Time: 21:42)



Next, if we analyze the design sequences at the amino acid composition level then we divide this into two parts. So, one is that all another core and surface as that the core of a protein is going to be more hydrophobic compared to the surface. So, hydrophobic residues which are they are going to be more on the at the core. So, if I use evolution based energy function and physics based then also you see that these isoleucine at the core, core is this black.

So, at the core isoleucine valine that increases whereas for the physics based they are not much methynin it is high for both cases the reasons is that methynin occurs very less in the nature, but when you are going for the design since it is getting more weightage compared to the its actual existence in the nature. So, if it occurs little more than also it magnifies and then it shows that it is occurring heavily. So, this is the analysis which indicates that at the amino acid composition level also what is the variation.

(Refer Slide Time: 23:11)



Now if we do more analysis by looking at the structure then the visualization effect will tell you that why I am vouching for this combined energy function physics-based energy and evolutionary based. This is an example of soluble human CD59 protein in this case, so, I am predicting using DSSP software on the wild type.

So, wild type means the native structure I am predicting on using DSSP the first line so E indicates it sheet C indicates coil and H is helix that is our same nomenclature we are using. Now, you need not have to look at this sequence alignment if you look at the structure then also you can see what is the variation. So, the first one is the wild type.

Next one is the output of the combined energy function and this one is the output of the physics based and here, so this combined this one, if you now look at the sequence alignment, then you will see that this stretch this part, this part, this part, this helix region, this sheet is almost same for say DSSP at the structural level when I computing PSS pred PSS pred is very good or on the wild type sequence or on the evolutionary based, or the combined prediction.

But when I am going for the physics based one, then I see that helix prediction is not correct. But that helix has a very good pattern at the hydrogen bonding level. So, predicting the helix is easy and that is why you will see most of the secondary structure algorithm prediction algorithm starting from this (25:27) man then you name any algorithm then you will see the accuracy on the helix prediction is very high.

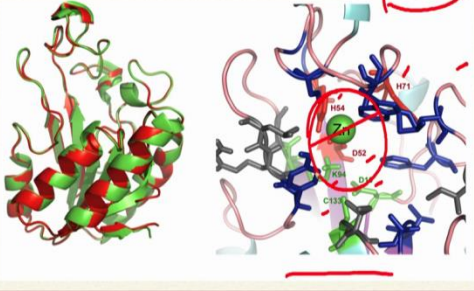
So, that way you should not have any doubt that because of the prediction system of the PSS pred PBM pills It is not like that it is because of the fact that PBM a physics based energy only is not good enough for the modeling. So, we have to have the combined or the energy based one.

(Refer Slide Time: 26:00)

Conservation of Functional Site

Crystal Structure of Pyrazinamidase of *Pyrococcus horikoshii* in Complex with Zinc (PDB ID: 1IM5)

RMSD between crystal structure and design structure is 0.28 Å

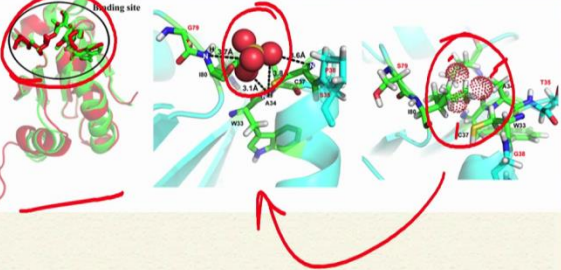


Pralay Mitra

Designing new binding site

Crystal structure from the thioredoxin C (PDB ID: 2H1U)

RMSD between crystal structure and design structure is 2.52 Å



Pralay Mitra

Next application level we will demonstrate the conservation of functional site. So, you look at the structure it is the crystal structure of protein and which are in complex with the zinc PDB ID is 1IM5 what is done, so, for this particular protein which is shown on the right hand side here for this particular protein. So, the zinc is a conserved one and it has one binding site, what are the residues at the environment of the zinc is also marked here.

So, here histidine I can see where here and here then aspartic acid I can see here and here the lysine and one cysteine is also present here. Now, during the design process, if you take out this zinc if you do not allow the zinc to be there, but its environment will be there then after the design process you will see that you see that the crystal structure of this one and the design sequence after you predict using some structure modeling software.

I mean protein folding software then you see that at the RMSD level almost no difference. So, I believe that if you compute the TM score you can get say about one if not 0.99. So, this also indicates that it will preserve the conserved residues if you do not allow them to alter. So, that is one application. So, where you wish to have a new sequence by preserving the functional site, then you can apply this one another application is designing new binding site.

So, for this crystal structure of thioredoxin C PDB ID is 2I1U then for this particular structure, so binding site is shown here on the left hand side here, this is my binding site. Now, if this is my binding site, then these dots small dots indicates that I cannot accommodate any small molecule at that particular position. But if during the design, I allow this region to mutate and allow them to give some flexibility there.

Then it is possible that this will give me a design sequence something like this, where after the mutation, I can very much accommodate one small molecule which can interact and have some function. So, for designing the new binding site or conserving the existing binding site, just by mentioning that one during your protein design process, you can do that one. So, these are some of the analysis that we have done. So in the next lecture, we will see a few more applications of this protein design algorithm. Thank you very much.