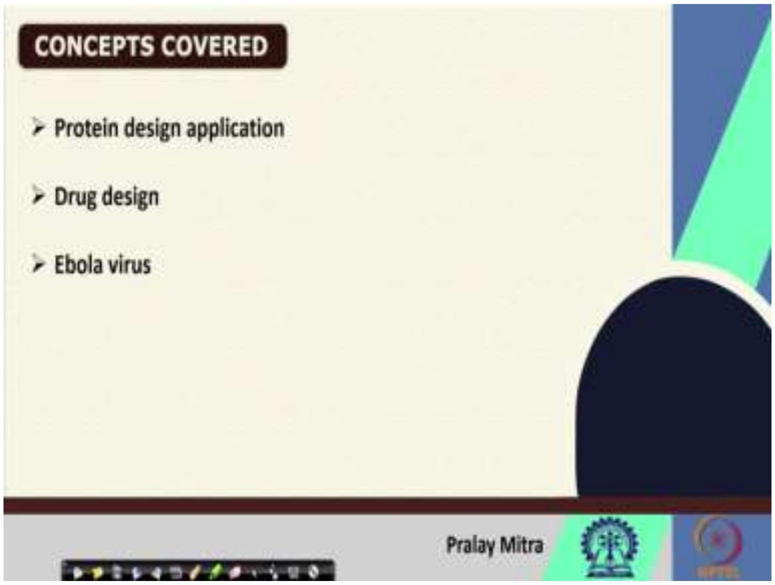
Algorithms for Protein Modelling and Engineering
Professor Pralay Mitra
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur
Lecture 47
Application of Protein Design on Drug Design

Welcome back. So, we are continuing with the application of replica exchange Monte Carlo in protein design. Now, in protein design we have noted that if we go for a replica exchange Monte Carlo then the chances are high that we will have a better optimized sequences and also we noted on the last lecture that there is a possibility to speed up the computation by implementing it in some parallel environment like shared memory programming for example, open MP.

So, in detail we did not discuss but we just give you a an idea that probably it can be implemented because there is as such no data dependency where local steps are running only dependency is that when during the simulation stage, there is a need of data exchange and that is after the execution of the local steps are over. So, we discussed that one now, in this slide we will see more applications of protein design specifically in drug design.

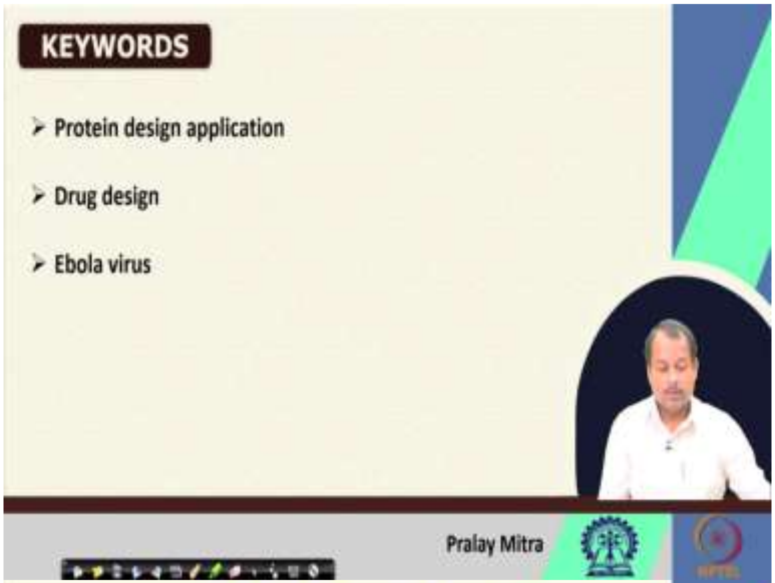
(Refer Slide Time: 01:20)



CONCEPTS COVERED

- Protein design application
- Drug design
- Ebola virus

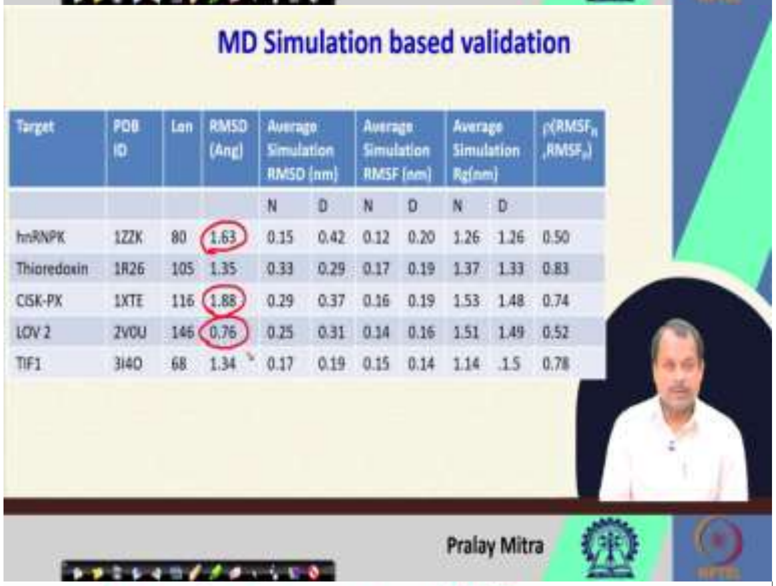
Pralay Mitra



KEYWORDS

- Protein design application
- Drug design
- Ebola virus

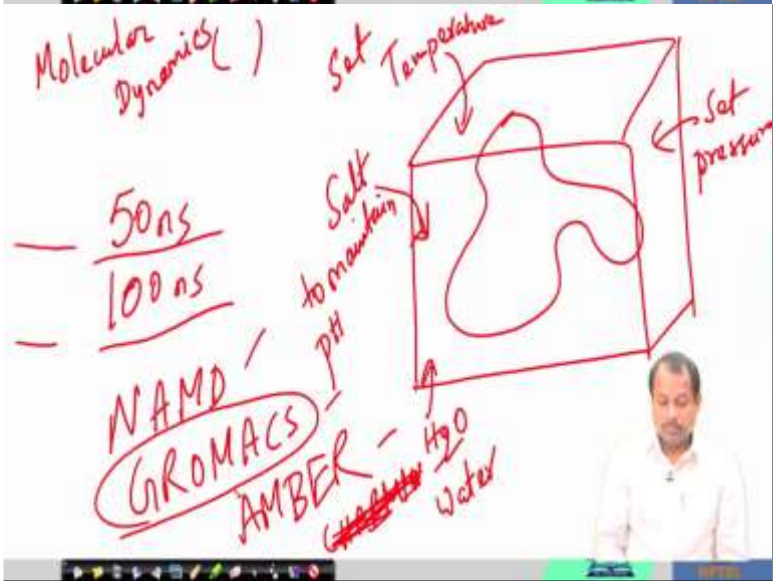
Pralay Mitra



MD Simulation based validation

Target	PDB ID	Len	RMSD (Ang)	Average Simulation RMSD (nm)		Average Simulation RMSF (nm)		Average Simulation Rg(nm)		$\rho(RMSF_N, RMSF_D)$
				N	D	N	D	N	D	
hnRNPK	1ZZK	80	1.63	0.15	0.42	0.12	0.20	1.26	1.26	0.50
Thioredoxin	1R26	105	1.35	0.33	0.29	0.17	0.19	1.37	1.33	0.83
CISK-PX	1XTE	116	1.88	0.29	0.37	0.16	0.19	1.53	1.48	0.74
LOV 2	2VOU	146	0.76	0.25	0.31	0.14	0.16	1.51	1.49	0.52
TIF1	3I4O	68	1.34	0.17	0.19	0.15	0.14	1.14	.1.5	0.78

Pralay Mitra



So, accordingly we decided to cover protein design application drug design and some application in Ebola virus proteins. Accordingly, the keywords are also same. Now, let us start with same seek five proteins that has been benchmarked on the last week using the Monte Carlo simulation of protein design, which is also there in Evo design web server.

Now, when we say extended that algorithm for replica exchange Monte Carlo technique, then it is expected that we need to benchmark the replica exchange Monte Carlo also for our same data set, the same 87 data set was benchmarked that result I am not showing you, but for this five case studies.

So, for which we did the analysis and experimental validation using the EVO design is taken for validation. Now, hnRNPK, Thioredoxin, CISK-PX, LOV 2, TIF1 those five proteins with the same PDB IDs and since the same protein so the scope classes will also be same that we consider, length is same.

Now, you see that if you remember correctly that in last class, so, our RMSD for these hnRNPK was very high it was 2.99. Now, you see it is 1.63 that is a tremendous reduction. So, 1.88 for this CISK is also reduction, fantastic reduction is for LOV 2 which is 0.76 only.

Now, from that point of view, we can say that probably replica exchange Monte Carlo is able to find out the better sequences with lower energy compared to the previous one. So, that is really a good news for us. And also, it is interesting to remember and to see that what will be the effect for this LOV 2 an hnRNPK.

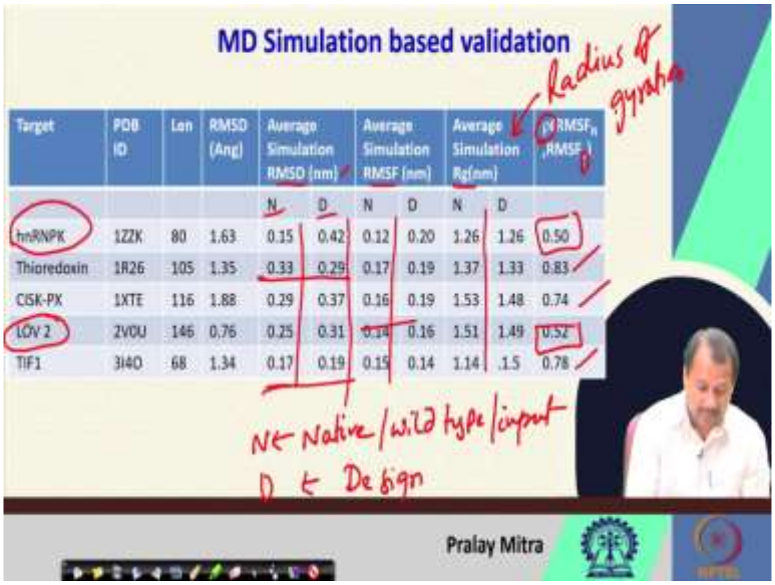
Because for these two particular proteins, we noted that it was expressed in the E.coli experimental relation I am talking about, it was expressed in E.coli, it was soluble, it show secondary structure, but it did not able to go for 3D complex structure. So, we did not able to get that 3D structure using NMR or X Ray crystallography.

So, that is why let us see. So, what we are now doing that instead of going for experimental technique, which is time consuming, so, let us do an another alternative computational way which is called as the MD or the molecular dynamics based simulation. Now, molecular dynamics based simulation, although is out of scope of this class.

But in short what I can tell you is that, so, for that there is one box you have to decide or select. So, this is a computational hypothetical box inside that one you put your protein structure and when you will put make sure that on all direction it is not penetrating the wall of this box or there is some gap between the wall of the box and the protein then you feel this with water molecule, you may optionally add some salt to maintain pH. Next you change temperature or you set better, set pressure and assume that it will keep on simulating, usually the simulation is done for nanoseconds.

So, you can go for say 50 nanoseconds or 100 nanoseconds simulations. So, there are molecular dynamics softwares like NAMD, GROMACS, AMBER, CHARM, et cetera. So, anyway so this NAMD, GROMACS, AMBER, CHARM. So, this among this these GROMACS is free and it is developed in C. So, it is most widely used specifically for protein structures. So, there is a software using that one you can do this one.

(Refer Slide Time: 07:09)



MD Simulation based validation

Target	PDB ID	Len	RMSD (Ang)	Average Simulation RMSD (nm)		Average Simulation RMSF (nm)		Average Simulation Rg(nm)		$\rho(RMSF_N ,RMSF_D)$
				N	D	N	D	N	D	
hnRNPK	1ZZK	80	1.63	0.15	0.42	0.12	0.20	1.26	1.26	0.50
Thioredoxin	1R26	105	1.35	0.33	0.29	0.17	0.19	1.37	1.33	0.83
CISK-PX	1XTE	116	1.88	0.29	0.37	0.16	0.19	1.53	1.48	0.74
LOV 2	2VOU	146	0.76	0.25	0.31	0.14	0.16	1.51	1.49	0.52
TIF1	3I4O	68	1.34	0.17	0.19	0.15	0.14	1.14	.1.5	0.78

Radius of gyration

N ← Native / wild type / input

D ← Design

Pralay Mitra

Now, if you do then the simulations are going on when the simulations are going on, then what you can do that you can take the trajectory of the simulation. So, let us assume if this is a protein then you can very much expect that during the simulation it will keep on say fluctuating or it can keep on changing a structure.

So, we will record that information in say one nanosecond interval nano means 10 to the power minus 9 second interval and say if it is running for 50 seconds, then up to 50 seconds if it is running for 50 nanoseconds, then up to 50 nanoseconds. So, starting from the initial one, so, 51 frame at each nanosecond one frame that way 51 frame I will capture then what you will do that you will compute the RMSD of the structures. So, the original structure is known to you and during the simulation at each nanosecond assuming 50 nanosecond simulations.

So, in total 51 structures or if you consider the initial one the original one so you can leave that one. So, 50 structures you have. So, with each structure and the original structure you compute the RMSD then you take the average of all the RMSDs, if you plot then you will get a plot like this where this indicates when you are simulating the native structure, then what is the root mean square deviation and if you are simulating the design, D stands for design, N stands for native.

So, N stands for native or wild type or input and D stands for design. So, what is the structure? Now, you see at the nanometer level I am talking about and there is not much change. So, the change although in the design sequence I can see except for the Thiredoxin so all are high but that high is also marginal in nature. Now, average simulation when we are calculating the root mean square fluctuation, fluctuation means the fluctuation of the atoms.

So, this root mean square deviation, root mean square fluctuation and this Rg is the radius of gyration. It is easy to compute this RMSD, RMSF and Rg if you are using say GROMACS software. So, there is inbuilt software through which you can analyze and it will give you not only the final value, but it will give you XVG file that you can use for plotting and you can see that what is the variation how the RMSD is changing.

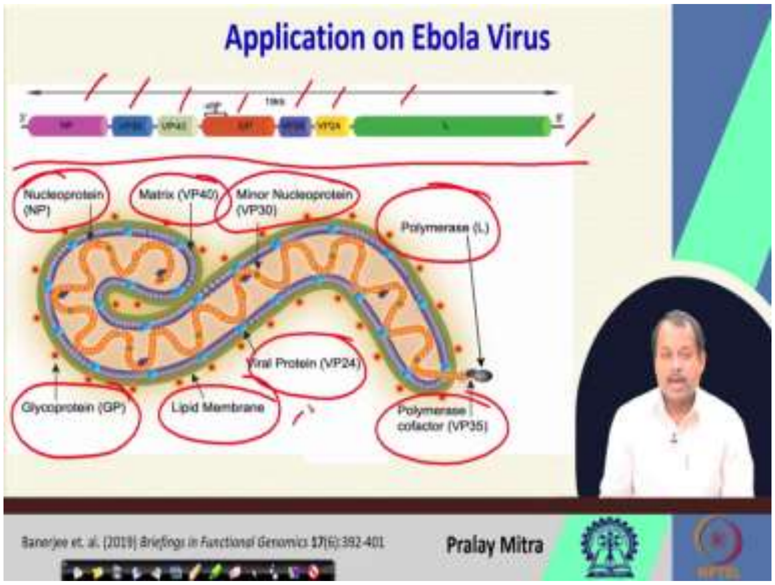
So, whether it is going to be saturated or it is going to keep on increasing. So, that trend also you can note from there and if you compare the RMSF and say RG for native and design then you will also see that not much variation and if you go for the row I mean the correlation between the RMSF of the native and this is D that is designed then we will see that very high correlation is observed for Thiredoxin for TIF1 for CISK-PX and these two guys are again the culprit.

So, you remember that for these two guys, last time we failed to get the 3D structure. Now, here also it demonstrates that the row value specifically indicates that row means correlation between this that probably this time also we are going to fail for this two structures. Now, you see if I go for computational design of these protein sequences on an average it will take few hours if I go for say molecular dynamic simulation for the validation.

Then in the latest GPU server it will take again few hours. So, combining that in say within a day corresponding to one protein structure you can conclude that what is the design sequences whether that design sequences is going to be stable experimentally or not. Well, I missed one thing after the design sequences you have to model the structure using some protein folding software because MD will take or molecular dynamics simulation will take one protein structure as an input.

So, you have to model that structure. But if I consider that protein structure modeling will take one more day. So, within two days, you can start with one protein structure and computational predict what are the protein sequences and whether these protein sequences are going to be stable experimentally or not. Then based upon that observation, you can conclude and provably you can infer that whether I will go for experimental validation or not. So, is not it a very, very effective application, you can think.

(Refer Slide Time: 12:59)



Application on Ebola Virus

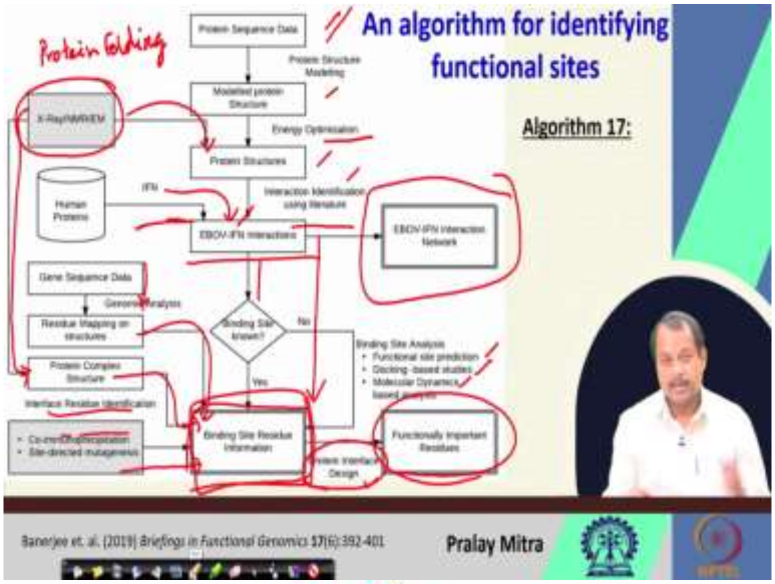
Banerjee et. al. (2019) *Briefings in Functional Genomics* 17(6):392-401

Pralay Mitra

An algorithm for identifying functional sites

Algorithm 17:

Protein Folding



Protein Sequence Data

Protein Structure Modelling

Modelled protein Structure

X-Ray/NMR/EM

Energy Optimisation

Protein Structures

IFN

Human Proteins

Interaction Identification using literature

EBOV-IFN Interactions

EBOV-IFN Interaction Network

Gene Sequence Data

Genomic Analysis

Residue Mapping on structures

Protein Complex Structure

Interface Residue Identification

Binding Site known?

No

Yes

Binding Site Analysis
- Functional site prediction
- Docking-based studies
- Molecular Dynamics based analysis

Binding Site Residue Information

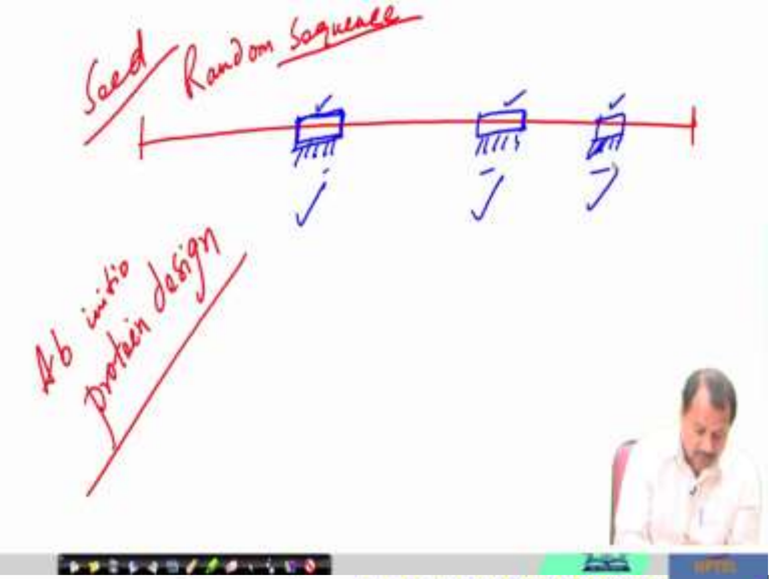
Functionally important Residues

Protein Interface Design

- Co-immunoprecipitation
- Site-directed mutagenesis

Banerjee et. al. (2019) *Briefings in Functional Genomics* 17(6):392-401

Pralay Mitra



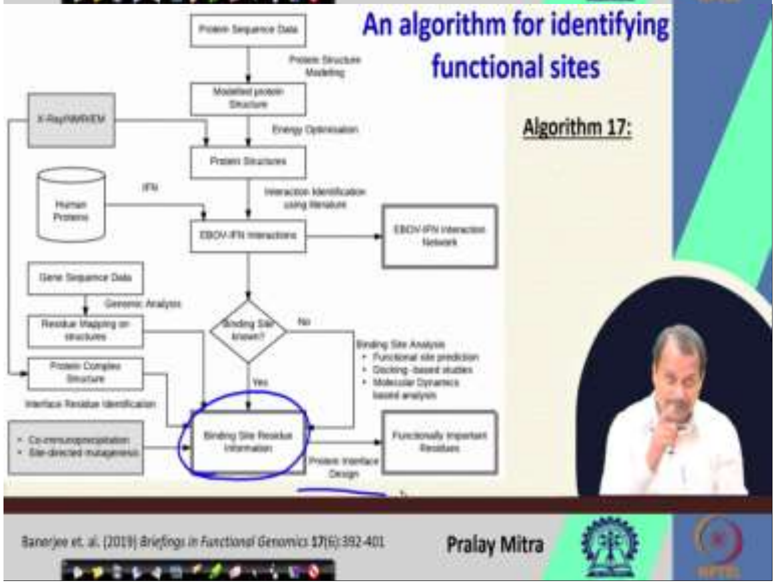
Seed

Random Sequence

Ab initio protein design

An algorithm for identifying functional sites

Algorithm 17:



Protein Sequence Data

Protein Structure Modelling

Modelled protein Structure

X-Ray/NMR/EM

Energy Optimisation

Protein Structures

IFN

Human Proteins

Interaction Identification using literature

EBOV-IFN Interactions

EBOV-IFN Interaction Network

Gene Sequence Data

Genomic Analysis

Residue Mapping on structures

Protein Complex Structure

Interface Residue Identification

Binding Site known?

No

Yes

Binding Site Analysis
- Functional site prediction
- Docking-based studies
- Molecular Dynamics based analysis

Binding Site Residue Information

Functionally important Residues

Protein Interface Design

- Co-immunoprecipitation
- Site-directed mutagenesis

Banerjee et. al. (2019) *Briefings in Functional Genomics* 17(6):392-401

Pralay Mitra

Next, let us go to another application of Ebola virus. So, this Ebola virus we can consider as a case study that we worked and published along back. So, there are, this is the Ebola virus and there are only seven proteins for this Ebola virus so, nucleoprotein here glycoprotein then three, then four, then five, six, seven and this matrix VP40.

So, if I consider then so 1 2 3 4 5 6 7 seven proteins are there and using these seven protein, you know that how deadly is the Ebola virus. So, the who situation report suggests that the fat mortality rate for this Ebola virus is very high yet, we have to design some drug or vaccine against this Ebola virus.

But if you look at the structure of this virus, then you will see it is very simple with respect to a human body. So, for us, for this human body, it is very sophisticated structure. So, several 1000s of proteins are there. Our so DNA sequences also billion base pair, 10 billion base pair. And for this, it is very small number size of gene, and also few numbers of proteins and that is enough to cripple down our immune system hijack our body's mechanisms.

So, that they will invade to our body, they will cripple our immune system, they will hijack our system and will replicate the cell. So, in order to design one drug molecule, it might be a good idea to look into the detail of this Ebola virus through some computational framework. As of now, we have studied protein design, protein docking and protein folding is it possible to combine those (())(15:35).

So, that computationally we can come up with some suggestion which can be taken by the biologist and they can work on that one because you know that this is very infectious virus. So, we cannot expect that every laboratory has a facility to perform the experiment for that, here we are proposing one flow diagram that flow diagram indicates that first I am starting with protein sequence data.

Then, we are going for protein structure modeling which means, I am applying protein folding, after applying the protein folding model protein structure I am getting I am doing some energy optimization because you know that structure that I have modeled is may have some steady classes and also all of them may not follow some or Ramachandran plot.

So, that way the model structure may not be very good candidate for experiment. It is better to go for some energy optimization and what it will do mostly it will use either replica exchange Monte Carlo or simulated aniline to adjust the side chain, So, this we discussed also in the context of fire doc.

So, it will adjust a side chain and after adjusting the side chain it will optimize its energy it will reduce the number of classes and it will fix if there is some bad phi psi angles after doing that one I got the protein structure then I will identify interaction using some literature information that whether that particular protein sequence, I am not exploding now or I am not informing now that what is going to be the protein sequence.

Right now, I am focusing on developing some general flow diagram now, that flow diagram may be extended for or maybe specialized or customized for one particular protein. Now, after identifying that now, I am focusing on EBOV-IFN interactions. So, this is the interferon protein with the EBOV we are locating that one and then we are identifying EBOV-IFN interaction network. So, when we identify that one then it is done.

So, here you can see that it is kind of a final step I have reached otherwise if I cannot go for that interaction network then I am checking whether binding site is known to me. So, if no then binding site analysis like functional site analysis, docking based studies, molecular dynamics analysis, all those I will perform and then I will look for some binding site residue information.

So, this is one information flow from here from the human protein the IFN proteins interferon proteins I am supplying to this box. Now, X ray NMR in EM. So, they are supplying the protein structure here as well as they are supplying protein complex structure here. Now, this gene sequence data after the genomie analysis and residue mapping on structures are contributing the information here then protein complex structure information is contributing here, interface residue identification, co-immunoprecipitation or CO-IP.

Then site directed mutagenesis all are contributing here with a hope that I will identify some binding site residue. Because you know that once I can identify the binding site residue then using protein design, sometimes I can do the miracle. This miracle you have seen when we discussed the glioblastoma on the last week in the application what we did, hexa protein matrix metalloprotein. We identified their binding site using the protein docking.

We picked those residues, we mentioned that those residues are relevant for the conservation of the function of the protein then we muted those residues and see that the functionality indeed are changing or previously the function was something and after we identify the crucial interaction sites by doing protein design, which will kill the interaction then we see that indeed the functionality has changed.

If it is then what we can propose design some drug molecule which will go and bind to that residue, which has been muted by protein design that way, it will kill the interaction of those residues with the MMP and the interaction will not take place. So, that was our proposal. So, that is why we are focusing more on identifying binding site information and we are exploiting as much information as possible from the literature, from the human protein, from the gene sequence data, from the PDB data X ray here it is given.

Then, we are looking for say other molecular dynamics and docking based functional site prediction. So, once we will identify that one then protein interface design or say using the protein design where we fixed that it is protein design, but you need not have to design everything, only the interface residues you need to design, using that one I will pinpoint the functionally important residues.

Now, at this point, I would like to mention a little bit more regarding this protein interface design. So, the same protein design technique either REMC version or Monte Carlo version you can use only thing you have to customize is given one protein sequence you need to design this one, once you need to design this protein sequence then you remember at the beginning, I mentioned you start with a seed sequence that is nothing but the random sequence and that is why the method was also Ab inito protein design.

So, all the residues in this protein are random in nature. Now, if you wish to design the interface, then for this particular protein you must be knowing what is the interface. So, it can be like this, because please note it down that interface all those structurally they are neighbor, but at the sequence they cannot be neighbor. So, the reason is very simple.

So, if you consider that this is my interface this side, so all the tips of my finger are the interface. So, this is the structural form now, if I open it and if I assume this is my sequence, this is the N terminal. So, N then going like this like this, like this, like this, and this is my C- terminal.

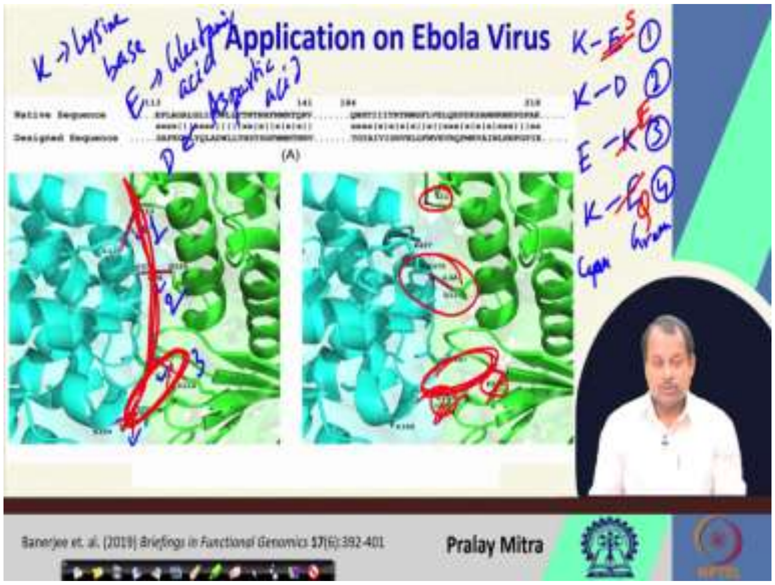
If I assume then you see this is my interface residue. This is my this is my this is my this is my and when they are in the structure form, then they are neighbor, but when they are like this then they are not neighbor. So, like that way so, here it is one stage, this is another stage this is another stage. So, you identify those stages. Once you will identify those stages then next step is that you keep all amino acids same except these interfaces.

So, what will be your seed sequence, your seed sequence will be the original sequence where randomly you mutate all the residues at the interfaces only. Then you give it to the protein design tool with an information that every time you mutate randomly in this sequence, the mutation will always take place either here or here or here in this blue regions only not in the red regions or I will block the mutations in the red regions and allow the mutations in the blue region.

If I do this one then you know that it will be finally come out one design sequence where red regions will be identical, only blue regions will be designed and that is what is expected. In this case also for this application. So, finally, when we will reach here after identifying binding site residue then protein interface will be designed and when protein interface will be designed only the blue regions just now I explained will be designed and rest of the red regions from the beginning will be keep intact will be kept intact they will not change they will be same.

So, if they are same then what will happen that once the design is over then only the blue regions will be designed, red regions will not be designed very simple. But doing some customization in your algorithm implementation you can achieve this one. So, you got functional important residue, you got binding site information, also you got EBOV-IFN interaction network. If it is useful for you for some purpose.

(Refer Slide Time: 26:20)



Application on Ebola Virus

Banerjee et. al. (2019) Briefings in Functional Genomics 17(6):392-401

Pralay Mitra

Now, using this one it is noted that there are some mutations indeed and at the interface, it is very interesting to note down on the left hand side you see that the 1, 2, 3, 4, four ionic bonds are there. So, ionic bonds are non covalent bonding. So, it occurs between acid and base. So, the first one is between so, K and E the second one is between K and D, third one is between so, this is 1, this is 2, this is 3, this is E and K, fourth one is between K and E.

So, this is 1, 2, 3, 4. So, this left hand side indicates cyan and right hand side indicates green. So, this is green and this is cyan. So, K means lysine base in all are cases and E Glutamic acid, D Aspartic acid, both are acid as the name suggests, on the right hand side you will see that what happens when the green one is allowed to muted at the interface region and then on the right hand side what you will find that here E has changed, this E has changed to S cyren with cyren licensed cannot form salvage then this D is retained.

So, there will be one salvage, then on the right hand side this 218 k has changed to E with E cannot form any salvage. Next at 203 position E has changed to Q with Q K cannot form salvage, there may be some small interaction but it cannot be a salvage. So, that way you see only one salvage retain rest has killed. Accordingly if you analyze the introduction energy computationally of course, you will see it has changed. So, you understand that this position, this position and this position are crucial for the interaction between these two proteins.

Now, if you design one small drug molecule which will go and bind at this region and that way, it will kill this in that way. So, go and bind at this region that way it will kill these two interactions definitely it will kill or it will inhibit their interaction. But this entire region is interface. So, try to understand. So, this entire region is the interface.

So, we are not going to design that. So that interface area is about 1000 angstroms square. So, that is huge, drug molecule cannot be large enough. So, only small thing I have to identify which is crucial for the interaction, if I kill, it will inhibit. That is it. So, that is it for today's lecture. So, we will again continue in the next lecture. Thank you.