Algorithms for Protein Modelling and Engineering
Professor Pralay Mitra
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur
Lecture 48
Application of Protein Design on Drug Design (Contd.), Protein Modification

Welcome back. So, we are continuing with this application of protein design. So, we are discussing the application in the context of drug design. So, in the drug design on the last lecture, I presented to you one algorithm which will combine protein folding protein docking and then protein design in order to identify the inhibitor binding site.

So, inhibitor binding site or say drug molecule binding site is the primary aim or main goal for any drug design technique. Once you will have that site then what you can do that you look for some drug molecule bind to that location again you can do this computationally and then you check that what is the effect of that binding in terms of the interaction energy, once the analysis is done probably you can go for experimental verification.

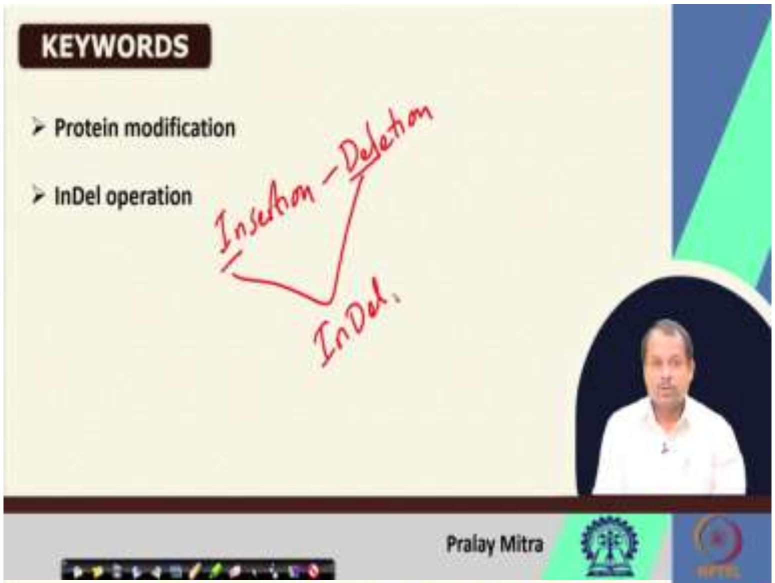
So, that we will consider, continue after that one then we will move on to protein modification which is part of kind of protein engineering. So, definitely protein design is a subset of the protein design, protein engineering, but other protein engineering techniques we will also discuss.

(Refer Slide Time: 01:34)

CONCEPTS COVERED

- Protein modification

Pralay Mitra



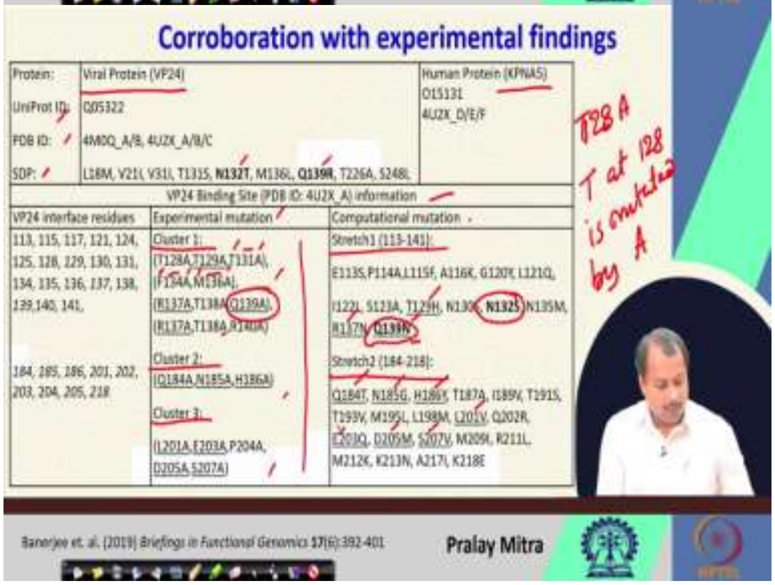
KEYWORDS

- Protein modification
- InDel operation

Insertion - Deletion

InDel.

Pralay Mitra



Corroboration with experimental findings

Protein:	Viral Protein (VP24)	Human Protein (KPNA5)
UniProt ID:	Q05322	O15131
PDB ID:	4M0Q_A/B, 4U2X_A/B/C	4U2X_D/E/F
SDP:	L18M, V21I, V31I, T131S, **N132T**, M136L, **Q139R**, T226A, S248L	

VP24 Binding Site (PDB ID: 4U2X_A) information

VP24 interface residues	Experimental mutation	Computational mutation
113, 115, 117, 121, 124, 125, 128, 129, 130, 131, 134, 135, 136, 137, 138, 139,140, 141,	Cluster 1: (T128A,T129A,T131A), (F134A,M136A), (R137A,T138A,Q139A), (R137A,T138A,R140A)	Stretch1 (113-141): E113S,P114A,L115F, A116K, G120Y, L121Q, I122L, S123A, T129H, N130K, N132S, N135M, R137N, Q139R
184, 185, 186, 201, 202, 203, 204, 205, 218	Cluster 2: (Q184A,N185A,H186A) Cluster 3: (L201A,E203A,P204A, D205A,S207A)	Stretch2 (184-218): Q184T, N185G, H186Y, T187A, I189V, T191S, T193V, M195L, L198M, L201V, Q202R, E203Q, D205M, S207V, M209I, R211L, M212K, K213N, A217I, K218E

T28 A

T at 128 is mutated by A

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

Pralay Mitra

So, the concepts we are planning to cover protein modification that will come after the we finish the discussion our protein drug design. Then keywords also we selected protein modification and InDel operation. So, InDel is the short from of insertion and deletion, that is the short form of the insertion and deletion.

So, that we will consider as of now, what we have done is just the mutation. But other kind of protein engineering like deleting one amino acid or multiple amino acids from some region or distributed over some sequences, those things can also be part of protein engineering. So, we will discuss that one. So, let us continue with the drug design part.

So, based upon the kind of algorithm we have designed, we noted that our prediction is corroborating with experimental findings. So, we identified one viral protein that is VP24 for Ebola virus and for human we identified KPNA5. So, they are UniProt ID, PDB ID. So, those are mentioned here. Now, VP24 binding site information is also provided.

Now, if I compare then I will see experimental mutations was performed and they identified three clusters here, here and here. So, this three cluster indicates three regions where the interface residues are there. So, one among them may be one of the possible drug targets. Now, what are the residues and what sort of mutations they have performed, they have mentioned it here.

Now, computationally, when we perform then we identify two stretch, stretch 1 and stretch 2 here and what is the sequence ID for them that is also mentioned here. Now, here, you will see with the SDP and the experimental mutation, so, when it is the computational mutation, as per the protein design method is matching with the experimental mutation, then those are marked by underline and then when it is matching with the SDP they are marked with bold.

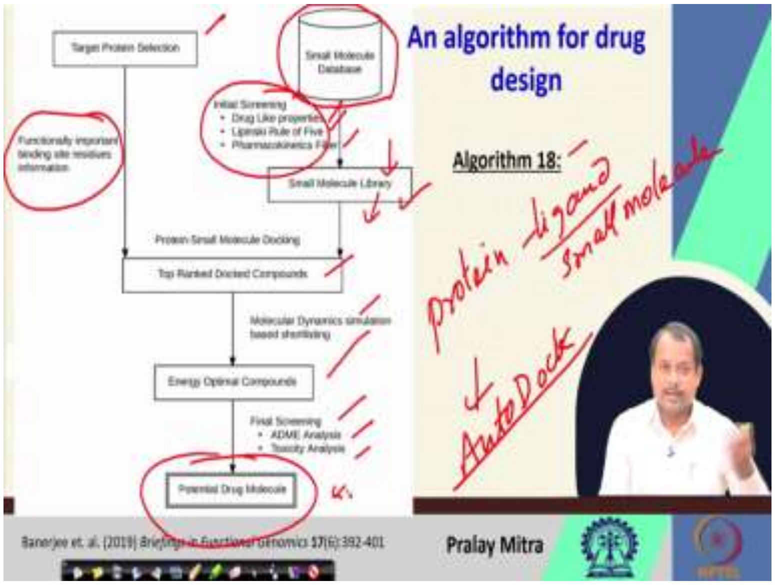
So, these two are matching with this one and this one are matching with SDP among which this Q139N is also matching with the experimental mutation. But one thing you should note that most of the experimental mutations are performed based upon the fact that one particular amino acid definitely that is non alanine amino acid will be mutated by alanine.

That is why if you look at this column, you see that all the mutations, so here, T128A indicates threonine at 128 position is mutated by alanine. So, like that way T129A indicates threonine at 129 is mutated by A, threonine at 131 mutated by alanine, phenylalanine at 134 is mutated by alanine.

Methionine at 136 is mutated by alanine that way. And last character for all the experimental techniques are A because an experimental technique most of the time the amino acids are mutated to alanine which is not the case for computational mutation. So, for computational technique there is as such no restriction in that way that is why you will find that several mutations are there and among those several mutations. These two are matching with the SDP and this underscore cases so, 1, 2, 3 for cluster 1 and for stretch 2 and 3 combined 1, 2, 3, 4, 5, 6, 7. So, they are matching

So, it is corroborating well with the experimental techniques. Now, the point is if you identify one small patch or region which will be good enough for one drug molecule binding then you can perhaps design some drug molecule that will go and bind there to inhibit the interaction and that way it can act as a drug for that particular disease.

(Refer Slide Time: 06:33)



An algorithm for drug design

Algorithm 18:

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

Pralay Mitra

So, here is your algorithm for drug design. So, I am starting with the target protein selection. So, I know that what are the functionally important binding site residues because just before this algorithm 18 I run algorithm 17 and the last lecture that algorithm 17 is there were started from the protein sequence then the interaction information either from the literature or from experimental technique then doing a lot of analysis, docking design, et cetera.

I identified functionally important binding site residues. And also the target protein structure is known to me. Then I will also have small molecule database where I will perform initial screening drug like properties, Lipinski Rule of Five, Pharmacokinetic Filter. So, after applying all those I will have or I will do prepare one small molecule library.

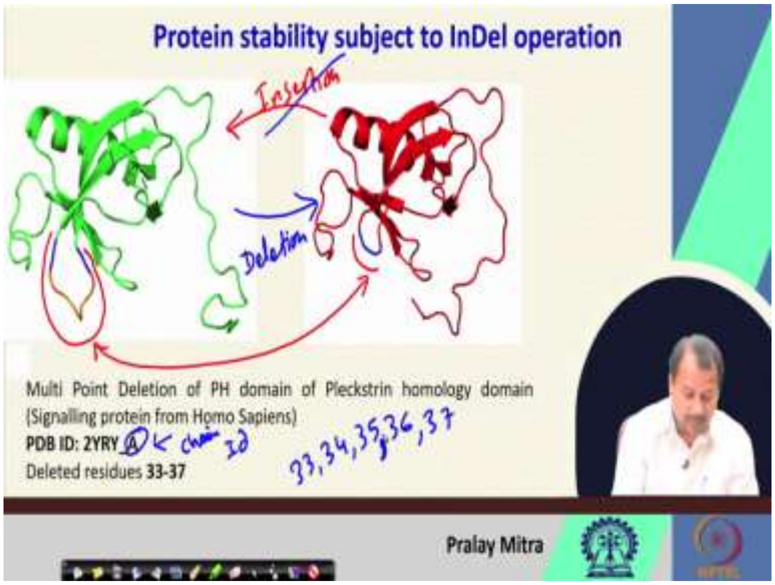
So, from the small molecule database using this initial filtering technique, I will prepare one small molecule library for my purpose then small molecule library and the target protein selection will be docked. So, these docking we did not discuss, this is protein ligand docking, this ligand is the small molecule. So, one of the most widely used method for this is Auto Dock.

Now, what is parallel version of Auto Dock is also available so, that you can go for first docking of the molecules. Now, after the docking procedure, you will have top ranked dock compound. So, this technique is similar to the last step of our protein protein docking technique also where after the scoring you will go for the ranking and then from the ranking you will get ranked dock compound.

Then you run molecular dynamics simulation based shortlisting to check that indeed that will optimize and that may be experimentally also stable then you will perform energy optimal compounds, then final screening where ADME analysis Toxicity analysis will be performed and then you will output potential drug molecule and this output is through computational framework or computational tool.

Now, after this if it is required then it may be synthesized in the chemical laboratory or you may purchase that one then you go for in vivo or incel or patient sample data testing and after that one you move on to pharmaceuticals company for marketing the drug molecule. So, these other steps.

(Refer Slide Time: 09:42)



Protein stability subject to InDel operation

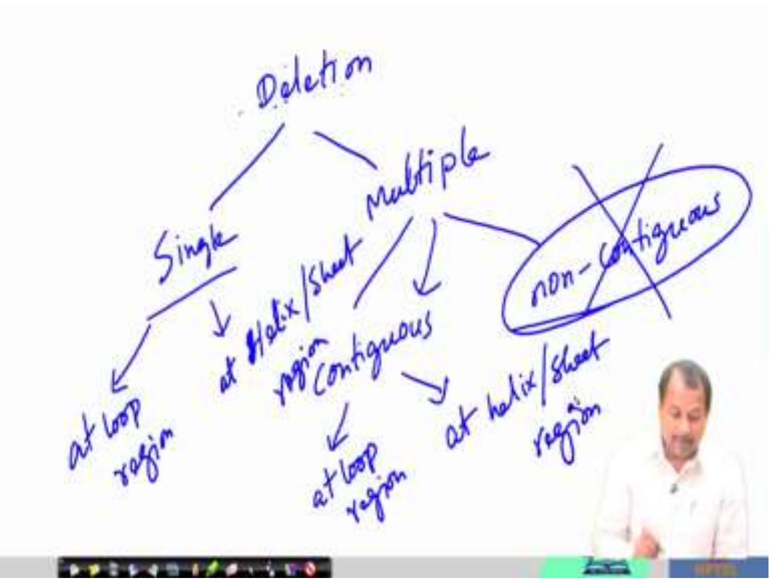
Multi Point Deletion of PH domain of Pleckstrin homology domain (Signalling protein from Homo Sapiens)

PDB ID: 2YRY A ← chain 32

Deleted residues **33-37**

33,34,35,36,37

Pralay Mitra



Next the topic we are going to discuss is another protein engineering problem that is insertion deletion operation or in short InDel operation. Now, given this protein on the left hand side the green color protein is given to you and on the right hand side the red color protein is given to you, if you look at this then do you able to notice any differences between these two protein molecules, if you look closely, you will see that this region has been sort in to this region.

So, this has been done. On the green, on the left hand side protein molecule the green loop region, sorry the golden loop region has been deleted. Now, in insertion and deletion, insertion indicates some amino acid has been inserted, deletion indicates some amino acid has been deleted. Now, you understand that if red is your input then and green is your output then that means insertion.

If you move this way then this is insertion. So, as if this golden color is being inserted in the red protein molecule in order to get the green one and this is your deletion. When golden loop region will be deleted from the green protein molecule then you will get red one. Now, it depends from which side you are moving on another side that is why instead of having both insertion and deletion, we will discuss only the deletion, insertion we will not discuss.

But you understand that when we will move from say green to red and we will call that as a deletion then when you will move from red to green then it will be the insertion. So, we will consider the deletion. In this case multiple amino acids has been deleted. So, example is taken from multi point deletion of PH domain of Pleckstrin homology domain signalling protein from

Homo Sapiens, PDB IDs 2YRY, chain ID is A this is my chain ID and deleted residues are 33, 34, 35, 36, 37. So, five residues was deleted.

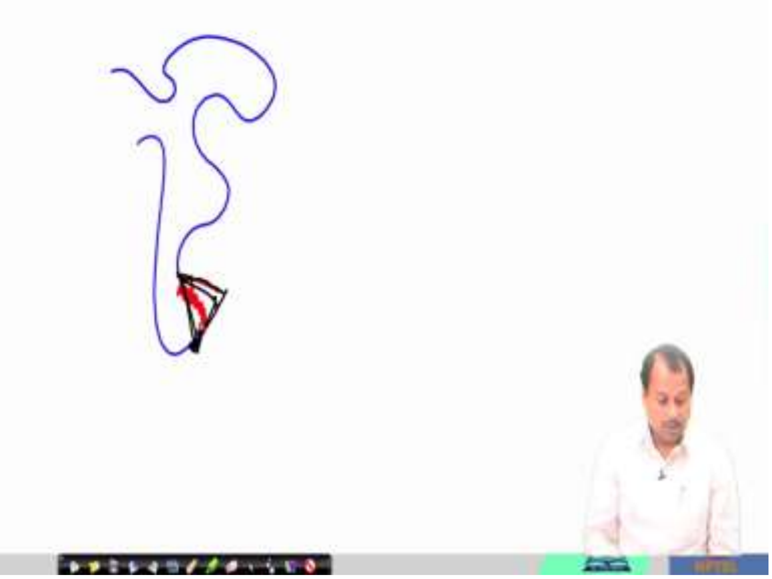
Now theoretically it may possible that one amino acid will be deleted deletion. So, single multiple, I mean the deletion may be in one residue only or it may be in multiple residues. Now, if it is only single then it is fine. But, if it is multiple then there maybe two possibilities theoretically, contiguous, non contiguous which means.

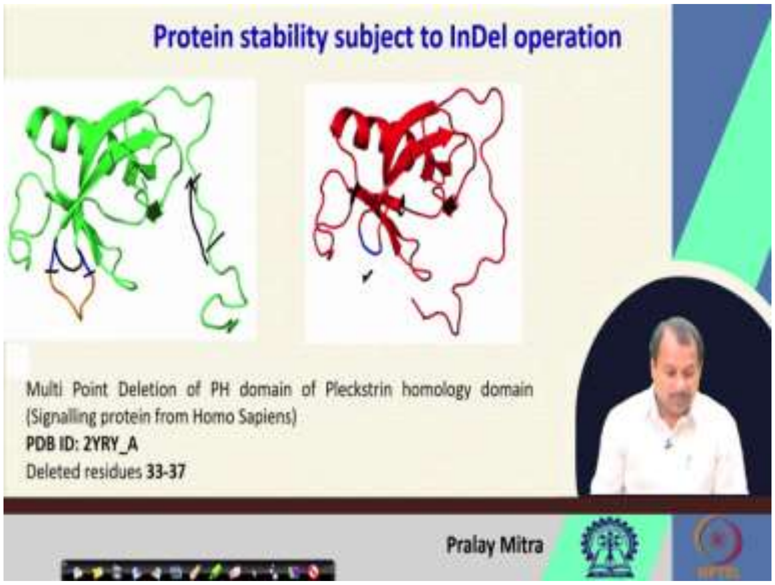
The deletion if say five residues has been deleted that five residue can be like that 33 to 38 that region or it can be say one residue from here, two residue from here, two another residue from there that way also that can be deleted. So, we will not consider this non contiguous work, because it is not that much simple to deal with.

We will focus on single and multiple contiguous and also if you look at the problem nature or the characteristic of the problem, then you will understand the single and single contiguous are not the same. Those two are two separate problems. Why I will come to that, but before that I wish to tell you even there can be two different variations for this single and contiguous both. One is at loop region at helix or sheet region.

So, it is needless to mention that if it contains helix and sheet both then it will be further complicated. So, these are the categories of the deletion that we need to deal with separately. But I mentioned that I will tell you that single point and multi points are not same, why?

(Refer Slide Time: 15:30)





Protein stability subject to InDel operation

Multi Point Deletion of PH domain of Pleckstrin homology domain
(Signalling protein from Homo Sapiens)
PDB ID: 2YRY_A
Deleted residues **33-37**

Pralay Mitra

So, let us consider one protein structure something like this. Now, if I remove one amino acid then after removing that one then what I need to do after the deletion I need to stretch this and this and then I need to join them because one amino acid has been deleted. So, I need to stretch that one.

Now, regarding the stretching, if it is in the loop region which is mostly floppy in nature, then stretching is not very difficult easily I can do that stretching. But if the stretching is at the secondary structure level, then it may not be that much easy also and also since there is a regularity or pattern in the hydrogen bond in helix and sheet.

So, if I delete some residue from in between, then the pattern of the secondary structure formation will also change which is also not expected. Now, this is about single point one. Now, instead of this small red, if I am interested to say delete this much it is a long one. So, multiple amino acid deletion if it is the case, then again, I have to do what I have to connect this one sorry, black color and this one.

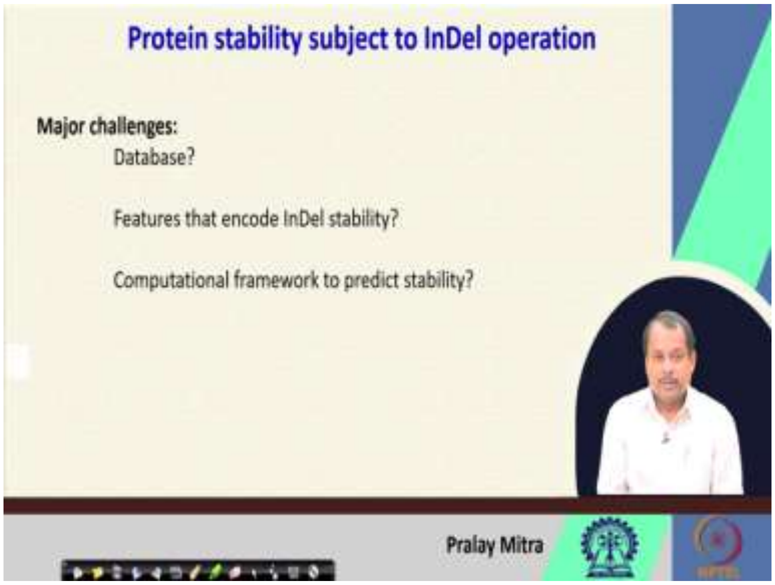
Now, for one single amino acid, it is bit easier compared to the multiple amino acid. Say, if five such amino acids are deleted, then if I stretch it, so, automatically it will not come. So, because you understand that the energy which is stored in the folded protein structure is enormous. If I wish to alter that structure, it is not very easy and also, I am trying to stretch it out so, that it can connect here.

So, that is also not a good thing. Good thing means the from stability point of view. Now, if I go back to the structure, now, you see that here so, it is at the loop region. So, if I delete from here to here and then try to connect then it will be loop it will be connected like this and you see the connection is something like this.

But instead, if I say wish to delete say from here to hear after deletion so, I have to take this all the way up to here how and also, I have to check the stability. Somehow, I have to do that one. So, that way single point and multipoint are two different stories and two different ballgames and also if I consider that because of the deletion say one small this sheet completely eliminated or deleted, then for single point deletion hardly we can expect that one secondary structure will be lost completely.

But for multipoint deletion, it may happen. So, that is why I am explaining that single point deletion and multipoint deletions are two different ballgames we need to tackle them separately. Needless to mention that we our say feature set or our methodology for the multi point deletion will be a superset of single point deletion. So, a lot of the features that we considered for single point deletion will be utilized for multi point deletions, but we need to deal with them separately. So, with this, let us move on to our single point deletion stuff.

(Refer Slide Time: 19:45)



Protein stability subject to InDel operation

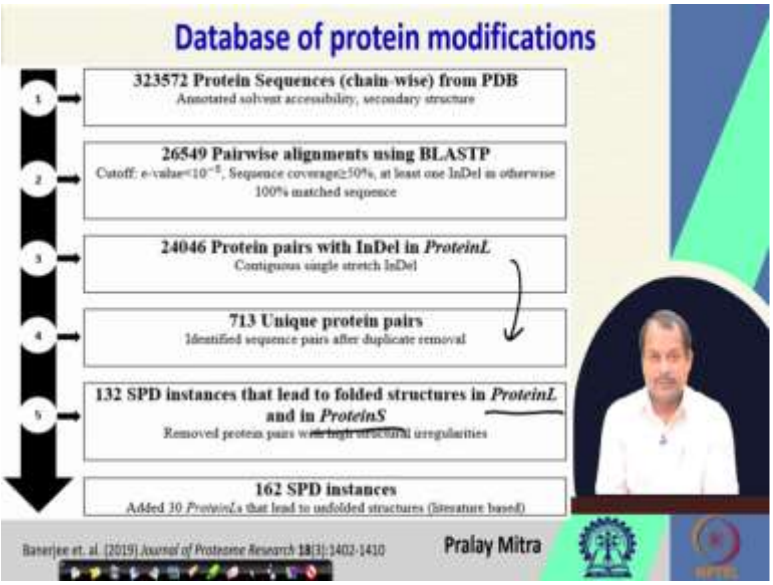
Major challenges:

Database?

Features that encode InDel stability?

Computational framework to predict stability?

Pralay Mitra



Database of protein modifications

1 → **323572 Protein Sequences (chain-wise) from PDB**
Annotated solvent accessibility, secondary structure

2 → **26549 Pairwise alignments using BLASTP**
Cutoff: e-value<10^{-5}, Sequence coverage≥50%, at least one InDel in otherwise 100% matched sequence

3 → **24046 Protein pairs with InDel in *ProteinL***
Contiguous single stretch InDel

4 → **713 Unique protein pairs**
Identified sequence pairs after duplicate removal

5 → **132 SPD instances that lead to folded structures in *ProteinL* and in *ProteinS***
Removed protein pairs with high structural irregularities

162 SPD instances
Added 30 *ProteinL*s that lead to unfolded structures (literature based)

Banerjee et. al. (2019) *Journal of Proteome Research* **18**(3):1402-1410

Pralay Mitra

When we are trying to go for this InDel operation irrespective of the single point or multi point there are a few challenges. First of all, databases are not there that much for this deletion operations. Now, you know that when there is no database, so, benchmarking your method, whatever method you will develop or if you go for some machine learning technique then training, testing so, that is not that much easy.

So, it is always a tough job next features that encode inDel stability, so, that is also not much available. So, you have to build it from the scratch, next computational framework to predict stability that is also you have to develop. So, we will address these three challenges during the development. So, first let us start with a database of protein modification.

So, what we can do. First, we will look in the protein databank and we will identify all the subunits or all the chains from the protein databank which are annotated solvent accessibility secondary structure, we will see that one during this process definitely, we will not consider whose resolution is say more than 2.5 or our factor is greater than 0.2.

Because we mentioned that if the resolution and R factor is not within a limit, then atomic level resolution does not guarantee that very good structure we are getting so, maybe there are some short contracts there may be some classes and refinement is not good. So, those kinds of problems may happen.

Now, after that, when we apply the blast for pairwise alignment of those sequences, because our intention is to identify two similar protein sequences as well as the structure who differs by only one amino acid, then only I can say, one is the deletion partner, single point deletion partner of the another. So, that is why we applied the BLASTP on all that data.

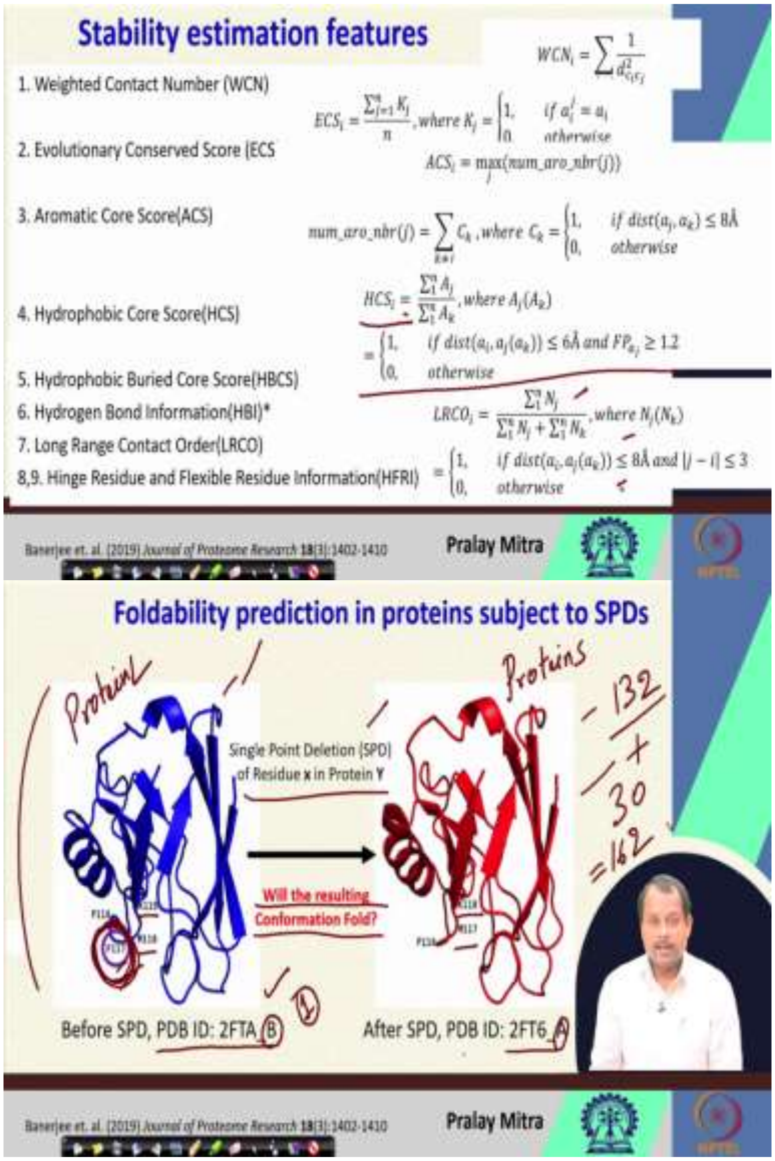
So, it is time consuming step nevertheless, it is the pre processing stage and you need to do only once during the creation or preparation of your data set. While I am doing the BLASTP, we so cut-off e value I am using 10 to the power minus 5, sequence coverage greater than 50 percent at least one InDel in otherwise 100 percent matched sequence.

So, those conditions we imposed and that way we are just reducing from 323572 protein sequences to 26,549 pairwise alignment, it will further reduce if I impose the condition thus contiguous single stretch InDel are required. So, it will reduce to 24,046 protein pair. Next, out of these 24,046 protein pairs, only 713 unique protein pairs are identified whose sequence pairs are not duplicate.

So, after duplicate removal, we are moving from third step to fourth step. Next from the fourth step, we have reduced to 132 SPD or single point deletion instances that lead to folded structure and we removed protein pairs with high structural irregularities. So, this proteinL and S indicates, so L indicates where there is no deletion, S indicates from the proteinL at least not at least exactly one amino acid has been deleted and then you got the proteinS.

So, proteinS and proteinL are a pair and they differ by only one amino acid which is removed or absent in proteinS and which is very much present in proteinL. Now, with this 132 SPD instances we added 30 proteinLs that lead to unfolded structure based upon the literature. Hence, we are having 162 SPD instances of which 132 instances are positive I mean, where there is a proteinL and proteinS pair and 32 are the unfolded structure which means they will be considered as a negative data.

(Refer Slide Time: 24:40)



Stability estimation features

1. Weighted Contact Number (WCN)
2. Evolutionary Conserved Score (ECS
3. Aromatic Core Score(ACS)
4. Hydrophobic Core Score(HCS)
5. Hydrophobic Buried Core Score(HBCS)
6. Hydrogen Bond Information(HBI)*
7. Long Range Contact Order(LRCO)
8,9. Hinge Residue and Flexible Residue Information(HFRI)

$$WCN_i = \sum_j \frac{1}{d^2_{c_i c_j}}$$

$$ECS_i = \frac{\sum_{j=1}^{n} K_j}{n}, where\ K_j = \begin{cases} 1, & if\ a_i^j = a_i \\ 0 & otherwise \end{cases}$$

$$ACS_i = \max_j(num\_aro\_nbr(j))$$

$$num\_aro\_nbr(j) = \sum_{k \neq i} C_k, where\ C_k = \begin{cases} 1, & if\ dist(a_j, a_k) \leq 8Å \\ 0, & otherwise \end{cases}$$

$$HCS_i = \frac{\sum_1^n A_j}{\sum_1^n A_k}, where\ A_j(A_k) = \begin{cases} 1, & if\ dist(a_i, a_j(a_k)) \leq 6Å\ and\ FP_{a_j} \geq 1.2 \\ 0, & otherwise \end{cases}$$

$$LRCO_i = \frac{\sum_1^n N_j}{\sum_1^n N_j + \sum_1^n N_k}, where\ N_j(N_k) = \begin{cases} 1, & if\ dist(a_i, a_j(a_k)) \leq 8Å\ and\ |j - i| \leq 3 \\ 0, & otherwise \end{cases}$$

Banerjee et. al. (2019) *Journal of Proteome Research* **18**(3):1402-1410

Pralay Mitra

Foldability prediction in proteins subject to SPDs

Single Point Deletion (SPD) of Residue x in Protein Y

Will the resulting Conformation Fold?

Before SPD, PDB ID: 2FTA (B) (1)

After SPD, PDB ID: 2FT6 A

Banerjee et. al. (2019) *Journal of Proteome Research* **18**(3):1402-1410

Pralay Mitra

Next, we will look at the features. So, stability in estimation features are like weighted contact number, then evolutionary conserved score, then aromatic core score, then hydrophobic core score, then hydrophobic buried core score, then hydrogen bond information, then long range contact order, then hinge residue and flexible residue information.

So, these are the features mostly is being considered for the machine learning based algorithm. So, weighted contact number is very simple. So, it is the summation of 1 divided by d square ci cj. Now, evolutionary conserved score or ECS is the summation of Kj divided by n where Kj is equals to 1 if aij equals to ai or 0 otherwise ACS is the aromatic core score.

So, it is the maximum of aromatic a number of amino acids there, then hydrophobic core score HCS is computed HCS is computed, HCS is computed, then hydrophobic buried core score that is computed here. So, all are simple equations and you can able to compute the score corresponding to one pair of protein structures.

Now, we have 162 protein structures of which 30 proteinL is present and its corresponding proteinS if you consider 1 then that does not exist I mean that when you delete one amino acid from there then it will unfold. On the other hand, 132 cases are there where from proteinL if you remove one amino acid then it will retain its structure.

So, that way you are defining so, this nine features then, we wish to have one prediction system where input will be blue color protein, this blue color protein and where there is an amino acids F117 we will ask the question will the resulting conformation fold means if I remove F117 then will it fold definitely after removing this one so, corresponding is blue color structure, you have a sequence.

Now, after removing this F117 you will have a new sequence if you feed that sequence to some protein folding software, then that software will give you one structure based upon that one you can have this conclusion whether the resulting conformation fold. But that is time consuming, it will take several hours.

On the other hand, this particular machine learning based algorithm what they can, it can able to do based upon the features that we have described on this blue it will compute that one it will take this F117 and then it will predict that whether after the deletion of F117 this red color structure is going to be stable or not.

Since I deleted say F117 that is why here instead of M118, it is M117 and instead of K119, it is K118. So, this is the PDB ID 2FTA chain B this is PDB ID 2FT6 chain A. Now, single point deletions of residues X in protein Y it is demonstrated here. So, what we are trying to do now is to come up with some machine learning technique that machine learning technique will take one protein structure as an input and this site or the amino acid that needs to be deleted and then it will compute the set of features that we have shown you on the last slide.

After computing that features then it will predict whether after deletion of these F117 this blue structure is going to be stable or not. So, that is the proposal and that we are going to perform. Now, we are having 132 positive instances where proteinL and proteinS here this is proteinL and this is my proteinS.

So, 132 such cases are there and 30 cases where proteinL is there and 1 for one particular position. If I delete then proteinS will cease to exist or it will unfold. Now with that 132 plus 30 that means 162 data and the features we need to come up with some machine learning techniques. So, now, so let us stop here with this lecture, we will continue this discussion to the next lecture. Thank you very much.