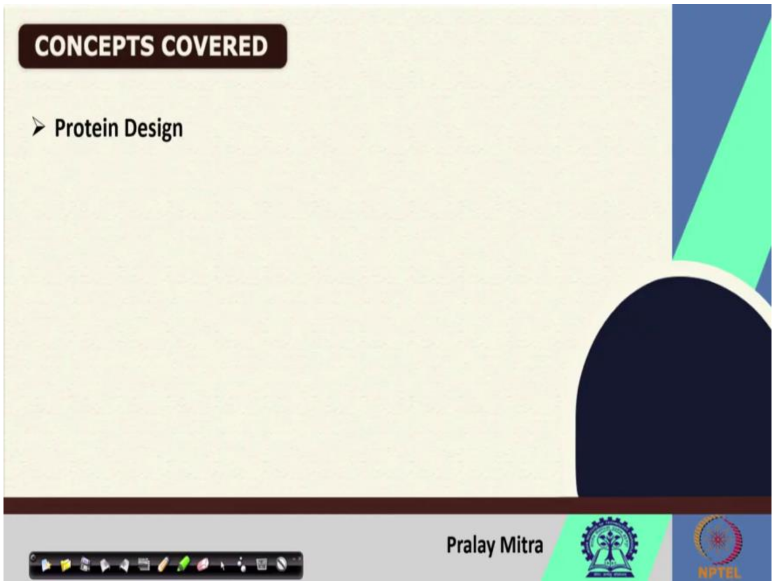
Algorithms for Protein Modelling and Engineering
Professor Pralay Mitra
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur
Lecture 41
Computational Protein Design (CPD)

Welcome back to the course on algorithms for protein modeling and engineering. So, in this week, we will start a very new and challenging topic that is protein design. So, people are working on these, protein design since long and most of the work was experimental. Now, for the last one decade or say starting from 2005 and 2006 people started to look whether there can be a computational alternative like say protein folding and protein docking for which computational alternative was already in place, is it possible that for protein design also we can have some computational technique needless to mention, as I said that, these computational techniques may not be very 100 percent accurate.

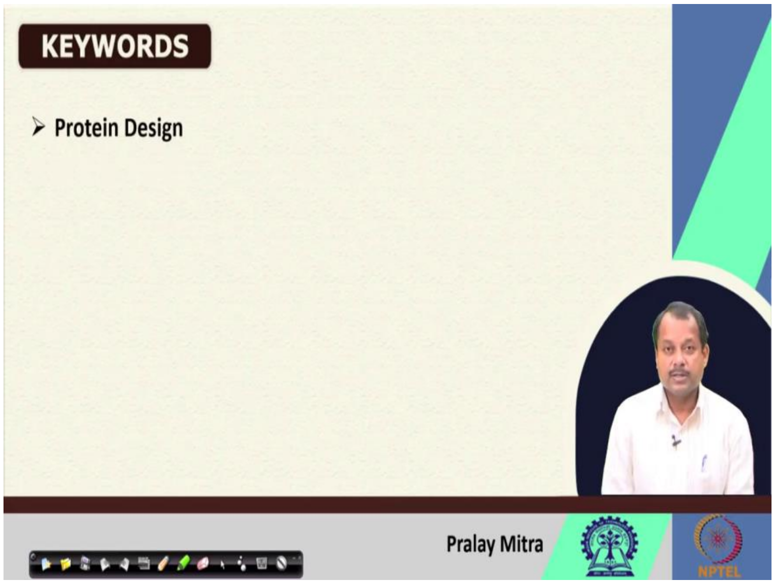
However, this computational technique can give you guidance or, can give you some probable solutions among which you can pick one for your own purpose. So, this protein design has a lot of applications, the application ranging from drug design new molecule design then in industry, so, everywhere you will find there is a application of protein design. Now, in this topic, in this week, we will start discussing about the computational framework or alternative for the protein design problem.

(Refer Slide Time: 01:49)



CONCEPTS COVERED

- Protein Design



KEYWORDS

- Protein Design

Pralay Mitra

So, we will cover the concepts on protein design and that is why the key word also I have picked as protein design, this protein design includes in detail the definition and the algorithm building, et cetera I will go shortly, but this protein design includes that given a protein structure, we need to design a sequence which will also fold to that structure.

So, in to make it more, simple for you that when one protein structure is given to you, which means the sequence is also known to you now, the sequence is also known to you, indicates that the sequence is there. Now, this protein design will look for some alternative sequence which may not exist or probably will not exist in nature, but that sequence will also fold to the given protein structure.

So, that is all about the protein design. However, there are some variations in the protein design problem. So, various includes that say when I will mutate I mean in one position I will replace one existing amino acid by another amino acid then what will be the situation if one amino acid if I delete one amino acid from some position if I delete more than one amino acid from some location, so, what will be the situation.

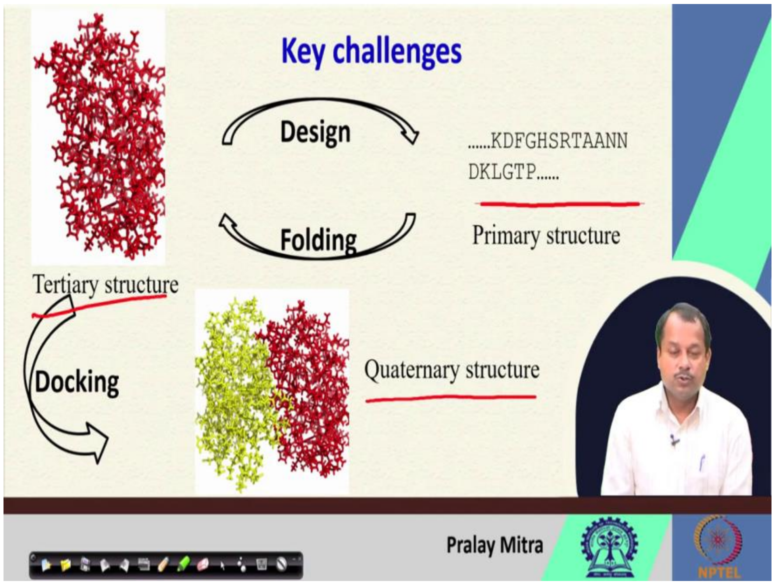
So, everything includes as part of the protein design. Now, these regarding deletion, et cetera you will see that during the evolution process mutation or deletion may happen and if it happens, then what possibilities are there that after that particular modification or such changes either say mutation or insertion or deletion the protein is going to be stable. So, if I know that one then probably I can take some precautions or I can take some idea.

So, why it is precautions? So, for example, let us assume that there is one virus and that virus is causing harm to human body we wish to design some vaccine or drug molecule corresponding to that virus, but that virus is changing it is evolving and because of that one it may possible that it will change its protein or sequence at the protein level I am talking about, but if there are some RNA viruses are there.

Then at the RNA level also they can change, but since we are discussing in the context of protein then let us assume that because of their evolution or mutation, so, at the protein level there will be some changes. Now, if we know a priori that what kind of changes may happen, and the changes is going to be stable then why designing the drug or the vaccine then beforehand we can take precautions.

I mean we can also look for those drug molecule or vaccine or what kind of drug molecule or vaccine we are designing we can include those possibilities in our discussion, and we can include those possibilities in our calculation and accordingly we can be prepared for that one.

(Refer Slide Time: 05:30)



Key challenges

Design

Folding

......KDFGHSRTAANN
DKLGTP......

Primary structure

Tertiary structure

Docking

Quaternary structure

Pralay Mitra

NPTEL

So, let us start so, we mostly revolving around three protein structures, primary protein structure, tertiary protein structure and quaternary protein structure. So, as I mentioned the primary structure indicates the protein sequence or it is also called as a primary sequence. Tertiary structure indicates that when one protein sequence takes some shape incidence of space and I represent that one at that atomic level resolution then what will be the structure.

So, that is my tertiary structure and the quaternary structure is the functional form one protein molecule I mean single chain or single subunit whatever I say they cannot have any function. So, for example, if I assume that I am sitting in this room and there is nobody present in this room and also I do not have say computer smart phone or any other means to do anything.

So, in that situation I will not able to function or interact with anything, but the moment some smart phone will be given to me or laptop will be given to me then I will start walking in that or if somebody else walk into this room then I can start talking or start discussing with him then the interaction will start.

Similarly, if there is only one subunit then it will have some structure it will have some shape but it cannot have function, function will happen when it will interact with other proteinous or non proteinous small or large molecule. So, in this case also you can see that on that top right corner, so, here there is primary structure.

So, here there is primary structure. Now, this primary structure if it will take some space (())(07:52) space then it will be changed to tertiary structure and when this two tertiary structure will be put together oil or will come in close contact then they will interact and that way I will get the quarterly structure.

Now, we discussed extensively that protein folding problem where the input is a protein sequence and output is a tertiary structure a lot of softwares a lot of algorithms exist out of which we developed I mean we discussed one Abinitio protein folding technique and we also discussed the replica exchange Monte Carlo method how it can be exploited in that context.

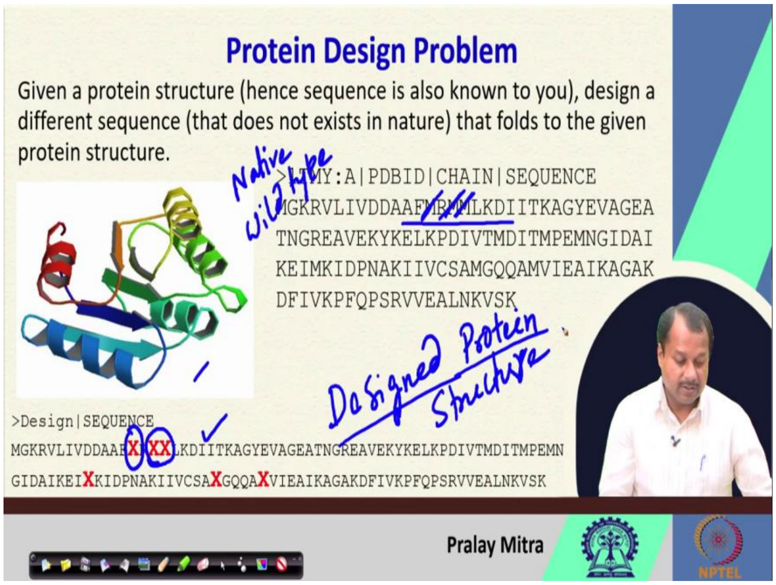
Then extensively we also discussed given to protein structure as an input then how can I compute the quaternary structure out of that one. So, that is called as a protein docking problem. So, we discussed several docking techniques starting from (())(09:02) dock, J Trank, then FT dock then patch dock then fiber dock, fire dock, then we discussed sim dock which is symmetry based docking. So, we discussed all those things.

Now, we are going to discuss protein design which sometimes is called as the inverse protein folding problem, why you see the diagram here input is a protein tertiary structure and output is going to be the primary structure or the amino acid sequences. It is doing just opposite of what protein folding is supposed to do where given one sequence you need to predict one structure, but you should keep in your mind that when say I am going for inverse protein folding or protein design problem. Then starting from one structure I am predicting one sequence which is not present in nature.

The primary reason is that when structure is known, then one sequence is also known that sequence fold to the structure. So, it is a mapping from structure to sequence, but the sequence is not the same as the sequence which represents the structure. So, it will be a changed version of the sequence and by this time we know that and anfinsen hypotheses is not always correct.

So, corresponding to multiple sequences there is a possibility that they will fold to the same structure. So, protein design probes to identify those sequences which will fold to the same structure out of which one sequence and the structure is known to you, you need to explore other sequences also which will fold to that structure.

(Refer Slide Time: 11:05)



Protein Design Problem

Given a protein structure (hence sequence is also known to you), design a different sequence (that does not exists in nature) that folds to the given protein structure.

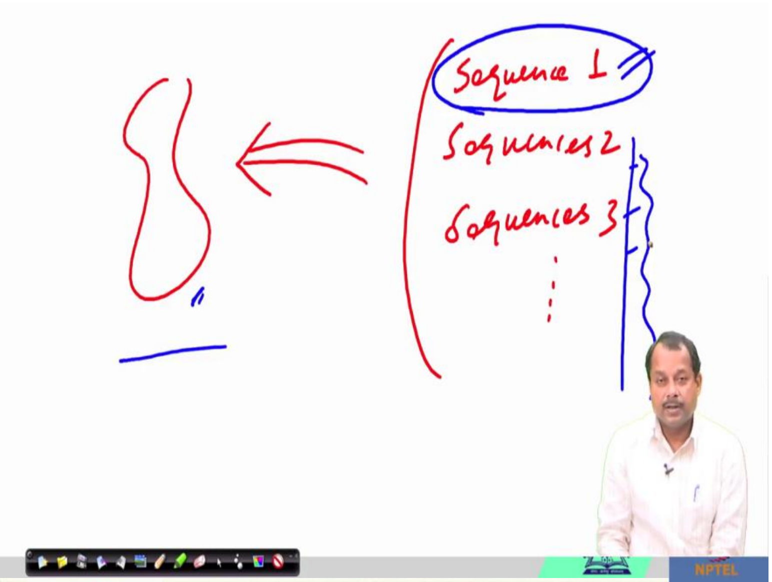
Native wild type

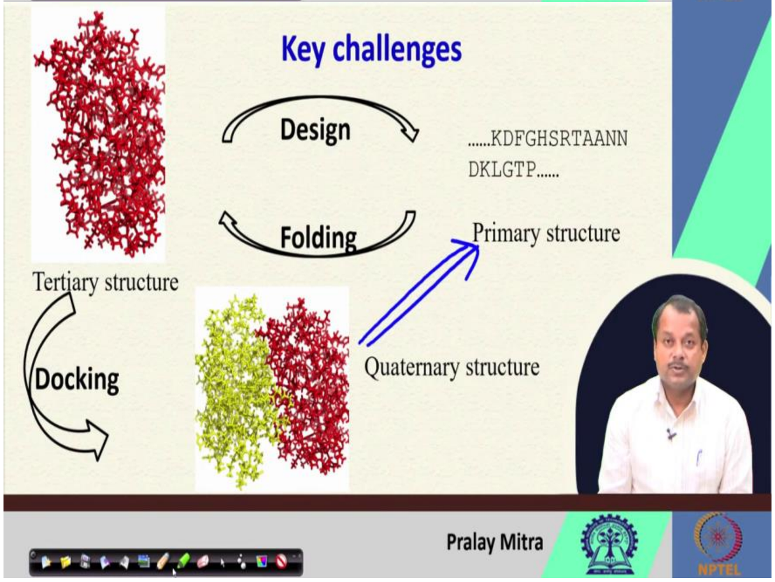
```
>1TMY:A|PDBID|CHAIN|SEQUENCE
MGKRVLIVDDAAFMRMMLKDIITKAGYEVAGEA
TNGREAVEKYKELKPDIVTMDITMPEMNGIDAI
KEIMKIDPNAKIIVCSAMGQQAMVIEAIKAGAK
DFIVKPFQPSRVVEALNKVSK
```

Designed Protein Structure

```
>Design|SEQUENCE
MGKRVLIVDDAAFXRXXLKDIITKAGYEVAGEATNGREAVEKYKELKPDIVTMDITMPEMN
GIDAIKEIXKIDPNAKIIVCSAXGQQAXVIEAIKAGAKDFIVKPFQPSRVVEALNKVSK
```

Pralay Mitra





Key challenges

Design

Folding

......KDFGHSRTAANN DKLGTP......

Primary structure

Tertiary structure

Docking

Quaternary structure

Pralay Mitra

Formerly we can say the protein design problem is given a protein structure hence sequence is also known to you please note it down design a different sequence that does not exist in nature that folds to the given protein structure. So, what I said previously is one structure is given to you hence, you have one protein sequence say let us assume sequence now, from the study of that homology and we also observe that there can be multiple sequences with some variations which will fold to this particular structure.

Now, out of those let us assume without any loss of generality that sequence one represent this particular structure. So, when this structure will be given to you as an input then by just looking at the sequence which represents the structure you will get sick one what at the same time there may be a number of other sequences also which is also supposed to fold to this particular structure our job is to identify all those other sequences.

So, as I mentioned there are a lot of applications we will explore one after another and we will see sometimes this particular structure sequence combination may not able to do one particular function which can easily be done by other sequences and also there is an extension that we will explain in the context of open areas or challenging problems that when say I wish to move from let me go back I wish to move from quaternary structure to sequences.

So, that we will pose as an open problem or challenging problem whose solution some solutions are there, but one unified computational framework as of today is yet to come. So, we define protein design problem pictorially it will be something like this given this particular structure whose which is taken from the protein databank its PDBID is 1TMY and only chain A is considered and what is the sequence it is denoted here it is in the FASTA format.

That is why the first line starts with greater than followed by my comment in the comment I prefer to keep track of from which PDB ID I have extracted that sequence that is why I kept 1TMY as the ID colon a indicates I took chain A then PDBID chain sequence. This is my sequence. So, protein design says, this structure which is given in this slide is given to you as an input which means, what is the sequence composition that composition is also known to you. It is on the right-hand side of the structure.

Now, your job will be to come up with some sequence which is given at the bottom where rate large cross indicates the positions where there will be some changes means at that position the amino acids which are there will not be there rather new amino acids will come after that change if I wish to allow the sequence to fold then also it will fold to that given structure. So, my point is that so, at this position at this position whatever is here. So, it is DDAE so, DDAF then MR so, this MR after that MM. So, these three positions in the native sequence so, this I am calling as native or wild type because it exists in nature.

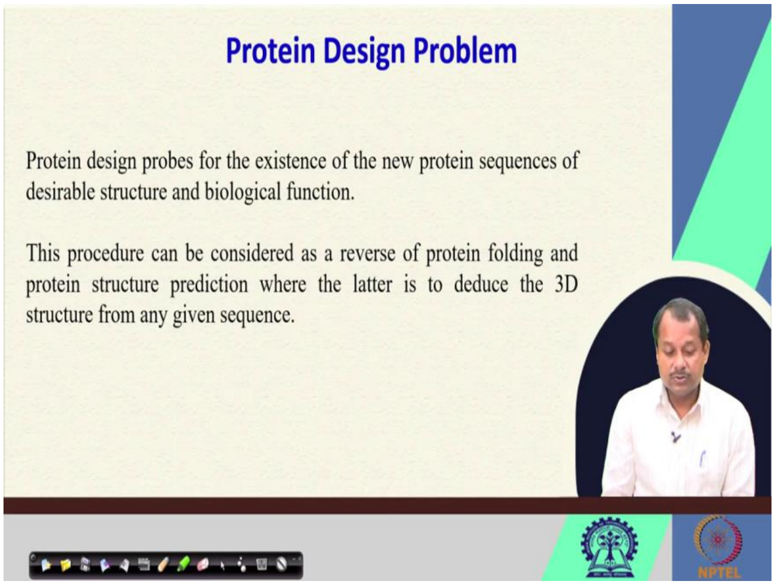
So, this M MM or methylene will be replaced by some other residue then also it was supposed to fold to this particular structure if he is then you understand that this sequence is going to be a new sequence this sequence after changing this M MM. So, this sequence is going to be a new sequence and that structure is called as that designed structure or you can say designed protein structure did you heard this design protein anywhere before.

Yes, you go back to your structural classification of proteins lecture scope there one extra field has been incorporated apart from alpha beta alpha and beta alpha or beta. So, that is designed sequences several other has been incorporated like membrane protein, et cetera but one is also designed sequences. Now, in this design sequence category you see that small changes are there because of that one it is supposed to take the same structure. So, as far the structural classification of protein its structure will be same.

However, we will see why we will discuss about the application of the protein design that because of the small changes and the position of the changes specifically the function may change and while our intention during the scope classification was to keep in one class, which has one particular function and since there is a direct relationship between the structure and function.

So, we are classifying according to the structure but it is also classifying according to the function but, when we will design the sequence then what will happen that the small changes may not affected structure but may change its function, that is why to keep a flag that these are kind of a new guys which may not respond correctly in the structure function relationship. So, one separate category with the name design sequence or design protein has been created there. Now, I believe you can able to make a link between that design protein at the scope and the definition of the design protein here.

(Refer Slide Time: 18:45)



Protein Design Problem

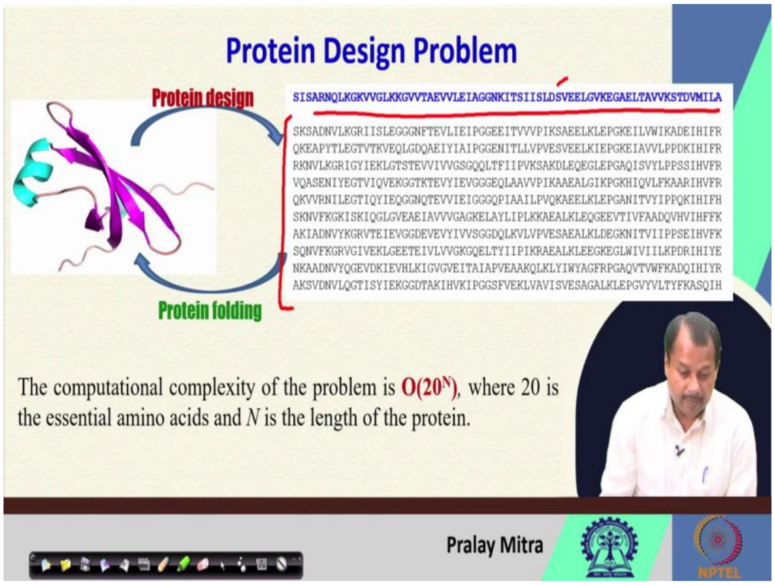
Protein design probes for the existence of the new protein sequences of desirable structure and biological function.

This procedure can be considered as a reverse of protein folding and protein structure prediction where the latter is to deduce the 3D structure from any given sequence.

So, protein design probes for the existence of the new protein sequences of desirable structure and biological function. So, when I say desirable structure and biological function that means you can very much expect that input cannot be only the structure but sometimes function also. So, by extending this protein design, we can also go for designing a protein function. And if I assumed it is all about the function of a protein which matters to us.

So, we are dealing with a very interesting problem in protein modeling and engineering that we are designing the function of a protein. So, this procedure can be considered as a reverse of protein folding and protein structure prediction, where the latter is to deduce the 3D structure from any given sequence.

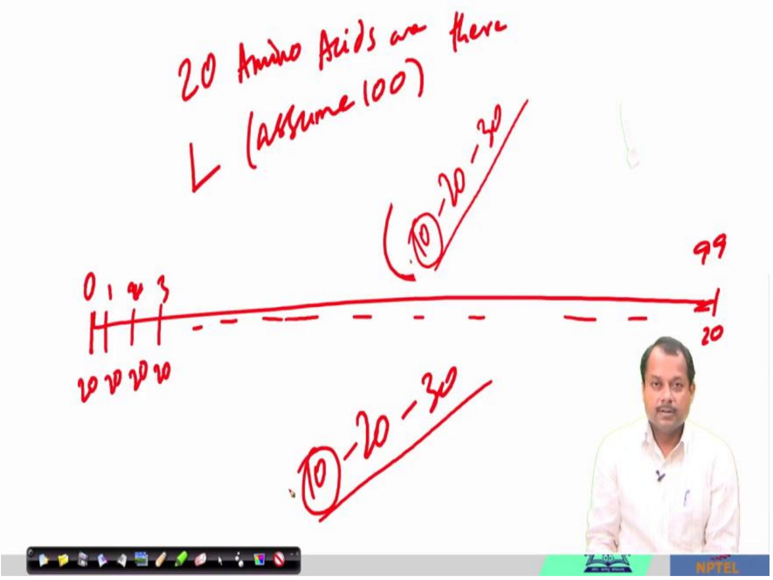
(Refer Slide Time: 19:42)



Protein Design Problem

```
SISARNQLKGKVVGLKKGVVTAEVVLEIAGGNKITSIISLDSVEELGVKEGAELTAVVKSTDVMILA
SKSADNVLKGRIISLEGGGNFTEVLIEIPGGEEITVVVPIKSAEELKLEPGKEILVWIKADEIHIFR
QKEAPYTLEGTVTKVEQLGDQAEIYIAIPGGENITLLVPVESVEELKIEPGKEIAVVLPPDKIHIFR
RKNVLKGRIGYIEKLGTSTEVVIVVGSGQQLTFIIPVKSAKDLEQEGLEPGAQISVYLPPSSIHVFR
VQASENIYEGTVIQVEKGGTKTEVYIEVGGGEQLAAVVPIKAAEALGIKPGKHIQVLFKAARIHVFR
QKVVRNILEGTIQYIEQGGNQTEVVIEIGGGQPIAAILPVQKAEELKLEPGANITVYIPPQKIHIFH
SKNVFKGKISKIQGLGVEAEIAVVVGAGKELAYLIPLKKAEALKLEQGEEVTIVFAADQVHVIHFFK
AKIADNVYKGRVTEIEVGGDEVEVYIVVSGGDQLKVLVPVESAEALKLDEGKNITVIIPPSEIHVFK
SQNVFKGRVGIVEKLGEETEIVLVVGKGQELTYIIPIKRAEALKLEEGKEGLWIVIILKPDRIHIYE
NKAADNVYQGEVDKIEVHLKIGVGVEITAIAPVEAAKQLKLYIWYAGFRPGAQVTVWFKADQIHIYR
AKSVDNVLQGTISYIEKGGDTAKIHVKIPGGSFVEKLVAVISVESAGALKLEPGVYVLTYFKASQIH
```

The computational complexity of the problem is O(20^N), where 20 is the essential amino acids and N is the length of the protein.



So, protein design problem let us looking into it again given input structure, I know the fast to blue color structure. This structure is known and that is attached with the structure which is given to me. So, this sequence is known to me I am proposing this could be possible sequences each of them can fold to that given structure.

Now, why this particular problem or computational approach is very important simple reason is that theoretically the, number of possibilities are very high why let us take an example simple one. So, we know that how many amino acids are there 20 amino acids are there. So, the degeneracy is 20.

Now, the length of the amino acid let us assume is L which you can also consider as assume 100 so, if 100 is the length of your amino acid then theoretically given a protein sequence 0 to 99. Now, one sequence is known to you, but how many possibilities are there in each position there are 20 possibilities if you go like this way up to here there are 20 possibilities that is astronomical out of which one is already known to you because the structure what is about the rest.

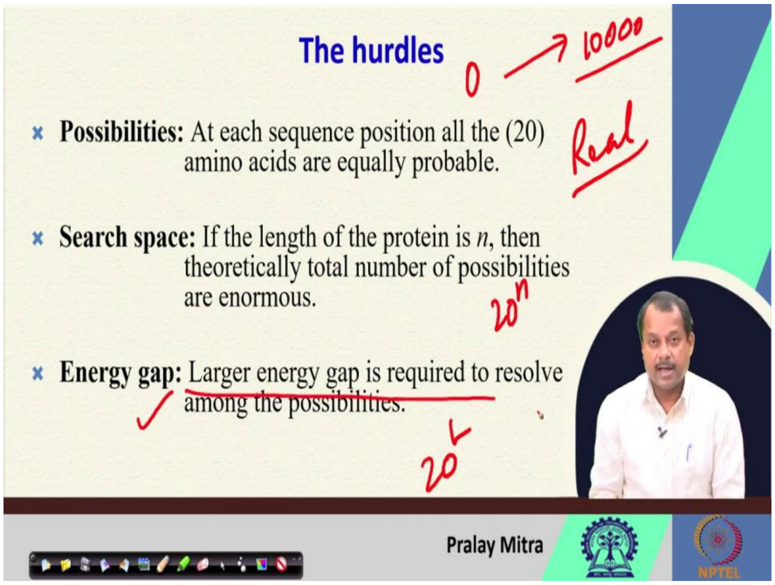
So, the idea that is why the application of the computational protein design so, out of these 20 to the power L possibilities which are theoretical in nature is it possible for you to give me say 20 to 30 max or even less than that is a 10 protein sequences on which I will do the biological experiment and confirmed that what is that design sequence?

Now, you think experimental technique needs a lot of resources that includes skilled manpower then laboratory space cost for the reagents also that time whereas, it is also not feasible that you understand that 20 to the power L possibilities will exist in nature only few definitely very few will be existing in nature.

But how do I know which one will exist then you need to have one very expert guy somebody like God who can guess by looking at the protein structure out of these 20 to the power L possibilities only these 10 or 20 possibilities are there and if it is a man human then the number of such humans are very less or may not be available.

So, we have to go for some alternative computational technique or we have to come up with an algorithm which can give you say 20, 30 or say 10 number of design sequences with some confidence so, that the experimentalists can pick say 10 or even lesser than that again based upon their knowledge and can do there experiment. So, we are trying to work in that area particularly. So, that is why this work is very much fascinating in nature.

(Refer Slide Time: 23:55)



The hurdles

0 → 10000

- **Possibilities:** At each sequence position all the (20) amino acids are equally probable.

Real

- **Search space:** If the length of the protein is *n*, then theoretically total number of possibilities are enormous.

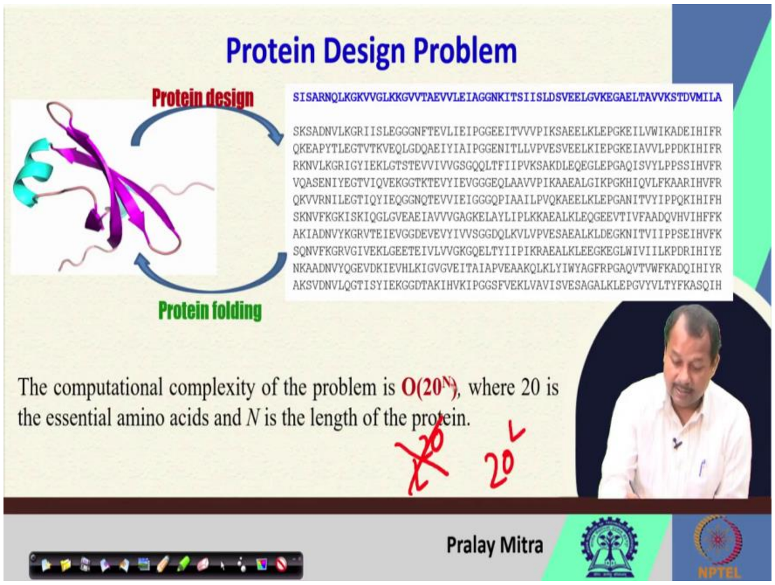
20^n

- **Energy gap:** Larger energy gap is required to resolve among the possibilities.

✓

20^L

Pralay Mitra



Protein Design Problem

```
SISARNQLKGKVVGLKKGVVTAEVVLEIAGGNKITSIISLDSVEELGVKEGAELTAVVKSTDVMILA

SKSADNVLKGRIISLEGGGNFTEVLIEIPGGEEITVVVPIKSAEELKLEPGKEILVWIKADEIHIFR
QKEAPYTLEGTVTKVEQLGDQAEIYIAIPGGENITLLVPVESVEELKIEPGKEIAVVLPPDKIHIFR
RKNVLKGRIGYIEKLGTSTEVVIVVGSGQQLTFIIPVKSAKDLEQEGLEPGAQISVYLPPSSIHVFR
VQASENIYEGTVIQVEKGGTKTEVYIEVGGGEQLAAVVPIKAAEALGIKPGKHIQVLFKAARIHVFR
QKVVRNILEGTIQYIEQGGNQTEVVIEIGGGQPIAAILPVQKAEELKLEPGANITVYIPPQKIHIFH
SKNVFKGKISKIQGLGVEAEIAVVVGAGKELAYLIPLKKAEALKLEQGEEVTIVFAADQVHVIHFFK
AKIADNVYKGRVTEIEVGGDEVEVYIVVSGGDQLKVLVPVESAEALKLDEGKNITVIIPPSEIHVFK
SQNVFKGRVGIVEKLGEETEIVLVVGKGQELTYIIPIKRAEALKLEEGKEGLWIVIILKPDRIHIYE
NKAADNVYQGEVDKIEVHLKIGVGVEITAIAPVEAAKQLKLYIWYAGFRPGAQVTVWFKADQIHIYR
AKSVDNVLQGTISYIEKGGDTAKIHVKIPGGSFVEKLVAVISVESAGALKLEPGVYVLTYFKASQIH
```

The computational complexity of the problem is **O(20^N)**, where 20 is the essential amino acids and *N* is the length of the protein.

$\cancel{L^{20}}$ 20^L

Pralay Mitra

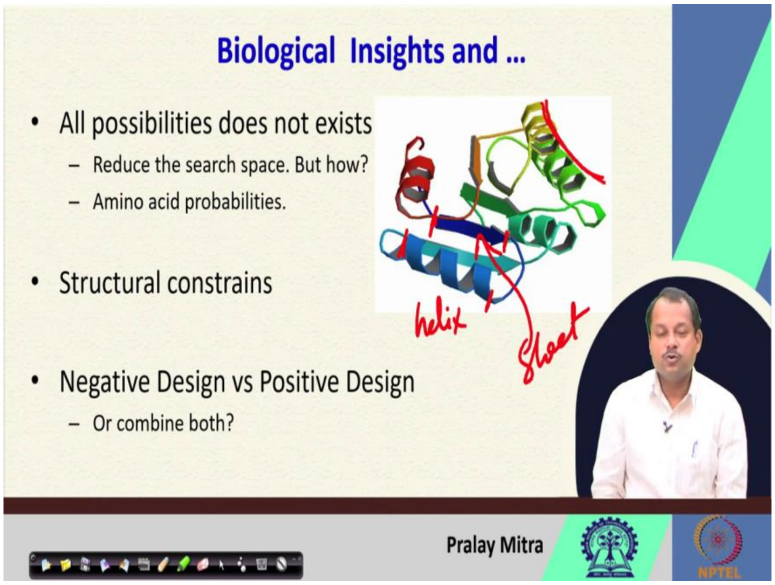
So, what are the hurdles? As I mentioned at each sequence position all that 20 amino acids are equally probable theoretically. So, the number of such possibilities as I mentioned on the last slide also is 20 to the power N where say 20 number of possibilities are there, sorry I make a mistake in that previous calculation.

So, it will be not L to the power 20 it will be it was wrong, it will be 20 to the power L in this case it will be 20 to the power N. Now, so 20 is the essential amino acid and N is the length of the protein. Now, the hurdle is at each sequence position all the 20 amino acids are equally probable, theoretically.

Next search space. If the length of the protein is n, then theoretically total number of, possibilities are enormous that we demonstrated larger energy gap is required to resolve among the possibilities. So, the last point specifically says that this last point that let us assume by some way even computationally also I have generated say 20 to the power L or L is the length of the sequence or in this case n 20 to the power n number of possibilities we have explored.

Now, if along with that one I need to have some score function which will score and rank my findings. Now, if I assume that my score function varies from say 0 through say 10,000 and even if this numbers are say real number in nature, then if you map then you will see that in for small change in the score function huge number of possibilities I mean in terms of the new design sequences will appear we will not get much larger energy gap to resolve among the possibilities. So, that is also going to be one hurdle for us.

(Refer Slide Time: 26:18)



Biological Insights and ...

- All possibilities does not exists
  - Reduce the search space. But how?
  - Amino acid probabilities.
- Structural constrains
- Negative Design vs Positive Design
  - Or combine both?

Then, what we need to exploit is the biology. So, if we provide some sort of biological insights then perhaps we will see all the 20 to the power n or 20 to the power L number of possibilities will not exist that is theoretical possibility. That is fine, but practically those do not exist that way if we (())(26:44) down and use some sort of biology, then it will be very much useful for fur thus coding and ranking.

That is why we need to know a little bit of biology or the domain knowledge without which we cannot design the algorithm. So, first of all, all possibilities does not exist reduce the search space, but how amino acid probability check with the amino acids probability. So, although I am telling that all the amino acids are theoretically equally probable at each position, but is it true, it may be the case that in order to get into that particular structure in one position, only few number of amino acids are probable, but not all that way I can reduce some of the possibilities.

Next, put some structural constraints, since my intention is to design one protein sequence which will fold to this given protein structure and the structure is known to me then why not fix that this particular stretch after the protein design also is going to be the helix this particular stretch is going to be the sheet.

So, from here I can mark say, this particular stretch is going to be the helix this particular stretch is going to be the sheet. So, why not mention that one in my score function or in my searching technique, so, that whenever I will screen for this then I will focus only into this. So, this is regarding the secondary structure, but the same thing we can do for say solvent accessibility which residues are on the surface.

So, here predominantly they are going to be the hydrophilic residues. Now, if it is on the surface and we identified that one because I know the structure then all those positions will be replaced by only the hydrophilic residues not by any hydrophobic residues it will be just opposite when I am considering the core of the protein structure where it is occluded from the solvent.

Then if I change that one by some hydrophobic residues then that is going to be more stable compared to changing to hydrophilic residues that way, the number of amino acids are not 20. So, it is only the hydrophilic residues which are very less compared to 20, so, why not use that one that is one. Third is that negative design or positive design.

So, positive design means that when we are we know that there is a energy function we have to go for stable design and we will keep on reducing our score function or we will move on to the stable or better solutions and pause in negative design indicates that we will generate and we will discard the negative cases means that these are not correct.

So, these are not correct so remove. So, this with us to go for positive design or for the negative design or combine them so, that after we generate something then we will decide whether we wish to go away from there or we wish to go towards there together we will combine that information in order to have a better scoring function. So, combining this we wish to come up with some algorithm which will be able to design a new protein sequence or a number of new protein sequences that will fold to the given protein structure that we will continue to the next lecture. Thank you