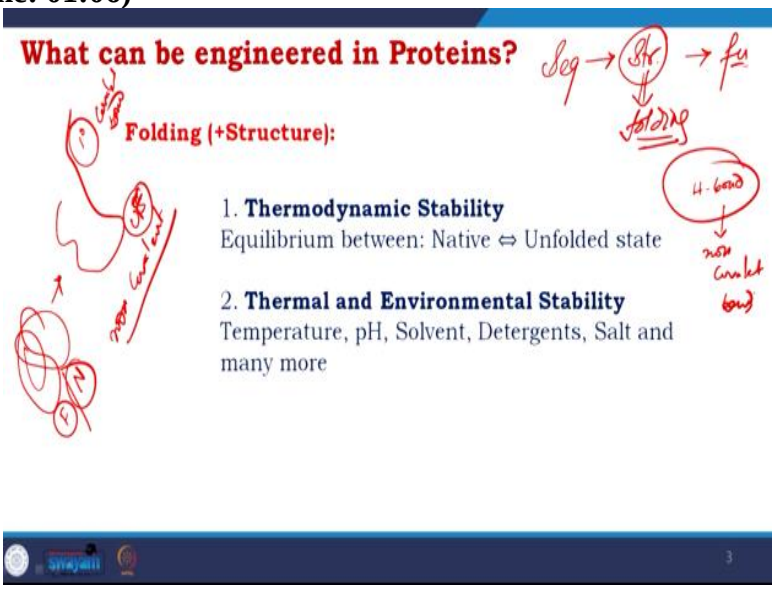
Structural Biology
Prof. Saugata Hazra
Department of Biotechnology
Indian Institute of Technology - Roorkee

Lecture - 52
How to Make Logical Protein Engineering: Process of Rational Design

Hi, everyone, welcome again to the course of structural biology; we are kind of at the end part of the course we are going through the 11th module. Furthermore, as you know, we started the protein engineering module; today, I will continue from that. So, today, the foremost thing we will discuss is protein engineering strategies; this is the core part, the intellectual part.

(Refer Slide Time: 01:06)



So, to start with, I have kept a few questions and answered them to understand how the process is going on; what can be engineered in proteins? So, we have the protein; what is our target? So, one thing is folding, and if you have gone through the course properly, you would understand the importance of folding the journey; what we talked about is the journey of sequence to function in between is structure; structure means folding. Moreover, that is why I wrote folding plus structure.

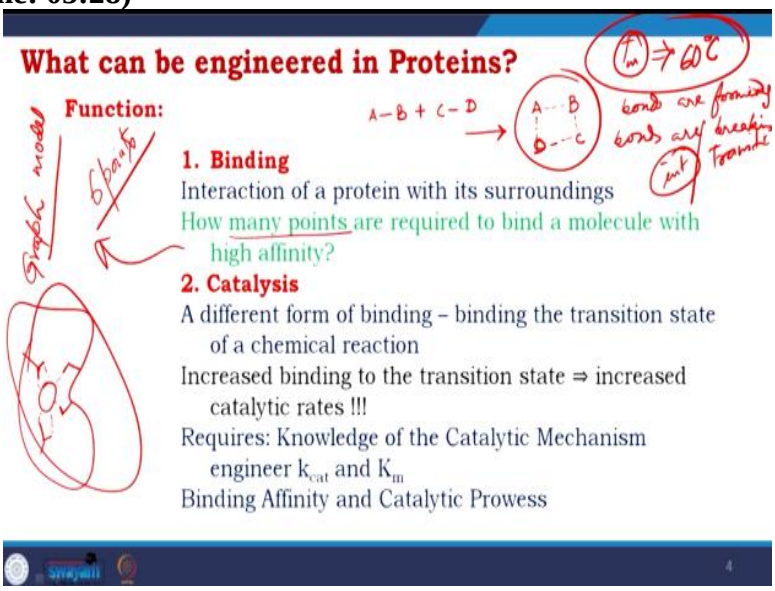
So, as I told you from the beginning of the course, getting the structure knowing the fold would help us now you will understand. So, one we will look for is thermodynamic stability. Thermodynamic stability deals with the native and unfolded state; if you remember, in the last module of MD simulation, we have discussed the opening of a protein. Opening a protein means opening a folded state to the unfolded state.

So, the folded state, we could also call it native state, show one of our close looks could be how the protein alters from the native to the unfolded state and can we keep it in the native state can we enhance the thermodynamic stability. The second one is in a broader sense, but thermal and environmental stability; under these, there would be temperature, pH, solvent, detergent, salt and many more.

So, you have a protein; if you all know that the protein has a primary sequence, the protein sequence now, then the sequence is made by a covalent bond. Now, this protein is doing what we call fold, where you get the structure by a non-covalent bond. Again throughout the course, we have discussed the non-covalent bonds; they are not as stable as the current bond, and there are many like hydrogen bond salt bridge we will name them.

However, you could consider as an example, which is hydrogen bond. A hydrogen bond is a non-covalent bond present in protein in many. So, their state, whether they are forming or not forming, have a huge role in swing the stability of the protein. So, one of the biggest targets of protein engineering is for stability because if we get to make the higher stability version, the thermostable version, I would say of a protein, it could be used more industrially.

(Refer Slide Time: 05:28)



Because normal protein has T_m, melting temperature around 60 degrees centigrade, whereas in the; industry, it goes above 70, 75 degrees centigrade. So, if you could make a protein stable above 75 degrees centigrade, this protein or this enzyme would be industrially suitable.

It would be part of the fermenter bioprocess and do not denature. The next thing is the function, and one of the aspects of function could be described by binding.

Binding is the interaction of the protein with its surrounding, the interaction of the protein with other macromolecules, the interaction of the protein with substrates inhibitors like and small molecule anything. So, how many points are required to bind a molecule with high affinity? A question I posed is generally used by people who develop graph models and further use those data towards the development of machine learning; this prediction is what I mean suppose, you have a protein like this.

Moreover, the protein has one amino acid, two amino acids, three amino acids, four amino acids; then there is this substrate, and the substrate makes non-covalent interaction with the protein. Now, you see that these non-covalent interactions would be measured as points, and the number measurement is straightforward. That is one of the reasons this is chosen definitely; it would not say everything if I said there are 6 points. I do not include chemistry or biochemistry here because these interactions are not the same.

One might be a hydrogen bond other might be a salt bridge; among the hydrogen bond, there are different strengths, but the point is the easiest and essential representation for the development of the model. Then you could further introduce complexity to the basic model that is how computational model development works. The second thing is that catalysis binding is happening now; how the enzyme converts the substrate to the product is called catalysis -a different form of binding that transitions trade of a chemical reaction.

So, we will discuss this point in detail in structure-based drug designing our last module, but just to talk about, when there is a substrate binding to the enzyme, the enzyme starts a reaction. So, what happened? You have A bond B + C bond D. When the reaction starts, the bond between A and B and the bond between C and D breaks the bond between B and C and A and D start forming; this is the state where bonds are formed.

Bonds are like forming and breaking; this state is called intermediate or transition state. The transition state is crucial because binding a substrate with an enzyme in its transition state shows the intrinsic property of the enzyme; why the intrinsic property is so important? Because if you see nature does the evolution, we will talk about that nature makes changes in

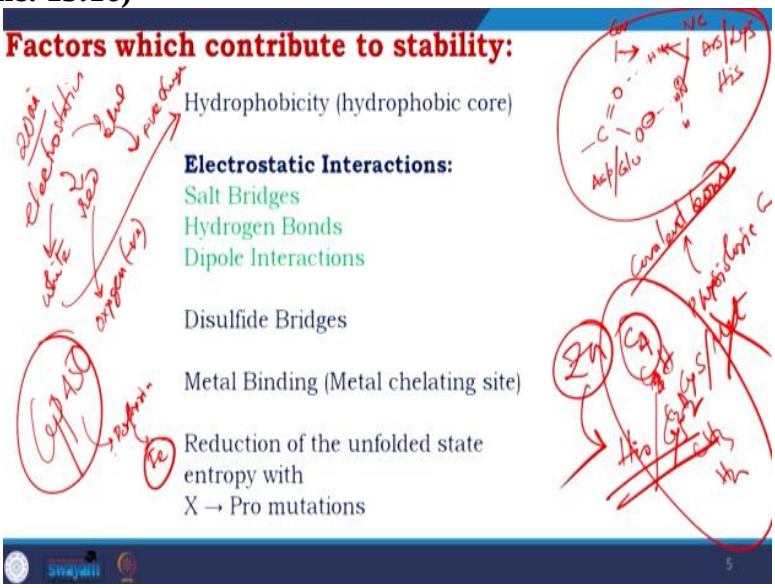
protein, because of many reasons, one of the significant reasons is where I research so, I talk about that is called antimicrobial resistance.

So, an enzyme changed itself so that it could not bind to the inhibitor, but it could bind its substrate, but the enzyme could not change its intrinsic property. I will come back to that on drug designing because this is a very critical concept in drug designing. So, catalysis engineering is possible targeting the enzyme altering its binding to transient state means altering its substrate specificity, altering its choice of substrate, and many other things.

So, as you all know, the enzymes are, in many cases specific, the specificity could be altered by studying the binding of the chemical reaction how the enzyme is binding to the transient state. However, you could ask me if you understand that transition state; as you say, bonds are forming and breaking. So, it is a hypothetical state there is no practical existence.

So, how do you study it? As I told you, I will go into detail in the drug designing module and show you how we study enzymes binding in its transition state or intermediate of the reaction. So, increased binding to the transition state, increased catalytic rate, altering catalytic rate, many things are possible. It needs us to engineer the protein towards catalyst by knowing the catalytic mechanism; we could engineer k_{cat} K_m, k_{cat} is the catalytic process catalytic power K_m is correlated to the affinity.

(Refer Slide Time: 13:16)



So, now, we are talking about factors that contribute to stability. As I talked about hydrophobicity, we have discussed 20 amino acids in one of the previous modules. We talked

about the 9th model of visualization. We talked about Pymol if you get a PDB downloaded from the RCSB protein databank. Moreover, load it in the Pymol and go and find the electrostatics; you will get the white colour that refers to the hydrophobic patch or the presence of hydrophobic amino acids.

Then you get red which will show you hydrophilic, oxygen and negative charge, and then you get blue, nitrogen or positive charge. So, in that way, when you look at the electrostatic presentation, you will get the hydrophobic amino acids, the hydrophobic amino acids, you know, valine, isoleucine, leucine, alanine, a lot of them the aromatic ones, the phenylalanine so, all of them are contributing to hydrophobicity. Electrostatic interactions are significant for different reasons.

Especially when you consider protein engineering, which is our current topic, we are more interested in electrostatic interaction because you could measure the electrostatic factors; one of them is the salt bridge; what is a salt bridge? A salt bridge is in its stability or power; between the covalent and non-covalent bonds, it is much stronger than a non-covalent bond and weaker than a covalent bond.

A salt bridge is formed between aspartate glutamate with arginine, lysine, and histidine; by introducing a salt bridge, you could enhance a protein's thermostability. As I told you, hydrogen bonds are not very strong; they are weak bonds, but these are many so, breaking and making of a set of hydrogen bonds help the protein get more stable, get protein changing its conformation and many things else. Then dipole interaction you could also you know look at the dipole interaction, and you engineer them.

A disulfide bridge is very powerful because it is the only covalent bond that could form in physiologic conditions. There are two amino acids, cysteine and methionine, which have sulphur methionine does not form a disulphide bond; it is all the cysteine. So, when two cysteines are in a proper position, they could form, but when you are engineering a thermostable protein, you could introduce a disulfide bond and design a disulphide bridge.

Metal-binding makes the domain of the protein more stable. So, what type of metal is coming in dye form like dye gene can all. So, these are all considerations if you consider zinc they generally bound to like histidine and cysteine it is mostly is called zinc finger where the

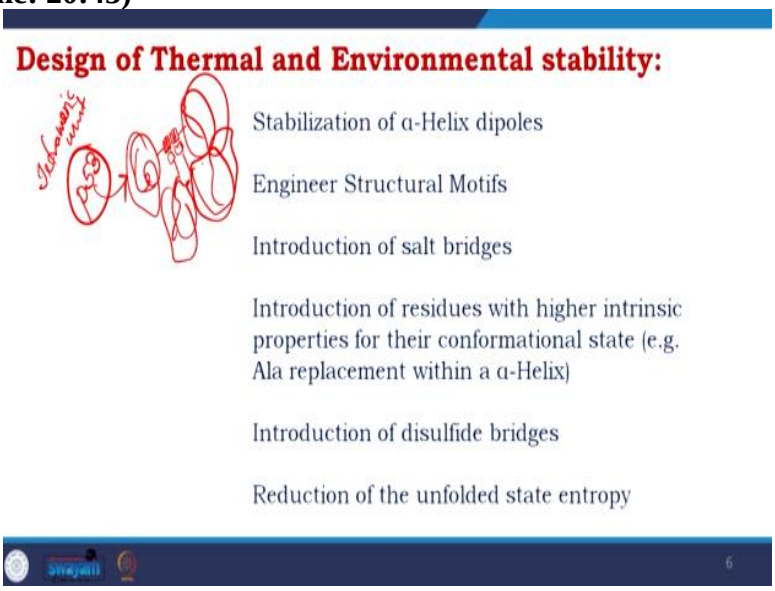
binding residues are C 4 all cysteine C 3 H, C 2 H 2, CH 3 or H 4. So, these are the general scenario. In the same way, there is a porphyrin ring that is binding to iron.

Here, there is an exciting thing I should tell you, which would be a perfect target of protein engineering if you see an enzyme called cytochrome p450. The enzyme had a porphyrin ring and an iron-bound to it now if you take out the iron if you alter the porphyrin rings so that it could bind to other metals, even other metals which are biologically not seen you could have carried out catalysis, which is not feasible in biological condition.

However, if the enzyme could agree to perform this, you will get the conversion very specific and probably in 1 step. So what I am trying to say like cytochromes, which are generally oxygenous types. If you could properly engineer it, you could use these biocatalysis in abiocatalysis to make 1 step, 2 step conversion very common example benzene to phenyl. You could perform those in significantly fewer steps and at a low cost.

So, these are the areas where protein engineering could take a critical role; also, reduce the unfolded state entropy you could work on how? Proline has a considerable rigidity that sometimes works on those purposes.

(Refer Slide Time: 20:43)

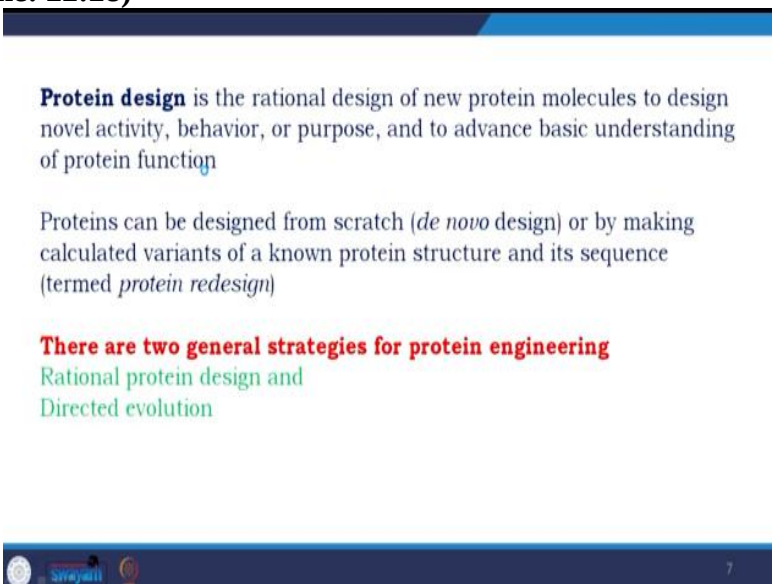


Design of thermal and environmental stability, stabilization of alpha-helix dipoles by stabilizing those dipoles you can increase the stability of the protein in significant amount, you could engineer structural motifs, you could do very interesting creative engineering, you could make protein oligomerized, a common strategy people use the tetrameric or tetramer

forming unit of enzyme p 53 and attach that with other protein so, they tag it. Hopefully, it would bring the same protein to form a tetramer, which generally forms a monomer. So, this type of thing is possible; you could also make a protein binding to a DNA that is not binding to DNA by introducing a structural motif that binds to DNA. So, anything you could do, you could introduce salt bridges.

As I told you, you could introduce residues with higher intrinsic property for their conformational state like you could replace alanine within an alpha helix to see the change of secondary structure change up the tendency of forming secondary structure. You could introduce a disulfide bridge to reduce the unfolded state entropy.

(Refer Slide Time: 22:28)

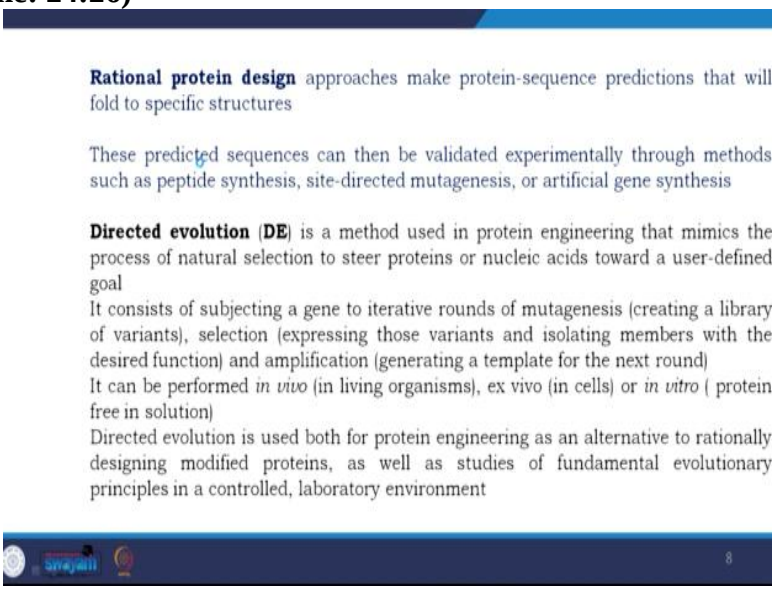


So, the protein designing process is the rational design of new protein molecules to design novel activity behaviour or purpose and advance a basic understanding of protein function. Proteins can be designed from scratch, called de novo design, or by making calculated variants of a known protein structure. Suppose, already, there is a protein you want to make the proteins stable that is called protein redesigned.

So, you could do a de novo design of a protein with a novel fold with a combination fold which is not present in nature, they are called de novo designing, but if you take one protein and alter it towards different activities called protein redesign, generally, the major strategies of protein engineering are divided into two major fields. One is rational protein designing. Rational protein designing is a process that is the actual process that people have adopted.

The other one is directed evolution is something that continues before humans; when I say we are born, I mean the existence of the human community. So, it is what nature used to do to optimize and make sure that the new organism or something coming out has a total system balanced and working.

(Refer Slide Time: 24:26)



So, rational protein designing approaches make protein sequence predictions that fold to specific structures. These predicted sequences can then be validated experimentally to methods such as peptide synthesis, site-directed mutagenesis or artificial gene synthesis. Directed evolution (DE) is a method used in protein engineering that mimics the process of natural selection to steer proteins or nucleic acid toward a user-defined goal. It mimics the process by which nature creates and maintains all the living organisms; directed evolution involves subjecting a gene to an iterative round of mutagenesis. So, you take a gene, perform random mutagenesis, create a library of variants, and then make the selection of what you want. So, selecting those variants, isolating members with the desired function, and then amplifying generate a template for the next round.

In directed evolution, one of the significant challenges is to get a proper selection method of the variants; it can be performed in vivo in a living organism like a mouse in the laboratory, ex vivo that means, in cells and in vitro that means, you had a gene, you look at the gene expression, or you take the gene product, the protein and make an assay out of this directed evolution is used both for protein engineering; As an alternative to rationally designing modified proteins as well as studies of fundamental evolutionary principles in a controlled laboratory environment. So, these two are what is, in practice, rational designing and directed

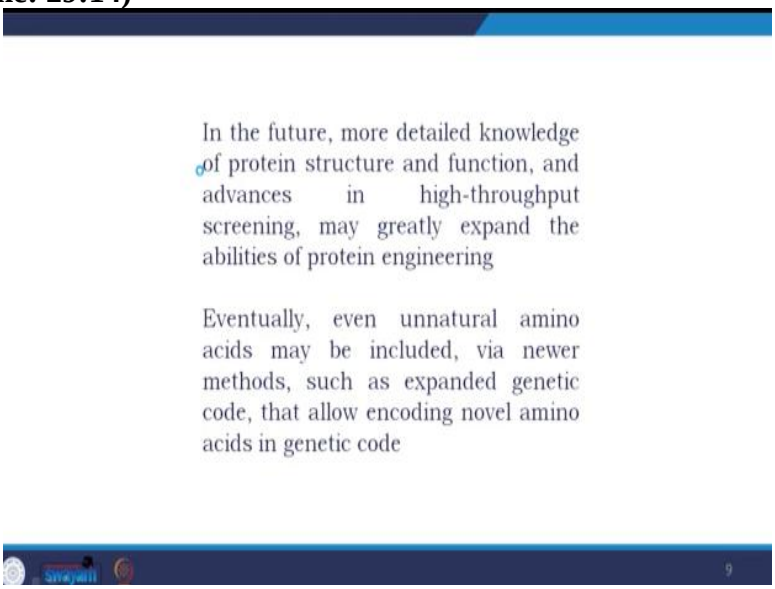
evolution. However, as I continue to talk about the innovation and optimization of next-generation sequencing technology, biology is looked at in a very different way.

Moreover, as I say, continuously to the younger generations, if you see, it is not very old, though it shows that NGS, the first human genome project, was performed in 2003; but, if you look out through the journey of every aspect related to NGS and structural biology, it is around 2015, when we start realizing the importance of NGS when the common scientists start using NGS, because, at the initial stage, NGS was very costly.

A good example was when I joined IIT Roorkee; the services we used to get in India used to charge more than 50,000 to 1 lakh for bacterial genome sequencing, and currently, it is reduced to below 15,000. So, these fivefold decrease like now, you spend 15,000, and you get the complete information about an organism that is where the revolution is happening, that is where the biology is changing biology is not looked at a gene level, you could have seen all other effects through genomics and transcriptomics.

You see the DNA RNA level through metabolomics. The small molecules generated to give you the entire picture in front of you would make protein engineering enzyme engineering much easier I would say much more creativity and more innovations are coming.

(Refer Slide Time: 29:14)

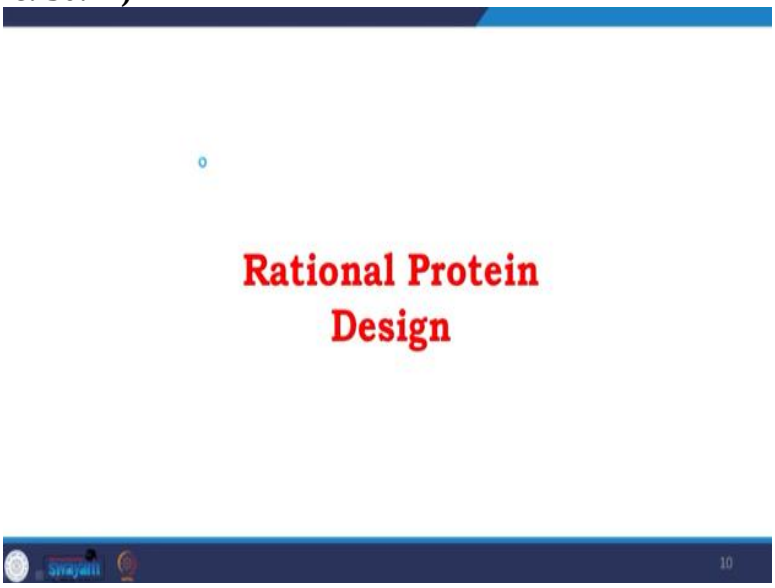


So we talk about the advances of protein structure and function in advance in high throughput screening, it may greatly expand the abilities of protein engineering, as I explained, there eventually even unnatural amino acids. We are only talking about 20 amino acids, but why

not modify amino acids into the protein designing? If it could happen via newer methods, such as expanded genetic code, which is already working, it would allow the encoding of novel amino acids in the genetic code.

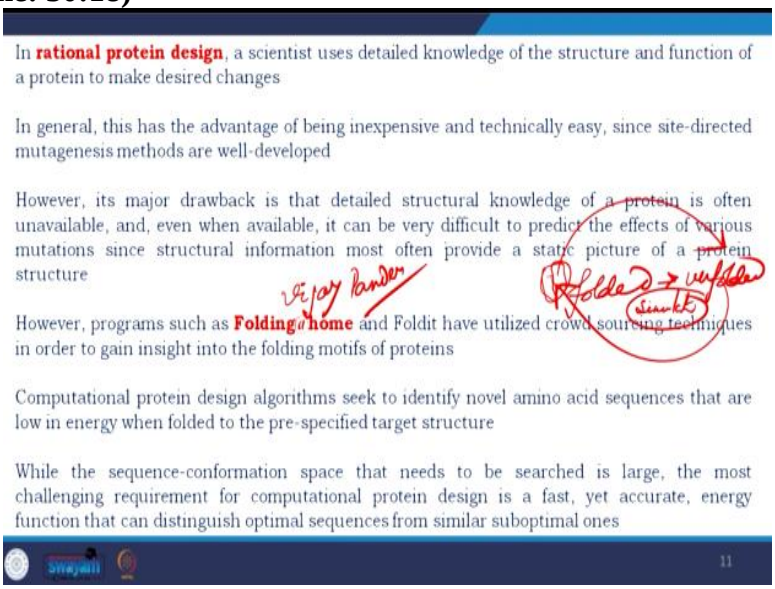
Moreover, what more you could do only with four bases, four nucleosides you have 61 codons each of one incorporation will make so many, and we could expand the chemistry.

(Refer Slide Time: 30:14)



So with that, I am coming to rational protein designing.

(Refer Slide Time: 30:18)



In rational protein designing, a scientist uses detailed knowledge of the structure and function of a protein to make desired changes towards getting an updated protein. I am keeping it common because updated, maybe anything. In general, this has the advantage of being

inexpensive and technically easy since site-directed mutagenesis methods are well developed. So, what you are doing, you look at the structure, design it and then make a change.

As I have already explained, whatever you do at the protein level, you cannot change the protein level; you have to go and make changes at the DNA level. However, its major drawback is that detailed structural knowledge of a protein is often unavailable as the initial description of this course, that challenges we do not have enough structure now, you know, we have 175000 plus structures, which is coming from around 35 to 40,000 proteins, whereas, millions of proteins are now available day by day newer proteins are coming.

So, that is one major challenge; if you do not have good structural information, then it is difficult. So, the major drawback is that detailed structural knowledge of a protein is often unavailable. Moreover, even when available, it can be tough to predict the effects of various mutations since structural information most often provides a static picture of a protein structure.

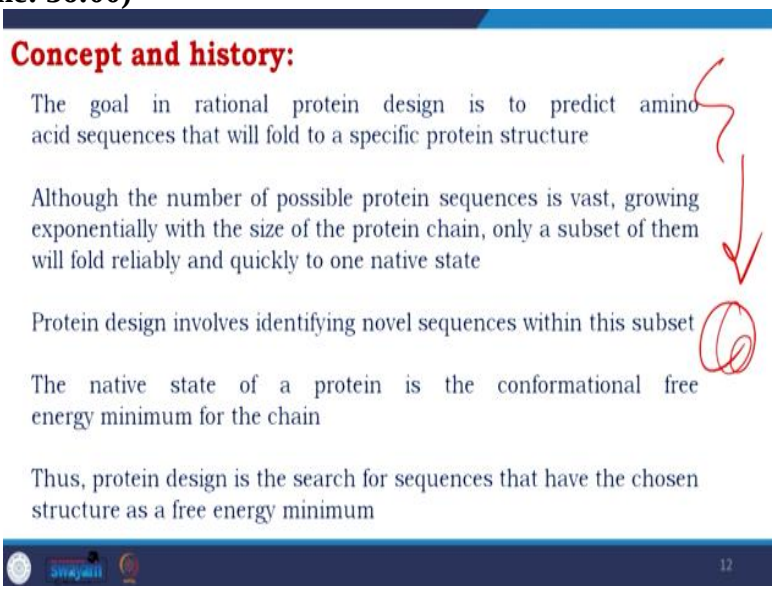
So, that is another thing we have talked about; most of the data is based on high-resolution X-Ray crystallographic data, representing protein in a static mode; but now, you want to get the dynamic data for making new mutants. However, programs such as Folding at home and Foldit have utilized crowdsourcing techniques to gain insight into the folding motifs of proteins.

So, if you had just gone through the module of MD simulation, you will see that making the dynamics of a protein takes huge time and as I told in the module, we do the folded protein to unfolded this is studied through simulation why? Because you have the folding information, you utilize it to unfold. There is a certain thing when something is zipped and opened up; there is one way around, but it could take much more computational power if you want to do this, and unfortunately, even our supercomputing is not good enough for that; but folding at home by Vijay Pandey is an amazing concept where people are giving access to their personal computers when they are not using it. Also, folding at home is taking the contribution of gamers; you know gamers always have high-performance computing.
So they are developing games where you make the team, and you start playing that so that you would make many permutations and combinations to achieve an unfolded protein to a folded one, and this initiative is coming up with solutions of protein structures that were not

shown before. Computational protein designing algorithms seek to identify novel amino acid sequences that are low in energy when folded to the pre-specified target structure.

While the sequence confirmation space needs to be searched in large, the most challenging requirement for computational protein designing is a fast yet accurate energy function that can distinguish optimal sequences from similar suboptimal ones. So again, this is the area where much research is going on. Artificial intelligence and machine learning are working a lot, but we are still waiting to talk between the theoretical studies and the experimental findings.

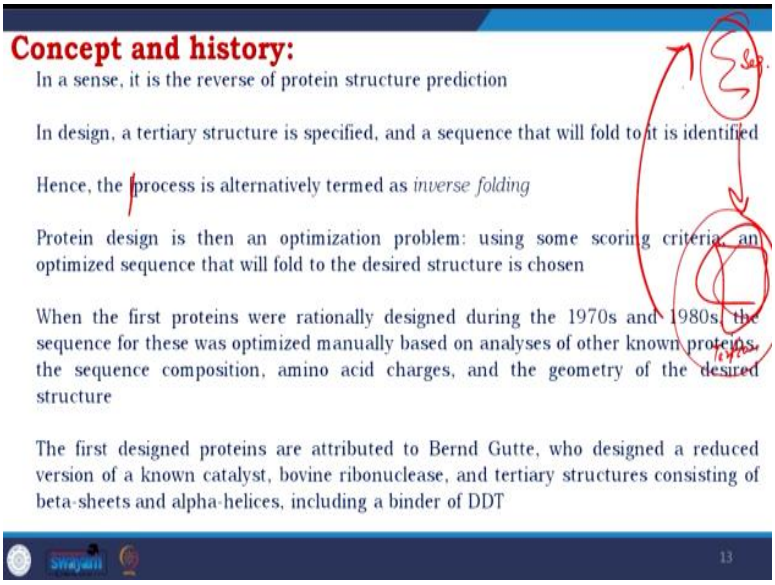
(Refer Slide Time: 36:00)



According to the concept and history, rational protein designing aims to predict amino acid sequences that fold to a specific protein structure. Although the number of possible protein sequences is first, as I told you, growing the size of the protein chain exponentially, only a subset of them will fold reliably and quickly to one native state. Protein designing involves identifying novel sequences within that subset; the native state is the conformational free energy minimum for the chain.

Now you understand what I am talking about; I have explained these when we have an unfolded sequence. We look at the folding by pushing it towards the lowest energy minima. Thus, protein design searches sequences with the chosen structure as a free energy minimum.

(Refer Slide Time: 37:06)

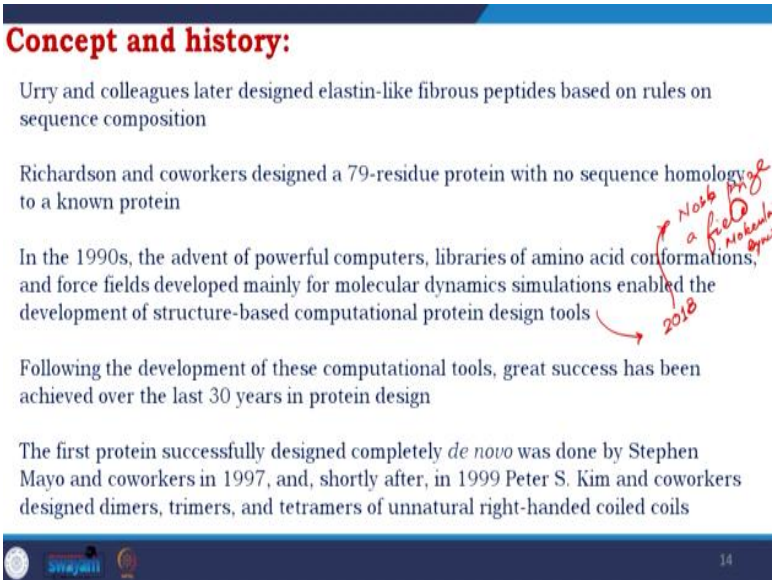


In a sense, it is a reverse of protein structure prediction. Because when we have protein structure prediction, we have the sequence, and we get the structure. Now we look at how we get a native fold from which sequences that are why it is a kind of a reverse process in designing a tertiary structure is specified first, and the sequence that will fold to it is identified.

So we have that tertiary structure, and we find a primary sequence which will provide or which get from these; hence, the process is alternatively termed as reverse or inverse folding, so this is folding, this is inverse folding. Protein design is then an optimization problem using some scoring criteria, and an optimized sequence that will fold to that desired structure is chosen. So here, we have the sequence, and we want to get the folded information here.

We take a folded structure and score to the sequence, given when the first proteins were rationally designed during the 1970 and 80s. The sequence for these was optimized manually based on analysis of other known proteins, the sequence compositions, amino acid charges, and the desired structure's geometry. The first design proteins are attributed to Bernd Gutte, who designed a reduced version of a known catalyst, bovine ribonuclease and tertiary structures consisting of beta seats and alpha helices, including a binder of DDT.

(Refer Slide Time: 39:08)

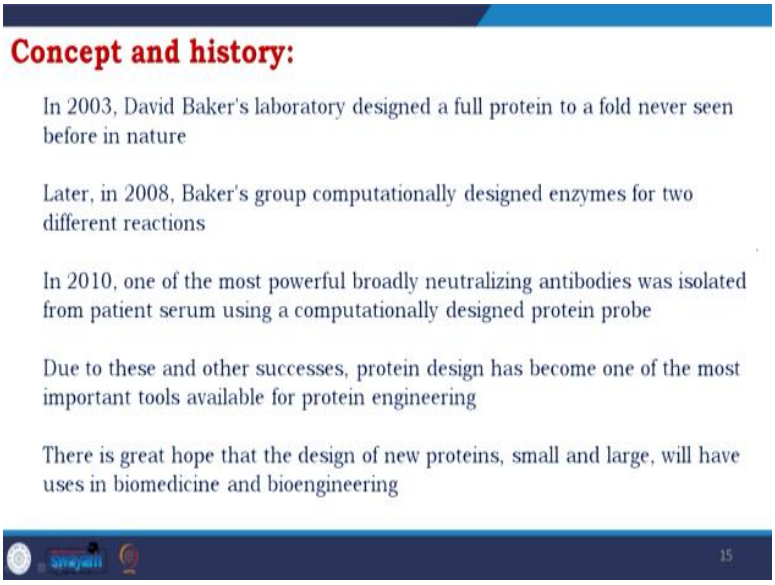


Urry and colleagues later designed elastin-like fibrous peptides based on rules on sequence composition. Richardson, whom we talked about spatially in the visualization time, and coworkers designed a 79 residue protein with no sequence homology to a known protein. In the 1990s, the advent of powerful computers, libraries of amino acid confirmations, and force fields developed mainly for molecular dynamics simulation enabled the development of structure-based computational protein designing tools.

If you remember, I told because of these, looking at this contribution in 2013 Nobel Prize was awarded to a field called molecular dynamics; it was the first time it was given to 3 people, but it was mainly because they have walked through the development of MD simulation and how molecular dynamic simulation contributed to the understanding of protein structure-function and helping in designing of new protein.

Following the development of these computational tools, great success has been achieved over the last 30 years in protein designing. The first protein successfully designed completely de novo was done by Stephen Mayo and coworkers in 1997. Moreover, in 1999, Peter S. Kim and coworkers designed dimer, trimer and tetramer of unnatural right-handed coiled coils.

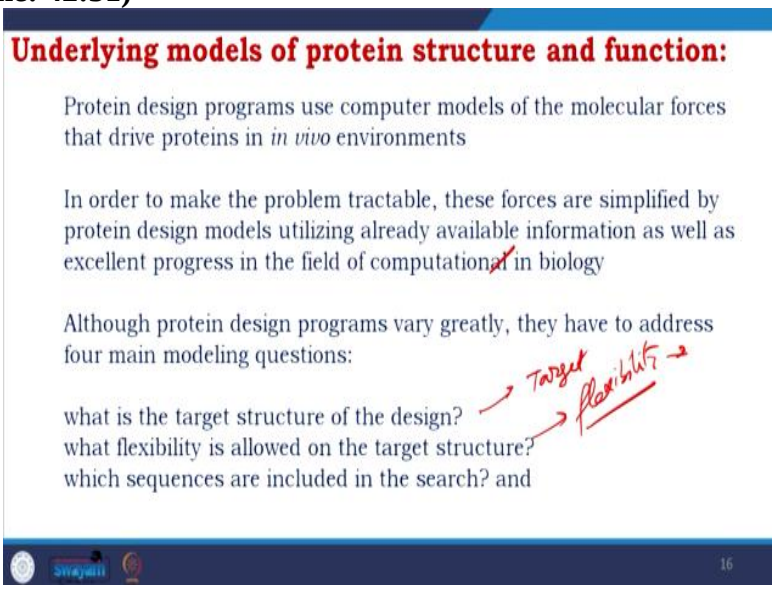
(Refer Slide Time: 41:26)



In 2003 David Baker's laboratory designed a full protein to a fold never seen before in nature. So, this is a new fold coming in nature; later in 2008, Baker's group computationally designed enzymes for two different reactions. In 2010, one of the most powerful broadly neutralizing antibodies was isolated from patient serum using a computationally designed protein probe. Due to these and other successes, protein design has become one of the most important tools available for protein engineering.

There is great hope that the design of new proteins, small and large, will have uses in biomedicine, bioengineering and it is just too broad filled, but it could be in any field not only in biology, it could also be you know, in the field of development of electronics, sensors, chemical reactions and whatnot.

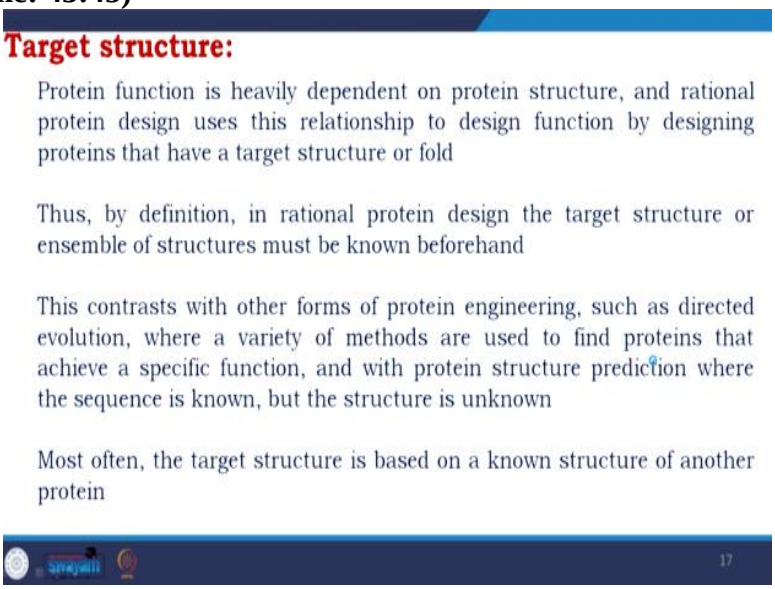
(Refer Slide Time: 42:31)



So, underlying models of protein structure and function, protein design programs use computer models of the molecular forces that drive proteins in vivo environments. In order to make the problem tractable, these forces are simplified by protein design models utilizing already available information and excellent progress in the field of computation in biology.

Although protein design programs vary greatly, they have to address four main modelling questions, what is the target structure of the designing? So, the primary target is the structure; what flexibility is allowed on the target structure flexibility imposed the dynamicity of the model what or which sequences are included in the search? So, what are the correlations? What is the sequence space? Which force field will be used to score sequences and structures?

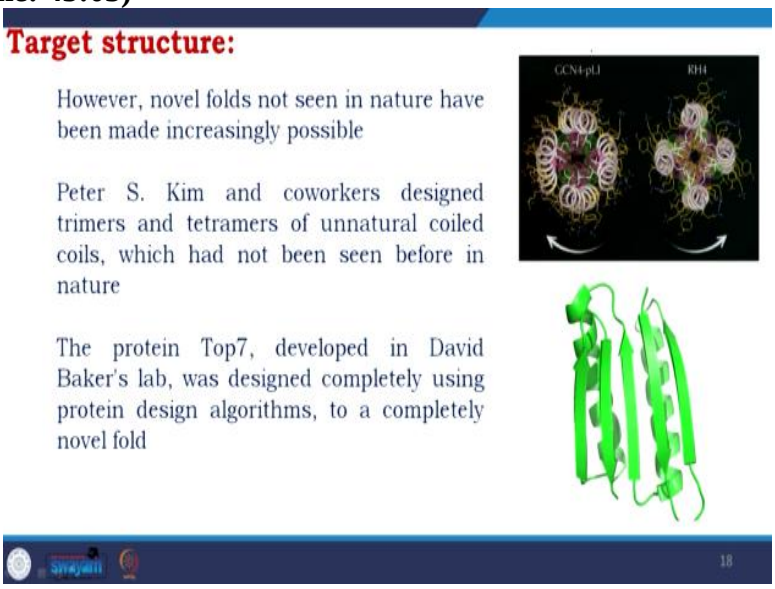
(Refer Slide Time: 43:45)



So, coming to target structure, protein function is heavily dependent on protein structure, as we have talked thoroughly in the course, and rational protein design uses this relationship to design function by designing proteins that have a target structure or fold. Thus, by definition, in rational protein design, the target structure or ensemble of structures must be known beforehand; This contrasts with other forms of protein engineering, such as another one which we told about directed evolution, where a variety of methods are used to find proteins that achieve a specific function and with protein structure prediction where the sequence is known. However, the structure is unknown here; it had to be that at least some structural information are available. The target structure is often based on the known structure of another protein.

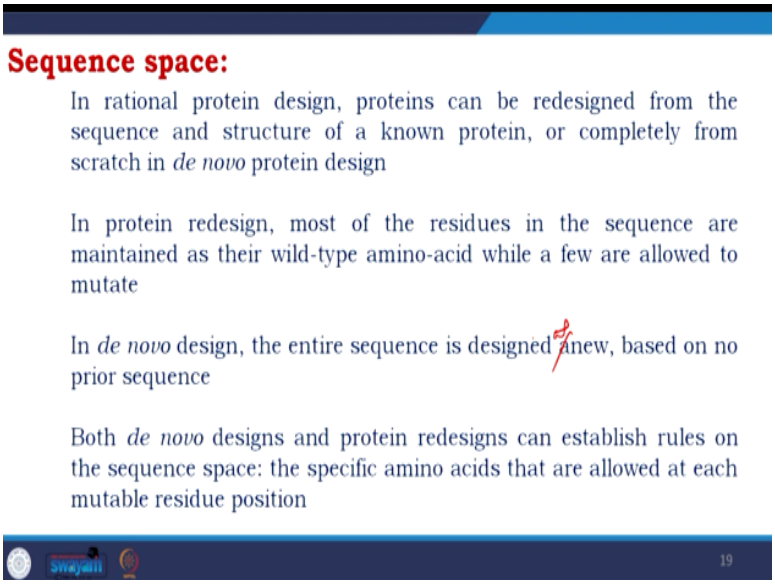
So, you might not have the availability of the exact direct structure of that protein, but you need structural information from the related proteins to use that to make a predictive model 3D model of the target.

(Refer Slide Time: 45:03)



However, novel folds not seen in nature have been made increasingly possible. So, more and more structures are coming less, and fewer new folds are available as it is estimated now that nature has a limited number of folds. Peter S. Kim and coworkers designed trimer and tetramer of unnatural coiled coils, which had not been seen before in nature. So, if you see, this is GCN4, the existing protein, and this is the tetramer RH4 designed by Peter S. Kim, you could see that the fold is very different, but these four alpha-helix you see are the exact alpha helix coming from this protein. The protein Top7 developed in David Baker's lab was designed completely using a protein design algorithm to a completely novel fold. So, this protein he was looking at is Top 7, and it is providing a fold, which is completely novel, not achieved or not made by nature.

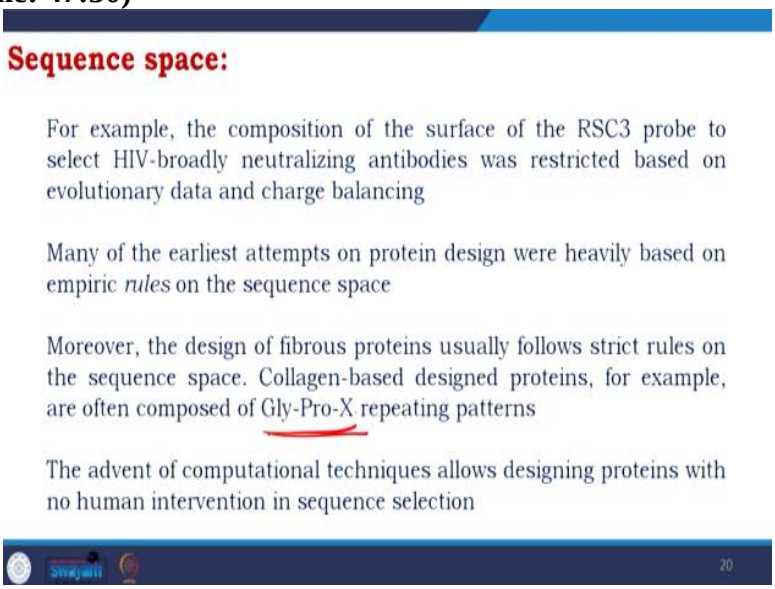
(Refer Slide Time: 46:33)



Coming to sequence space, as I was talking about in rational protein design, proteins can be redesigned from the sequence and structure of the known protein or entirely from scratch in de novo protein design. In protein redesign, most of the residues in the sequence are maintained as the wild type amino acid, while a few are allowed to mutate and see the effect of those changes.

In de novo design, the entire sequence is designed new or based on no prior sequence. So one is de novo, and another is redesigned. Both de novo design and protein redesign can establish rules on the sequence space and the specific amino acids allowed at each mutable residue position.

(Refer Slide Time: 47:30)

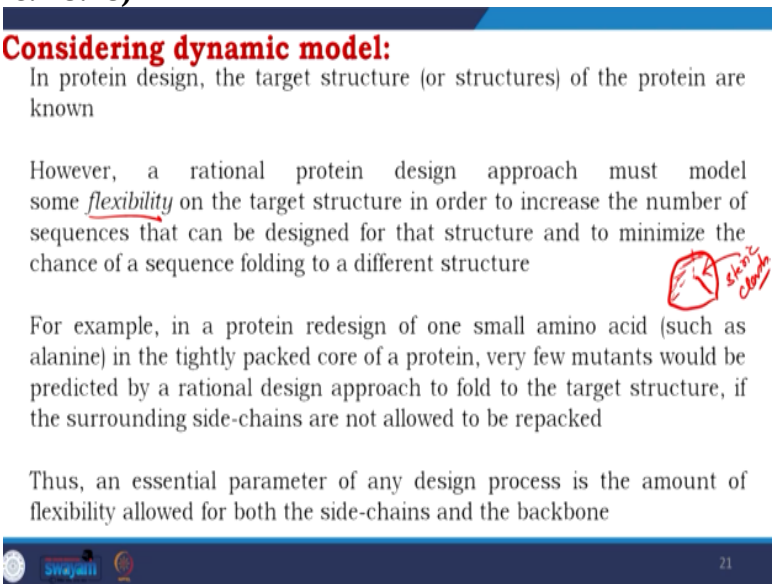


For example, the composition of the surface of the RSC3 probe to select HIV broadly neutralizing antibodies were restricted based on evolutionary data and charge balancing.

Many of the earliest attempts on protein design are heavily based on empiric rules on the sequence space. Moreover, fibrous protein design usually follows strict rules on the sequence space collagen-based designed protein.

For example, collagen-based designed proteins are often composed of having the repeating pattern of Gly Pro X . We see in that collagen fibre the advent of computational techniques allow designing proteins with no human intervention in sequence selection, the whole sequence selection thing i

(Refer Slide Time: 48:28)



In protein designing, the targets structure of the protein is known. So, I am saying that the static information of the 3D structure is known, but now we are talking about a dynamic model. However, a rational protein design approach must model some flexibility on the target structure to increase the number of sequences designed for that structure and minimize the chance of a sequence folding to a different structure.

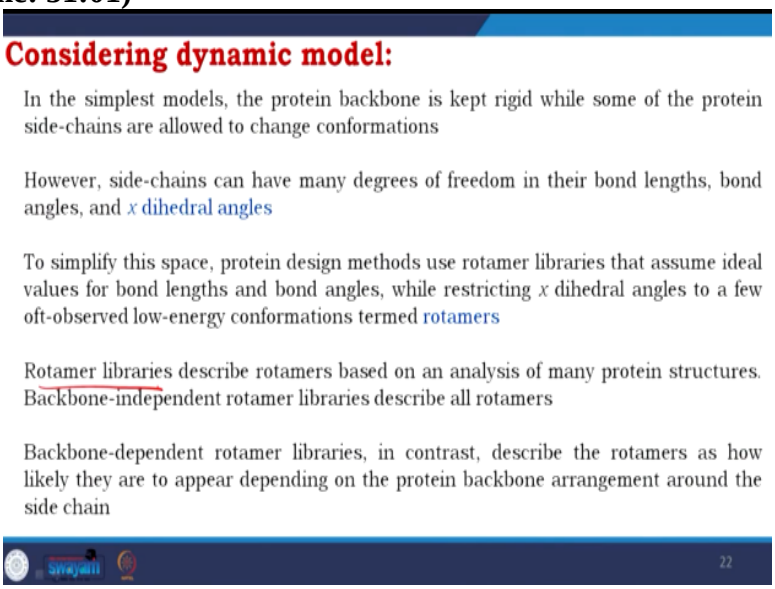
So, the introduction of dynamicity would be critical for optimizing the design approach. For example, in a protein redesign of 1 small amino acid such as alanine in a tightly packed core of a protein, very few mutants could be predicted by a rational design approach to folding the target sequence if the surrounding side chains are not allowed to be repacked.

So what I am trying to say is that suppose you have small amino acid, which is your structural periphery. So you cannot allow designing a bigger amino acid because, in this static environment, there is a possibility of a steric clash. After all, in the environment where the

small amino acid alanine or glycine exists, an arginine, in theory, is not possible. However, if the model could be swing dynamicity, we could understand what amino acid could be back there.

Thus, any design process's essential parameter is the flexibility or the dynamicity allowed for both the side chains and the backbone. If we had a dynamic model, the designing would be much more accurate.

(Refer Slide Time: 51:01)



In the simplest model, the protein backbone is kept rigid while some of the protein side chains are allowed to change conformations. However, side chains can have many degrees of freedom in their bond length bond angles and that chi dihedral angle the sidechain dihedral angles, which you have talked about before. So, what is the problem? You have phi and psi; you work on phi and psi you calculate you get Ramachandran plot.

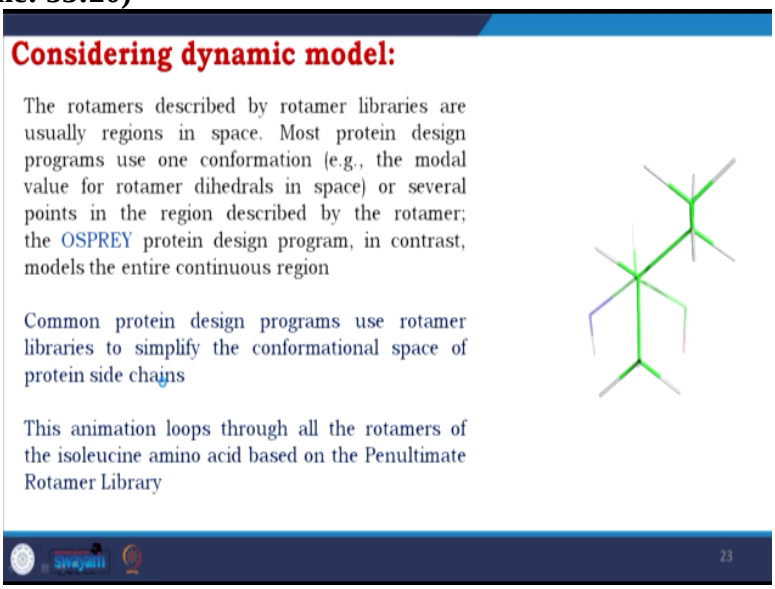
However, when you are allowed to consider the chi's, there are so many chi's that your calculation would be very complicated. You need to simplify the case and use rotamer libraries to simplify this phase protein design method; what is a rotamer library? Rotamer libraries assume ideal values for bond length and bond angles while restricting chi dihedral angle to a few observed low energy conformations termed rotamers.

So, some of the rotamers selected on different criteria are there, and you cannot change beyond that to make the process simplified rotamer libraries describe rotamer is based on an analysis of many proteins structures. So backbone independent rotamer libraries describe all

rotamer. So, what is these rotamer libraries? They take similar structures and see the possibility of the rotamer position of amino acid and the statistical consideration choose the winners.

Moreover, they are in the library, whereas there are backbone independent rotamer libraries that have backbone dependent rotamer libraries; in contrast, describe the rotamer as how likely they are appear depending on protein backbone arrangement around the side chain.

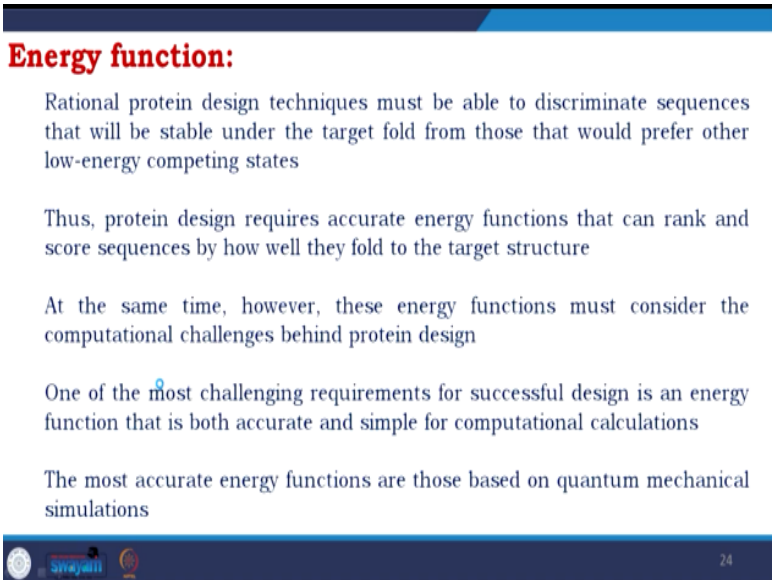
(Refer Slide Time: 53:20)



The rotamers described by rotamers libraries are usually regions in space. Most protein design programs use one conformation: the modal value for rotamer dihedral space or several points in the region described by the rotamer. The OSPREY protein design program, in contrast, models the entire continuous region. So if you see, you would see that there is a rotamer.

So the common protein design programs use rotamer libraries to simplify the conformational space of protein side chains if you look at this animation loop through all the rotamers of the isoleucine amino acid based on the penultimate rotamer library, which is one of the premier libraries at the initial stage of protein designing provided by Richardson.

(Refer Slide Time: 54:25)

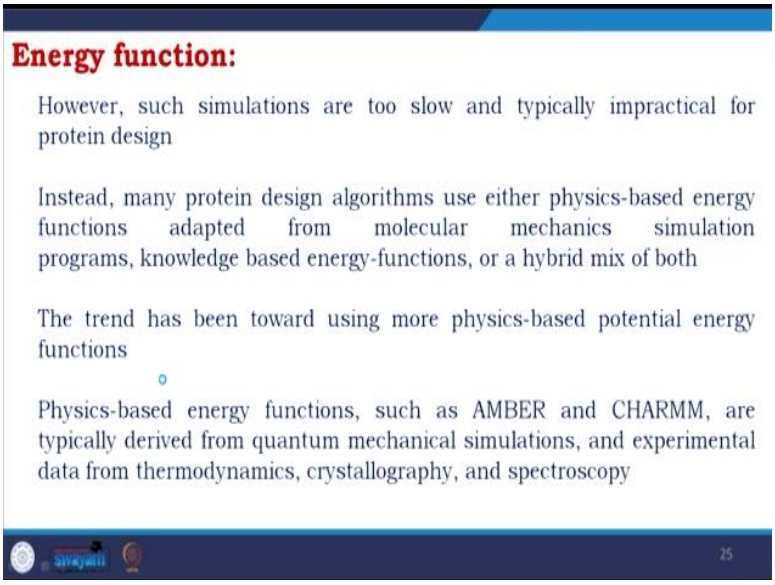


Coming to energy function, rational protein design techniques must discriminate sequences that will be stable under the target fold from those that would prefer other low energy competing states. Thus, the protein design requires an accurate energy function that can rank and score sequences by how well they fall to the target structure. At the same time, however, this energy function must consider the computational challenges behind protein designing.

So more accurate, it would be more complicated the calculations or more time-consuming. So that is the challenge, and accuracy and optimization are required. One of the most challenging requirements for successful design is an energy function that is both accurate and simple for computational calculations; the most accurate energy functions are those based on quantum mechanical simulations.

If you remember, I talked about MD simulation when I was discussing MD simulation; in MD simulation, you provide classical physics, whereas in quantum mechanics, what you are doing is the electronic level calculation, which is sometimes more accurate than the experiment like a real-life experiment.

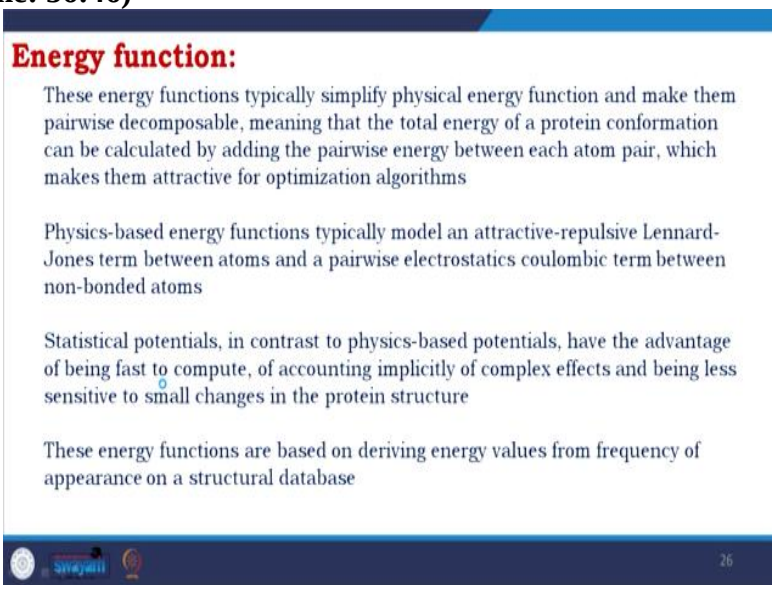
(Refer Slide Time: 55:58)



However, such simulations are too slow; as I told you, it is time demanding and computation demanding. It is typically impractical for protein designing, considering the enormous size of a protein as a macromolecule. Instead, many protein design algorithms use either physics-based energy function adopted from molecular mechanics simulation programs, knowledge-based energy function or a hybrid.

The trend has been toward using more physics-based potential energy functions. Physics-based energy functions such as AMBER and CHARM are typically derived from quantum mechanical simulations and experimental data from thermodynamics, crystallography and spectroscopy.

(Refer Slide Time: 56:46)

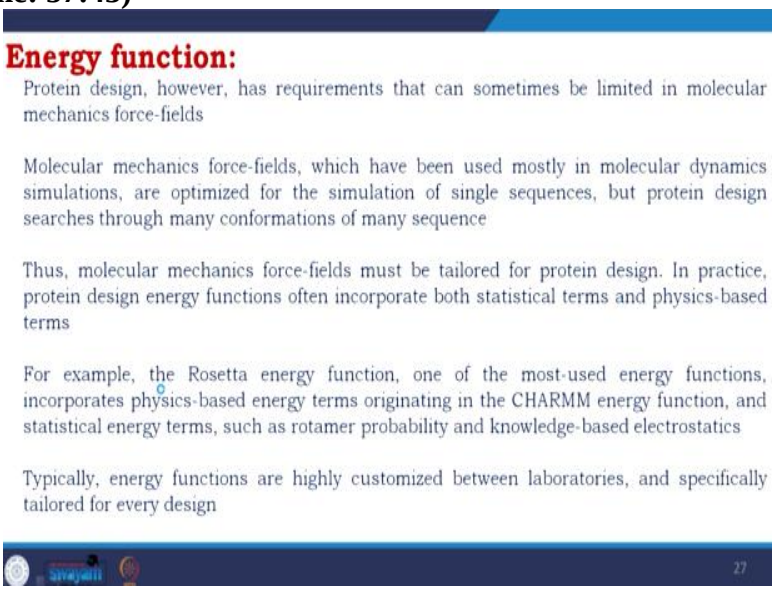


These energy functions typically simplify physical energy functions and make them pairwise decomposable. The total energy of the protein conformation can be calculated by adding the

pairwise energy between each atom pair which makes them attractive for optimization algorithms. Physics best energy functions typically model an attractive, repulsive Lennard Jones term before atoms and appear wise electrostatic coulombic term between the non bonded atoms.

In contrast to physics-based potentials, statistical potentials have the advantage of being fast to compute accounting implicitly having complex effects and being less sensitive to small changes in the protein structure. These energy functions are based on deriving energy values from the frequency of appearance on a structural database.

(Refer Slide Time: 57:43)

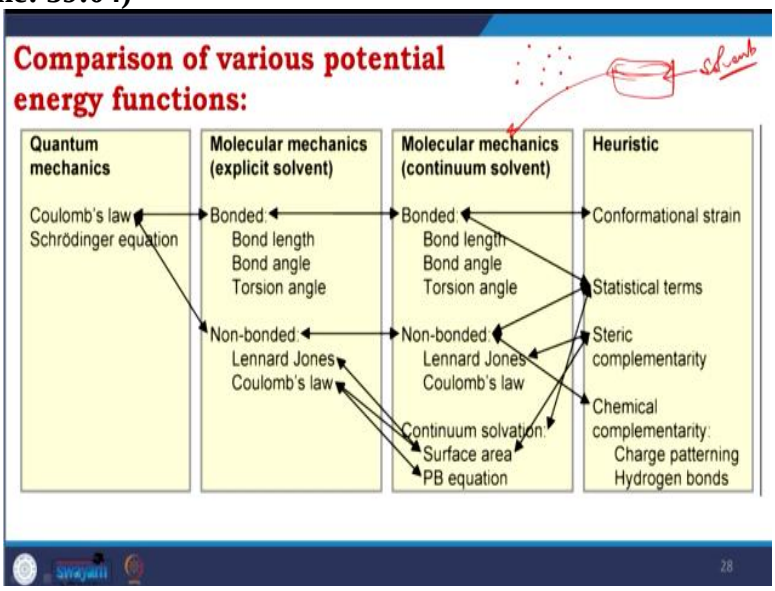


Protein design, however, has requirements that can sometimes be limited in molecular mechanical force fields. Molecular mechanics force fields, which have been used mostly in molecular dynamics simulations, which we have discussed in detail, are optimized for the simulation of single sequences, but protein design searches through many conformations of many sequences.

Thus molecular mechanics force fields must be tailored to protein design. In practice, protein design energy functions often incorporate statistical and physics-based terms. For example, the Rosetta energy function, one of the most used energy functions in corporate physics-based energy terms, originated in the CHARMM energy function and statistical energy terms such as rotamer probability and knowledge-based electrostatics.

Energy functions are highly customized between laboratories and specifically tailored for every design. So, we are trying to develop many force fields; we are trying to develop a lot of matching, but depending on cases, you have to make modifications that are protein-specific or target specific.

(Refer Slide Time: 59:04)



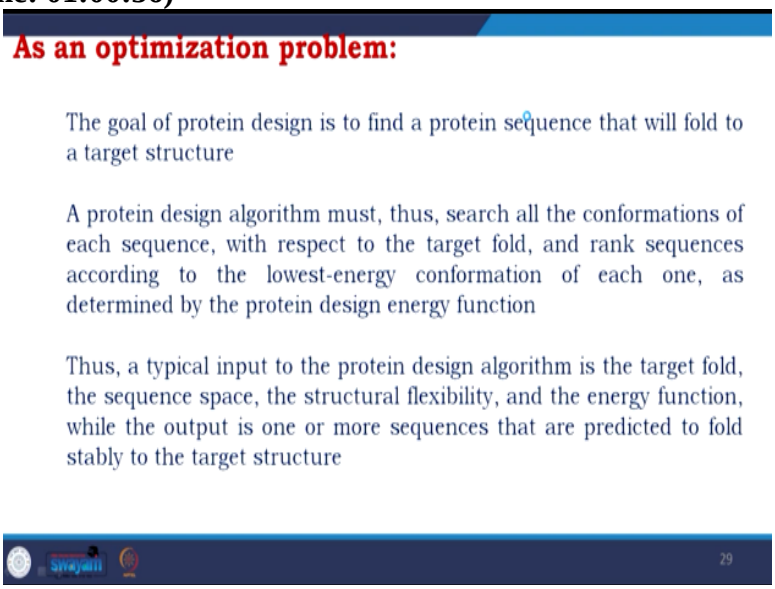
So, here is a comparison of various potential energy functions taking quantum mechanics, molecular mechanics with an explicit solvent with continuum solvent and heuristic. So, what is explicit solvent? If you remember, I talked about these in simulation; when you consider, let us say, the water model, you could consider them as discrete water molecules as individual existence water molecules.

These are explicit solvents, and when you consider them a continuous solvent with average property, it comes to continuum solvent. So in quantum mechanics, there would be coulomb's law reading the equation for the bond you have to use bond length which is stretching bond angle which is bending torsion angle or dihedral angle for non bonded use Lennard Jones potential and coulomb's law in molecular mechanics.

This is molecular mechanics with explicit with continuum solvent bonded and non bonded are same for continuum solvation you have to calculate the surface area and PB equation for heuristic you have to take the conformational strain the statistical terms which will come from different areas and which could be distributed from the information of the different areas, you have to consider the steric clashes you have to consider the complementarity you

have to consider chemical complementarity which is charged patterning and hydrogen bond distribution.

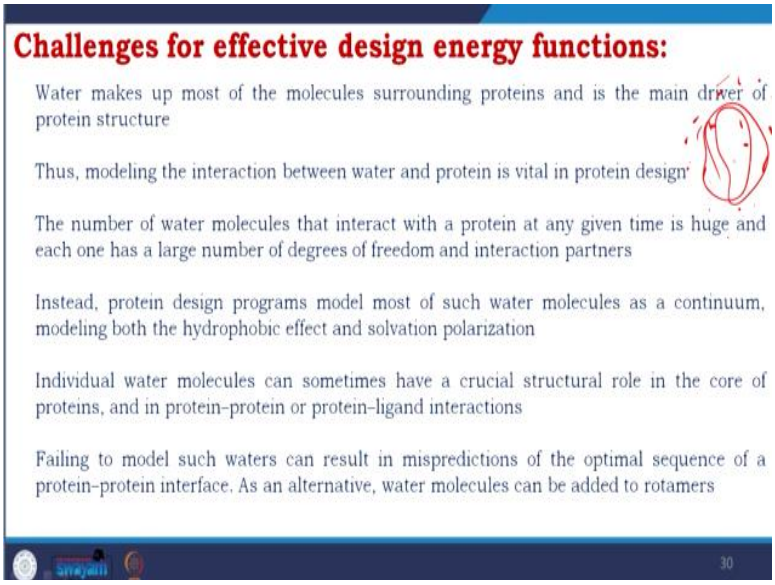
(Refer Slide Time: 01:00:56)



Protein folding design or protein design as an optimization problem, as we know, the goal of protein design is to find a protein sequence that will fold to a target structure. A protein design algorithm must thus search all the conformation of each sequence concerning the target fold and rank sequences according to the lowest energy conformation of each one as determined by the protein design energy function.

Thus, a typical input to the protein design algorithm is the target fold the sequence space, the structural flexibility and the energy function, while the output is one or more sequences that are predicted to fold stably to that target structure. So, we have the target structure; we have the sequence space, structural flexibility and energy function.

(Refer Slide Time: 01:01:55)

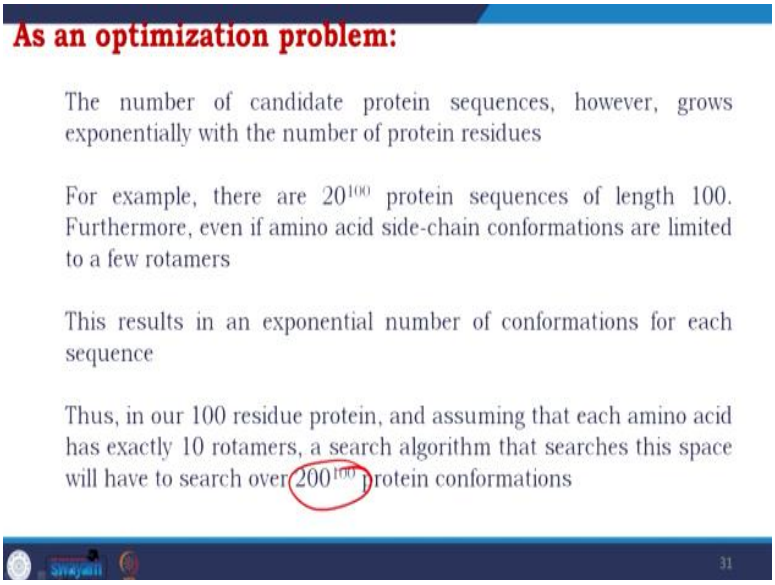


Challenges, water makes up most of the molecules surrounding protein and is the main driver of the protein structure. If you look at the protein structure we have shown you in the molecular visualization, you have seen the red dots, which are water. Thus, modelling the interaction between water and protein is vital in protein design. The number of water molecules that interact with a protein at any given time is huge, and each one has a large number of degrees of freedom and interaction partners.

Instead, protein design programs model most water molecules as a continuum modelling both the hydrophobic and solvation polarisation. Individual water molecules can sometimes have a crucial structural role in the core of the protein and protein-protein and protein-ligand interaction. Failing to model such waters can result in misprediction of the optimal sequence of a protein-protein interface as alternative water molecules can be added to rotamers.

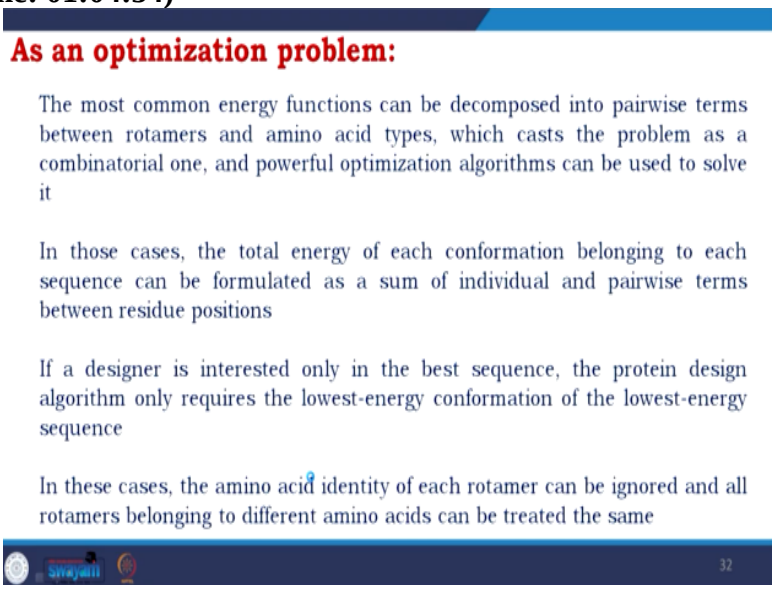
So, if you understand the problem, there is a protein with many water molecules inside and outside the protein. If you consider all of these water molecules individually, your calculation would be hugely enhanced. So, you take it as a continuum effect, but some water molecules have a catalytic or binding role. If you could identify those water molecules through previous studies, that is the best situation; if not, then there is a risk of those things taking part in the misprediction; an alternative is there to take the water and including the rotamer library those water molecules.

(Refer Slide Time: 01:03:59)



However, the number of candidate protein sequences grows exponentially with the number of protein residues; for example, there are 20 to the power 100 protein sequences of length 100 amino acids. Furthermore, even if amino acid sidechain confirmations are limited to a few rotamers, this results in an exponential number of confirmations for each sequence. Thus, in our 100 residue protein and assuming that each amino acid takes exactly ten rotamers is a search algorithm that searches this space, we will have to search over 200 to the power 100 protein conformation. So this number says what the problem is or the computational challenge.

(Refer Slide Time: 01:04:54)

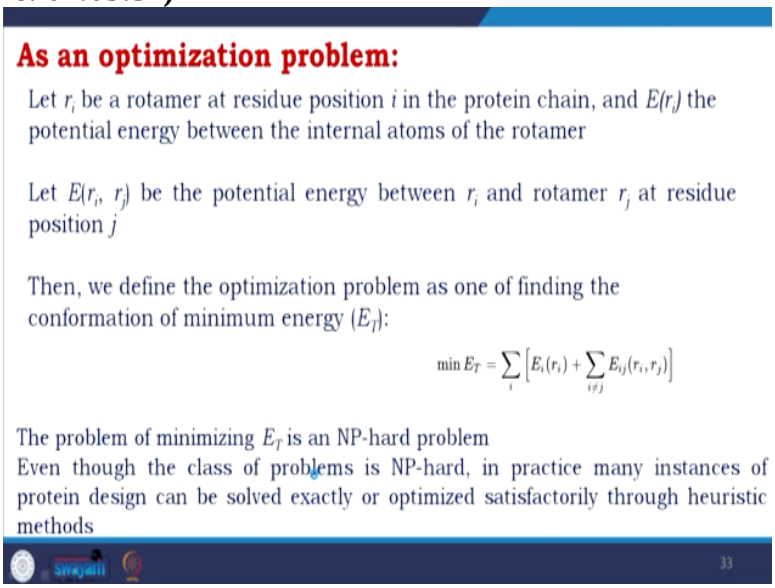


The most common energy function can be decomposed into pairwise terms between rotamers and amino acid types, which costs the problem a combinatorial one, and a robust optimization algorithm can be used to solve it. In those cases, the total energy of each conformation

belonging to each sequence can be formulated as a sum of individual and pairwise terms between residue positions.

If a designer is interested only in the best sequence, the protein design algorithm only requires the lowest energy conformation of the lowest energy sequence. So, if you fix them, you could get rid of many calculations. In these cases, the amino acid identity of each rotamer can be ignored, and all rotamers belonging to different amino acids can be treated as the same to reduce the computation.

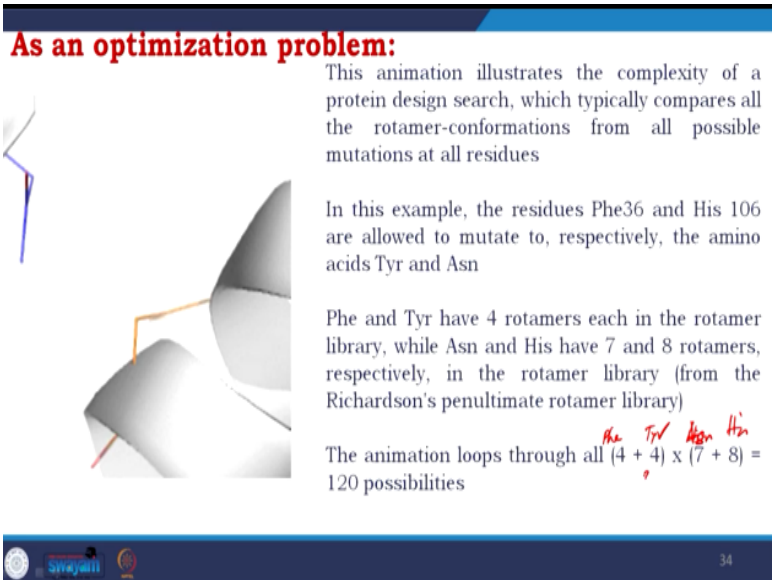
(Refer Slide Time: 01:05:52)



If you consider r_i be a rotamer at the residue position i in the protein chain and $E(r_i)$, the potential energy between the internal atoms of the rotamer. Let $E(r_i\ r_j)$ be the potential energy between r_i and rotamer r_j at residue position j. Then you define the optimization problem as one of finding the confirmation of minimum energy (E_T)= summation of E_i (r_i) + i not equal to j E_{ij} ($r_i\ r_j$) summation.

This minimization is considered an NP-hard problem. The NP-hard problem is where we consider the macromolecule of forming a polynomial working as a hard one, but even though the class of problem is NP-hard, in practice, many instances of protein design can be solved exactly or optimized satisfactorily through heuristic methods.

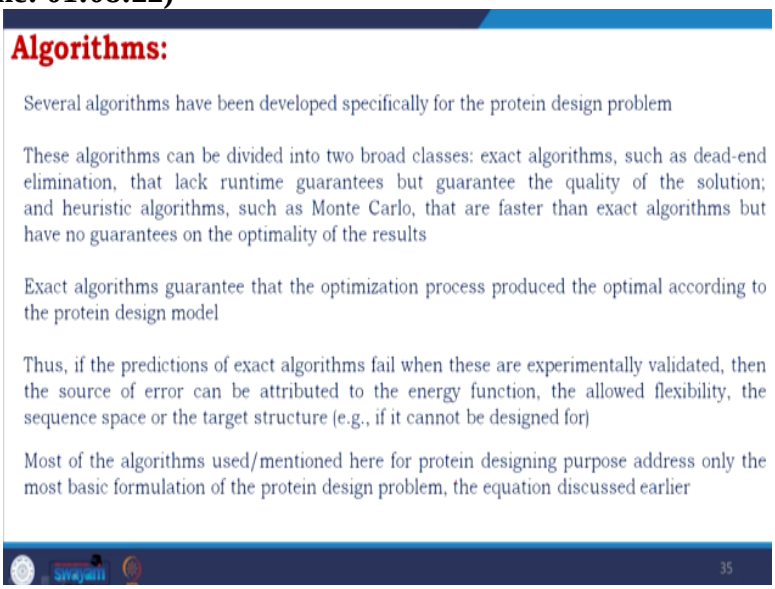
(Refer Slide Time: 01:07:03)



So if you look at here, these animations illustrate the complexity of a protein design search. By the way, this is taken from a structure of myoglobin; This typically compares all the rotamer confirmation from all possible mutations at all residues. In this example, we are taking phenyl 36 and his 106; they are allowed to mutate to, respectively, the amino acid tyrosine and asparagines, phenylalanine and tyrosine have four rotamers each in the rotamer library. In comparison, asparagine and histidine have 7 and 8 rotamers in the library.

As I told you, this is Richardson's penultimate rotamer library. So the animation loop through 4 for phenylalanine, 4 for tyrosin and then 7 for asparagine, and 8 for histidine, so eight into 15 = 120 possibilities.

(Refer Slide Time: 01:08:22)

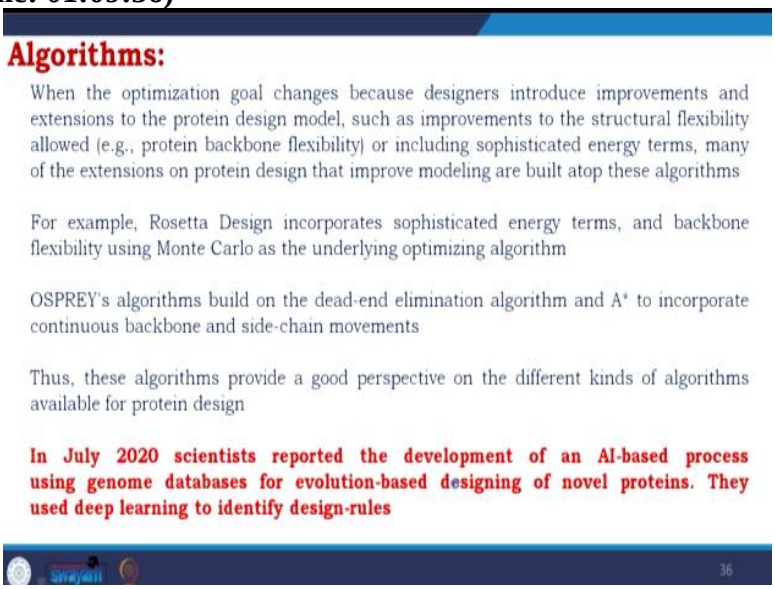


Coming to the algorithms, several algorithms have been developed specifically for protein design problems. These algorithms can be divided into two broad classes, exact algorithms,

such as dead-end elimination, that lack runtime guarantees but guarantee the quality of the solution and heuristic algorithms such as Monte Carlo, which we have explained in time of MD simulation that is faster than exact algorithms but has no guarantees on the optimality of the results.

So, two types of algorithms: exact algorithms take a lot of time but give you accurate results, and other heuristic algorithms such as Monte Carlo, which would be time-saving but does not guarantee the optimization. Exact algorithms guarantee that the optimization process is optimal according to the protein design model; thus, the; prediction of exact algorithms fails when these are experimentally validated; the source of error can be attributed to an energy function that allows flexibility, the sequence-based or the target structure. Most of the algorithms used or mentioned here for protein designing purposes address only the most basic formulation of the protein design problem; the equation is taken, which we have already discussed for measuring the minimum energy ET.

(Refer Slide Time: 01:09:56)

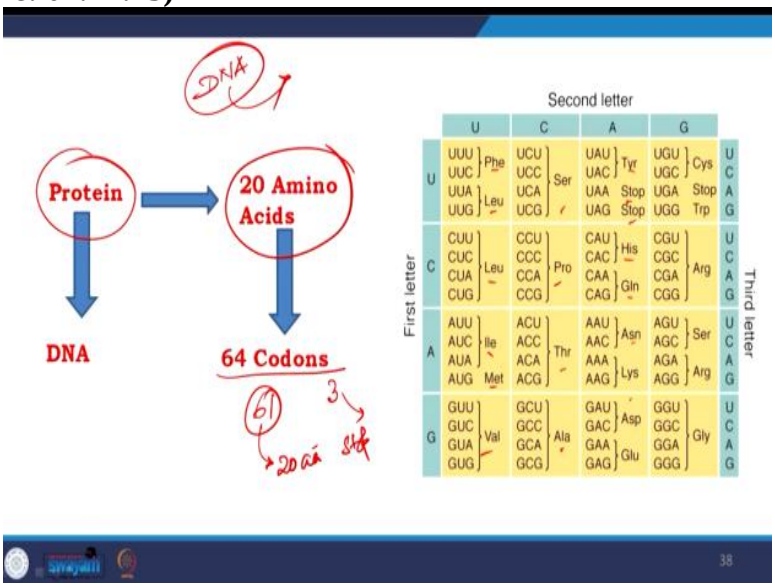


When the optimization goal changes, designers introduce improvements and extensions to the protein design model, such as improved structural flexibility (e.g. protein backbone flexibility) or including sophisticated energy terms. Many extensions on protein design that improve modelling are built atop this algorithm. For example, Rosetta design incorporates sophisticated energy terms and backbone flexibility using Monte Carlo as the underlying optimizing algorithm.

OSPREY's algorithm is built on the dead-end elimination algorithm and A' to incorporate continuous backbone and sidechain movements. Thus these algorithms provide a good perspective of the different algorithms available for protein design? Good to mention, though; I will discuss that in detail later. In July 2020, scientists reported the development of an AI-based process using genome databases for evolution based designing of novel proteins; they use deep learning to identify design tools; we will talk about that in detail.

(Refer Slide Time: 01:11:09)

(Refer Slide Time: 01:11:13)

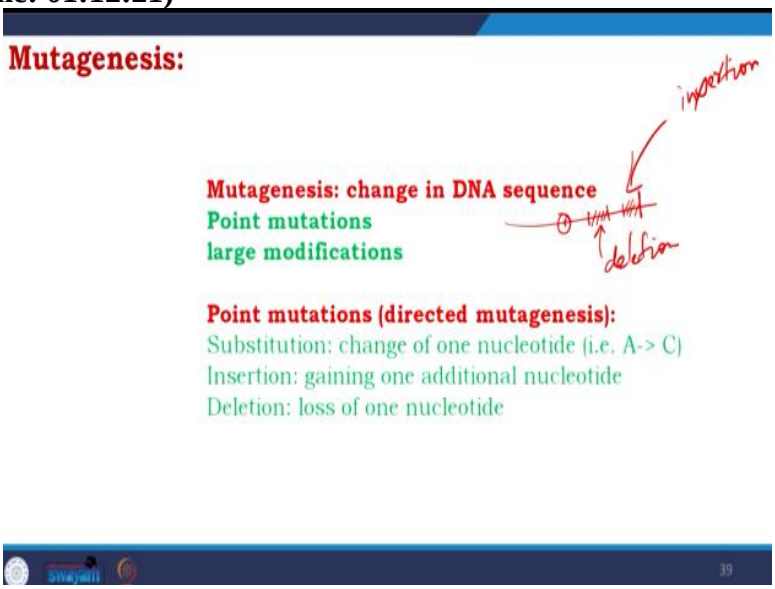


So, coming to wet-lab experiments, what we have done now, we have designed a protein. As you know, when we do that wet-lab experiment, you cannot do that on protein. So, we have to go to the DNA. As we all know, we have 20 amino acids, and for these 20 amino acids, there are 64 codons 61 specifically for these 20 amino acids and three is stop codon; this is

the codon library, where you see the relations between the amino acids which are here these are stop histidine and with the nucleosides.

So, further, we will work on modifying DNA because in biology, the only change that goes further is DNA, as we have discussed in previous classes.

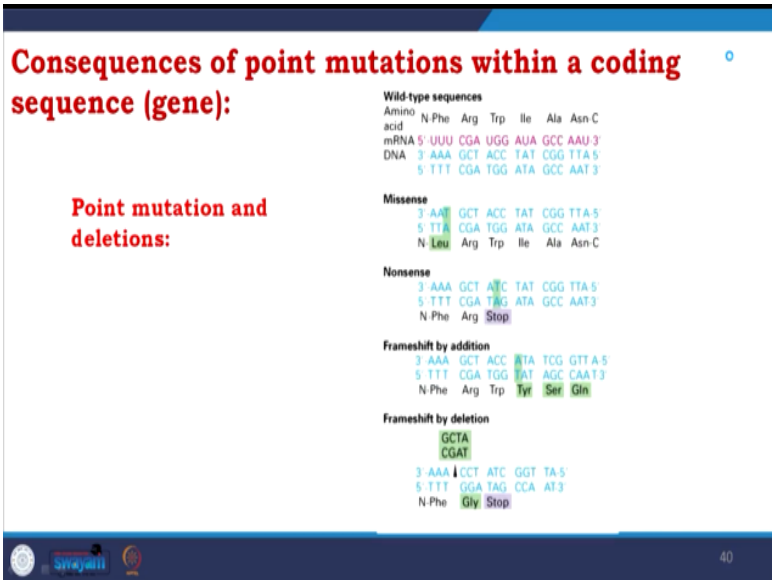
(Refer Slide Time: 01:12:21)



So, we have to perform mutagenesis; as we all know now, mutagenesis is the change in the DNA sequence; it might be point mutation, it might be large modifications; I talked about them while discussing protein structure. So, this is a point mutation if you want to cut this part. So, this is deletion; if you want to incorporate a new part, you incorporate an extra part; this is insertion.

So, in that way, large modifications could be done. So, point mutations or directed mutagenesis substitution, change of 1 nucleotide insertion could also be pointed, gaining one additional nucleotide deletion loss of 1 nucleotide they are coming under point mutation.

(Refer Slide Time: 01:13:22)



So we will discuss different types of point mutations within the coding sequence and how that is incorporated towards mutagenesis and that effect. So, today, we have discussed basic principles about what we target towards protein engineering, how we plan, what is our goal, what type of things we want to see change and the major process of development of protein designing, one is rational designing, where we go and do our designing.

Another is directed evolution, where we mimic nature; we have discussed how this rational designing happened, and in the next class, we will talk about the further consequence of those protein mutations on rational designing, and we will start the other one, the direct evolution. Thank you very much for listening. Thanks.