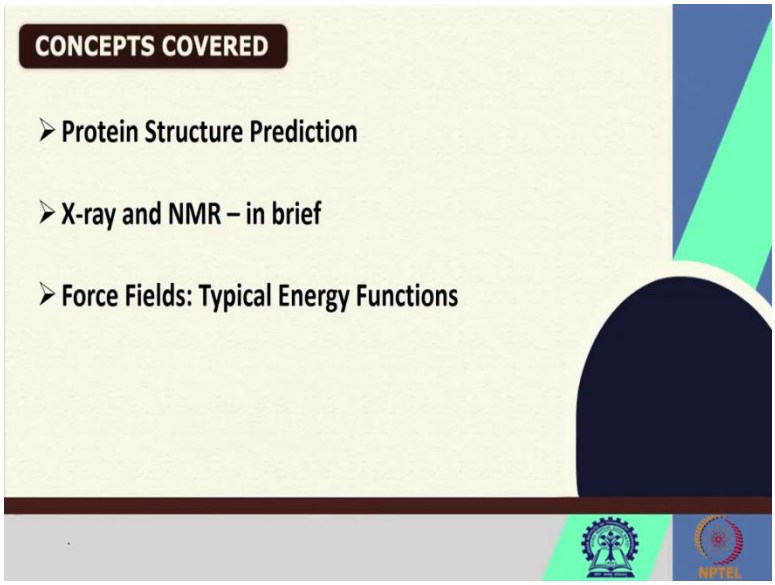
Fundamentals of Protein Chemistry
Prof. Swagata Dasgupta
Department of Chemistry
Indian Institute of Technology, Kharagpur

Module - 03
Protein Structure and Stability
Lecture - 13
Protein Structure Methods

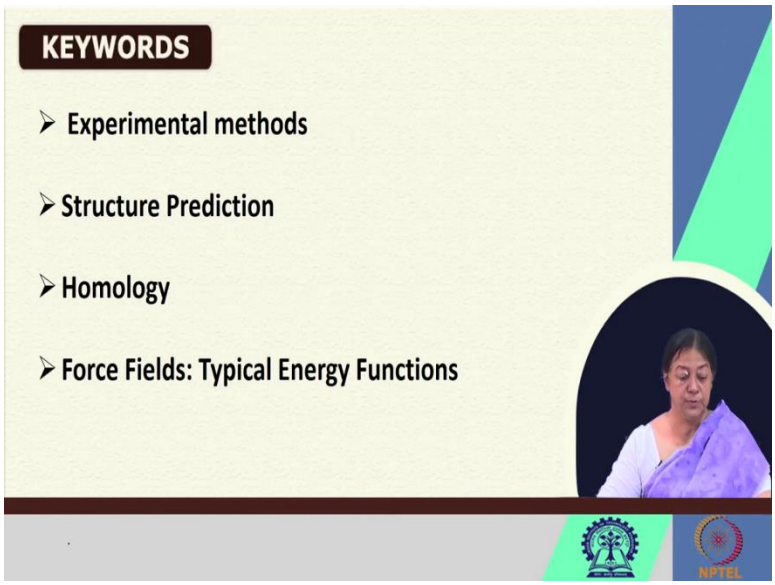
In this lecture of the module on protein structure and stability, we will be talking about protein structure determination methods. In this lecture we will not be going into the details of any of the methods, but just have a brief overview of what methods are followed for protein structure determination.

(Refer Slide Time: 00:38)



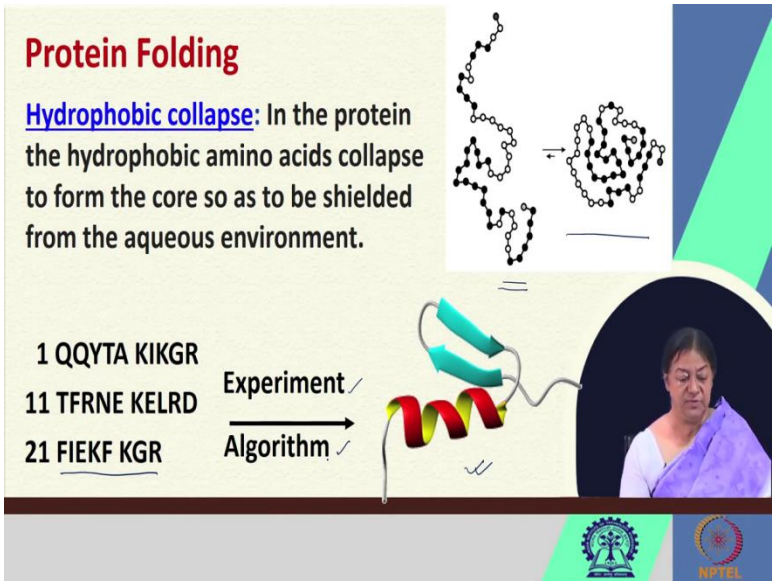
We realize the importance of protein structure, because we know that this folded structure is going to give us information about the function of the protein, which is important to design say any molecules as inhibitors for enzymes and so on and so forth.

(Refer Slide Time: 00:53)



We will now look at the different types of ideas related to protein folding.

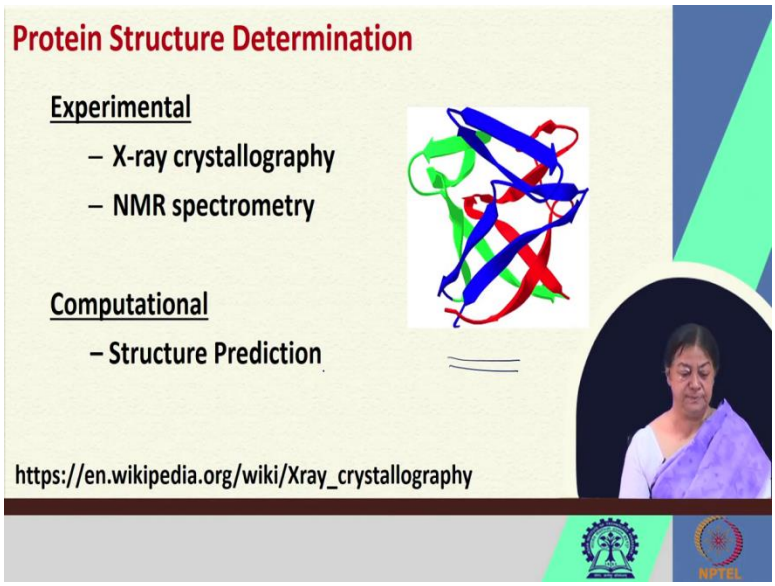
(Refer Slide Time: 00:56)



Now in the protein, the hydrophobic amino acids we know collapse to form a hydrophobic core. So, if you just imagine that we have a protein chain that looks like this [refer to slide] and the circles that are black are the hydrophobic ones. We would expect that in a folding process all the black ones, would tend to come to the center part of the protein, forming a structure like this. But we realize that this can have umpteen possibilities in the way it folds.

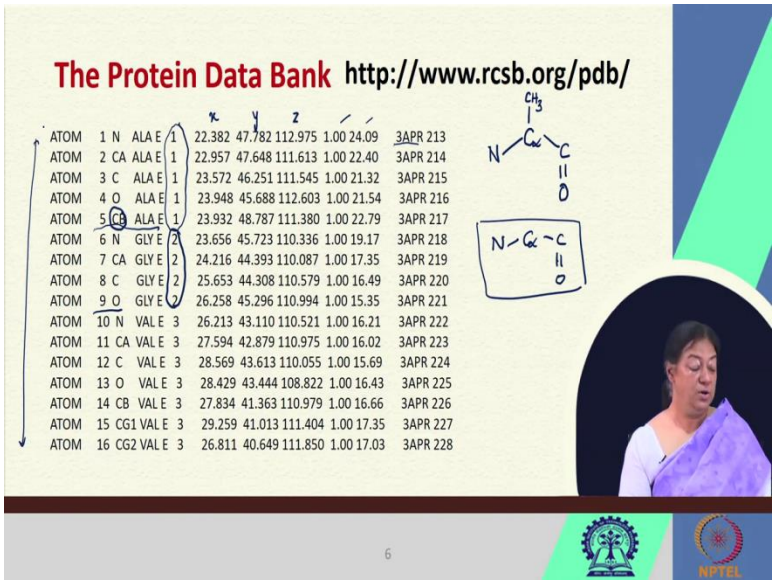
This is our one letter amino acid sequence, now to go to the final folded structure what are the methodologies that we can use. We can have experiments, we can have algorithms, or prediction methods that are going to tell. And we will briefly touch upon some of these.

(Refer Slide Time: 02:03)



If we look at the experimental techniques that are known, we have X-ray crystallography and NMR spectrometry. They tell us about the folded structure and the computational methods tell us about structure prediction, from the amino acid sequence or from known folds of proteins.

(Refer Slide Time: 02:22)



The interesting thing about having a protein structure is, we can look at a three-dimensional structure on a visualization software, where we have the x y z coordinates of every atom of the protein.

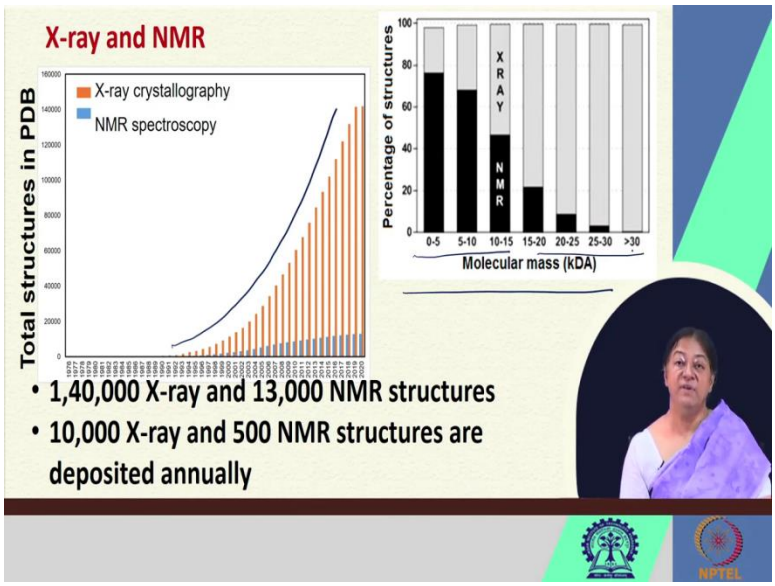
This [refer to slide] is a typical protein data bank file; the file name here is 3 APR. It is an alphanumeric code. And if you look at the particular set of atoms, say 1 through 5, they are all

marked 1, indicating that this is residue number 1. And what we have here is we have the 5 atoms, the hydrogens are not seen here, we have the nitrogen of the alanine, we have the Cα that is marked CA, then we have the C = O, and we have what are called the x, the y, and the z coordinates. These are the occupancy and the b factor, but we do not need to worry about that now.

When we are talking about a protein structure the information that we want is the x, y, z coordinates of every atom of the protein. We know that alanine has a CH_3 as its side chain, so we have a Cβ that is a CB. Then we have the glycine, where we see residue number 2 is glycine and that there is no side chain for glycine. All we have is the N, the Cα, the C and the O. This indicates that this is glycine.

This is how we have the protein data bank representations. And the solving of a structure means we deposit our coordinates to the protein data bank. So we have this information for visualization on a visualization software.

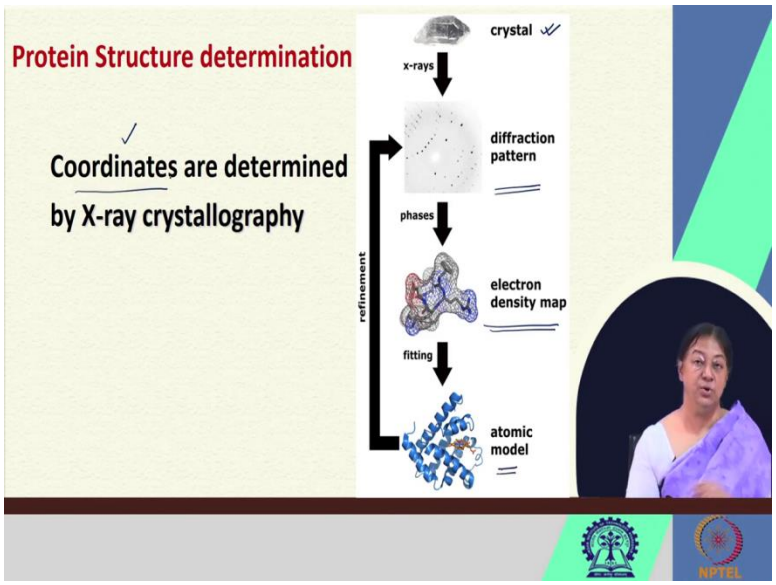
(Refer Slide Time: 04:34)



If we look [refer to slide] at the experimental techniques, the X-ray and the NMR, there are around 140000 X-ray structures and 13000 NMR structures. And if we look at the exponential rise over the years, this is in 2020 the last one, where the X-ray are in orange and the NMR are in blue.

And when we look at the percentage of the structures in terms of the molecular weight, we see that in the smaller molecular weight range we have NMR predominantly. But as we get to larger and larger molecular weight, then we have X-ray structure determinations predominate.

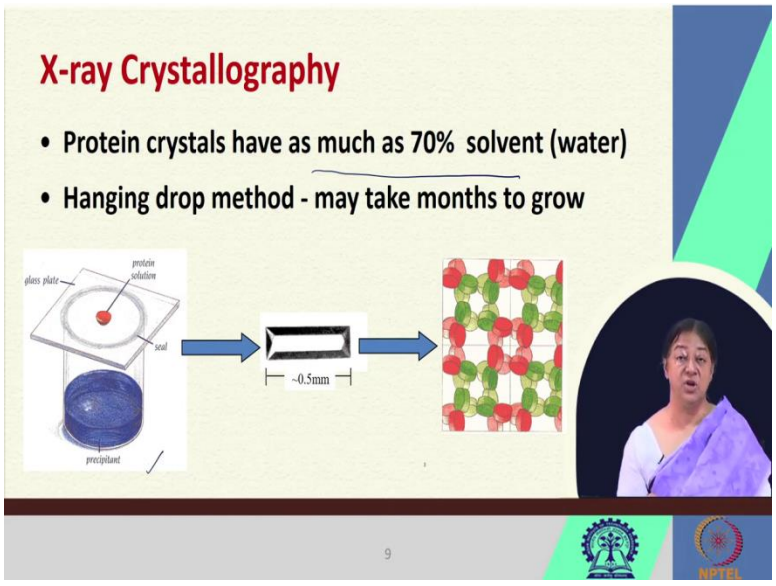
(Refer Slide Time: 05:18)



When we look at the coordinates and at an X-ray structure, we see a snapshot in time. So, we have the crystals and we have a diffraction pattern (this is a very simplistic way in which this structure is obtained), but it takes years to do this - better ways to read electron density maps as they are called.

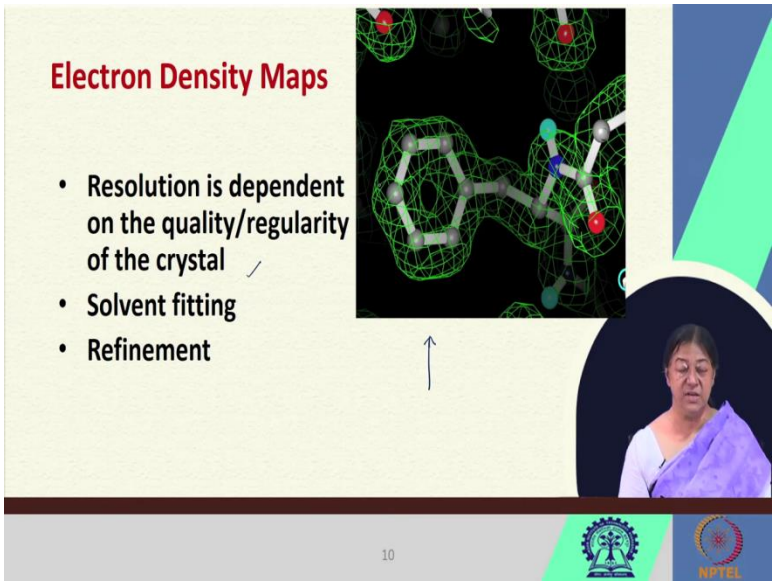
So we have a diffraction pattern depending upon the location of the atoms in the crystal. This then has different electron density maps that are created from this, giving us an atomic model. There is constant refinement of this atomic model to lead to the coordinates that are going to be determined by X-ray crystallography.

(Refer Slide Time: 06:10)



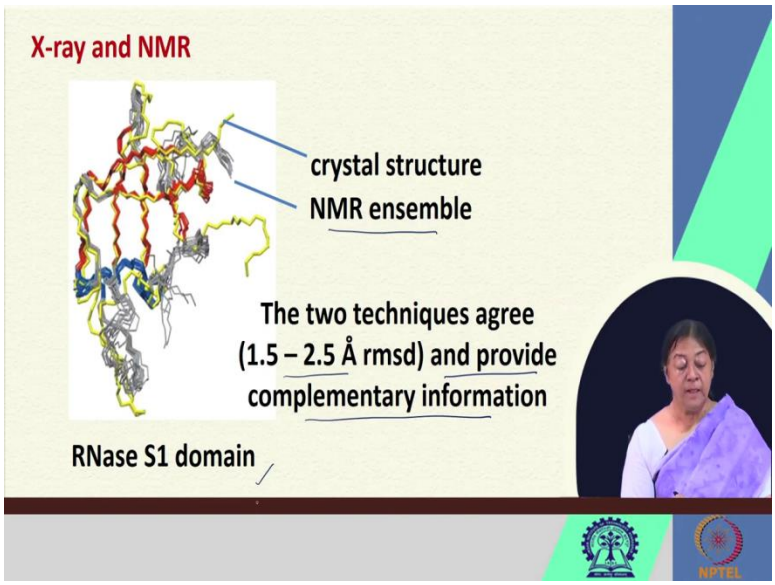
Now, protein crystals themselves have as much as 70 percent of water, which makes them very difficult to grow. One of the methods of growing is called a hanging drop method. They may take months to grow and then on top of that to solve the structure is also very difficult.

(Refer Slide Time: 06:29)



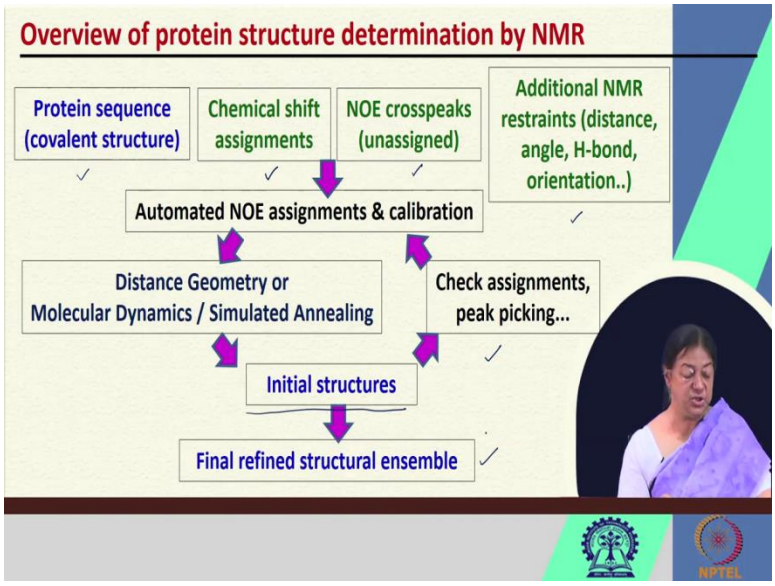
So, we have the resolution that is dependent upon the quality and the regularity in the crystal, there is solvent fitting, there is refinement and this [refer to slide] is a typical electron density map where given the electron density, the specific atoms are tried to map to the electron density. Here what we see a phenylalanine residue being mapped to the electron density. So from the diffraction pattern we have the phases, we have the electron density map, we get the atomic model with constant refinement, giving us a three-dimensional structure. And it is definitely not as easy as it sounds.

(Refer Slide Time: 07:11)



In this [refer to slide] case we are looking at a comparison between the X-ray and the NMR. The two techniques agree at a lower resolution and they provide complementary information. And this is the comparison of a crystal structure and the NMR ensemble as we see for the RNase S1 domain.

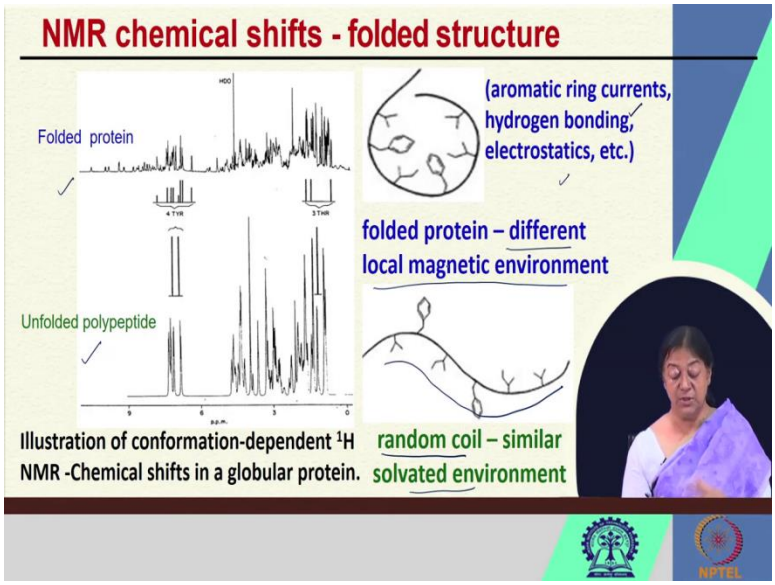
(Refer Slide Time: 07:36)



When we look at an overview of the protein structure determination by NMR methods, we know the protein sequence, we have chemical shift assignments, we have NOE cross peaks (which we will briefly know what they are) and we have additional NMR restraints that tell us that these are specific interactions that we can look for in our NMR spectra. Then there are assignments, there is distance geometry, there is peak picking and a lot that gets into the initial structures.

Finally, we have a refined structural ensemble and because we are looking at a solution, what we see is we see different motions that are possible in NMR, but for X-rays we look at a snapshot in time; but the b factor tells you that some variations are possible.

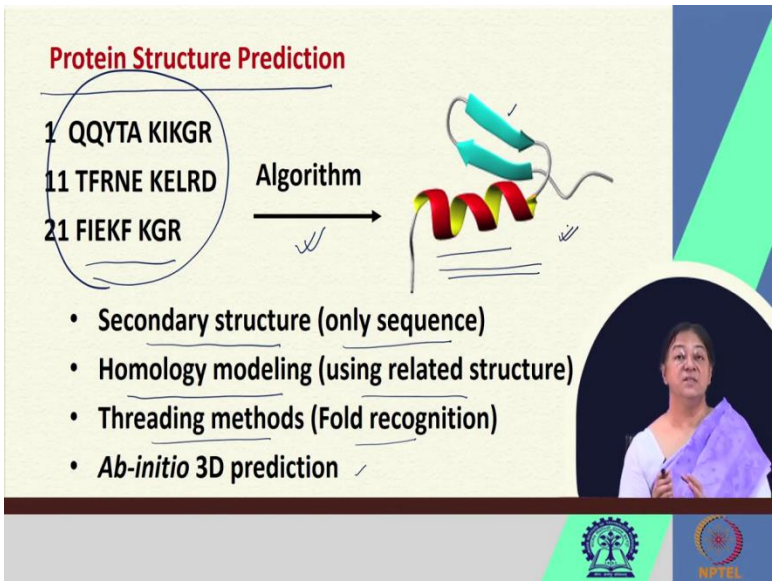
(Refer Slide Time: 08:32)



In a folded protein, NMR looks at chemical shifts. So, when we look at a folded protein as opposed to an unfolded protein, they are in a different environment. In the folded protein, we have a different local magnetic environment that arises from the aromatic ring currents, the hydrogen bonding, the electrostatics; but when we open this fold, when we have the random coil, we have a similar solvated environment for the polypeptide chain.

We have a conformation dependent NMR that results in a chemical shift of these proteins. So, when we are looking at a folded protein, we have a different environment as opposed to a random coil that we observe here [refer to slide].

(Refer Slide Time: 09:21)



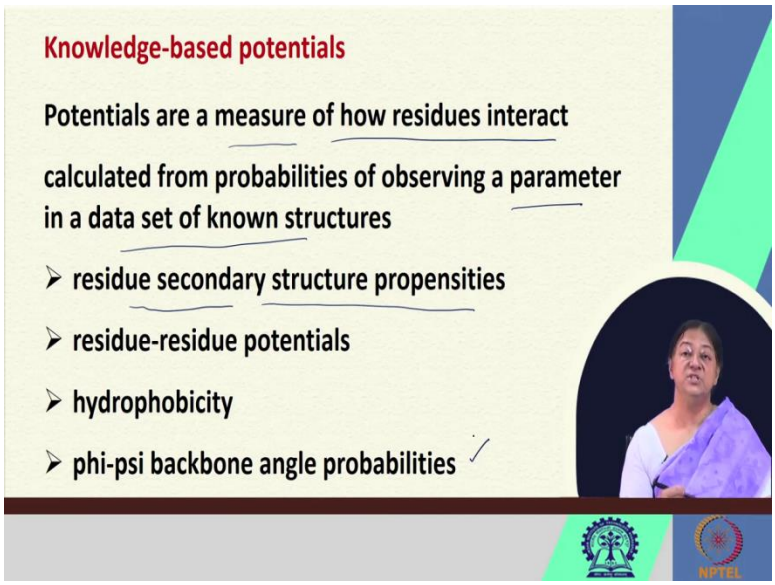
When we are looking at methodologies for our protein structure prediction, we can do it through experimental methods. The experimental methods from the nuclear overhauser effect where we looked at the chemical shifts, the pairing, the restraints; these give us NMR structural possibilities, a whole ensemble of structures.

When we look at X-rays, we are looking at a snapshot in time, we are looking at a specific crystal that has to be formed, from the crystal we get the X-ray diffraction map, from the X-ray diffraction pattern rather we get the different phases that we can observe, and from that the electron density, and from the electron density we get the final atomic model from several refinements. When we look at a structural prediction method, we have our one letter amino acid sequence and we want to get to our folded structure. So there must be some prediction methodology that could tell us how we can get to the folded structure.

Now, when we look at the secondary structural method and we look at the sequence only, there are software that tell us whether a specific sequence will form a helix or strands of a β-sheet.

We have homology modeling, we have threading methods that look at fold recognition, homology modeling that looks at related structures and Ab-initio 3D predictions which, from scratch looks at the sequence of the protein and from specific force fields or energetics of protein folding try to get us the final folded structure.

(Refer Slide Time: 11:16)

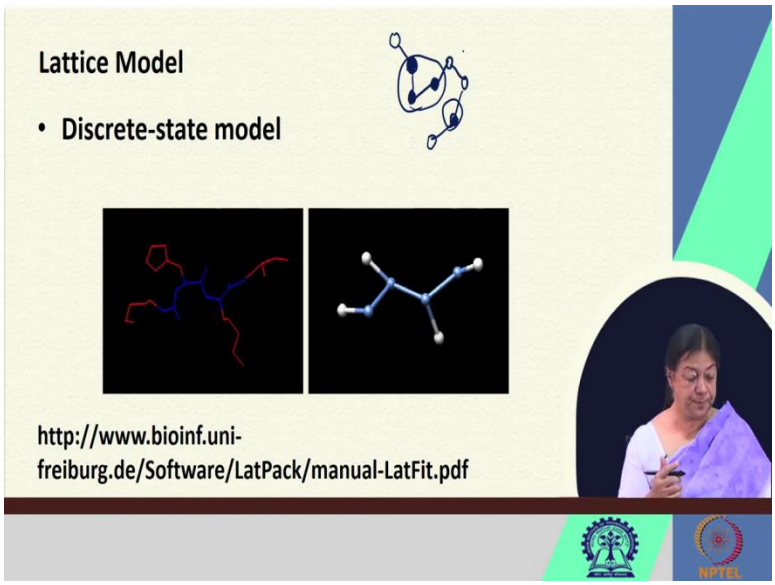


So, when we look at what are called knowledge based potentials, they are potentials of how residues interact with each other. So, they are calculated from probabilities of observing a specific parameter of a data set of known structures.

For example, we look at residue secondary structure propensities. What do we mean by this? From the statistics that we can do on the available database, we realize that there are some residues that would prefer to remain in an α-helix, there are some residues that would prefer to remain in a β-sheet and there are some residues that are helix breakers or helix makers and so on and so forth. So these would contribute to our knowledge of the structural aspects of the specific residues.

In addition we know the hydrophobicity factor, there is a steric factor, there are residue-residue potentials and there are phi-psi backbone angle possibilities, that we know. We have specific regions of the Ramachandran plot that are allowed for specific conformations. This gives us some idea of what the structure might or might not be in terms of any steric interactions, in terms of any residue-residue propensities, in terms of residue-residue potentials, or in terms of what we might expect a particular residue to be in a particular structure.

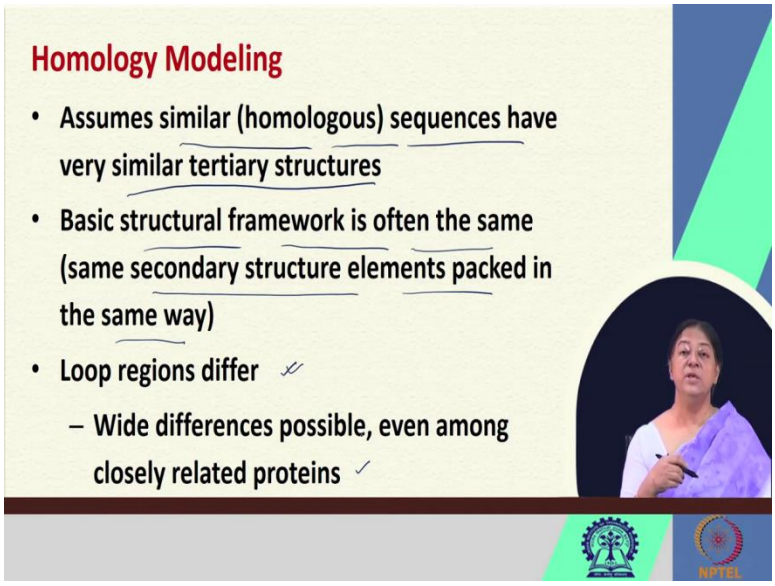
(Refer Slide Time: 12:57)



So when we have a discrete state model, we have what is called a lattice model, where the hydrophobic interactions sequester in the middle of the protein.

If we have say circles like this [refer to slide] and we have the darker ones that we can consider to be hydrophobic in nature, when we have a chain that is a random chain like this, all these dark ones would tend to come together in the central core of the protein, then being hydrophobic in nature.

(Refer Slide Time: 13:41)



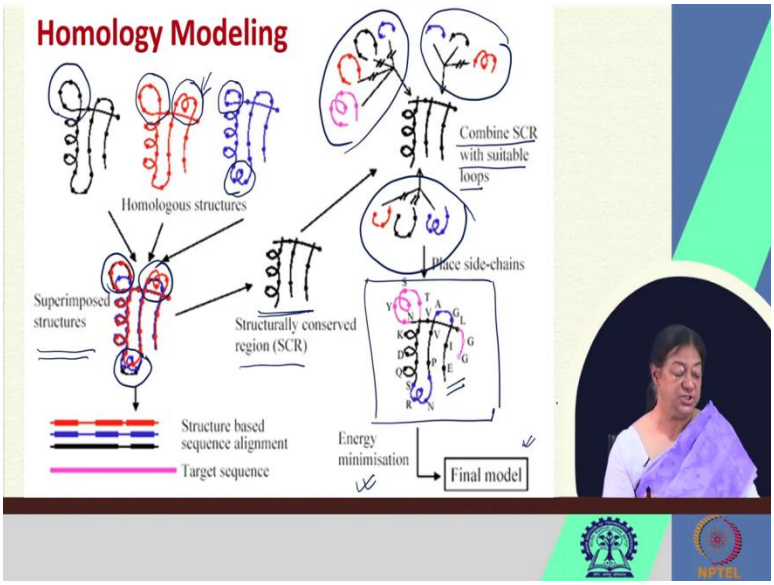
In homology modeling, there is an assumption that the similar homologous sequences would have similar tertiary structures. In the sense that if we have a family of proteins, it is likely that given their sequence because this family of proteins say performs a similar function, it might be

expected that these homologous sequences or similar sequences would have similar tertiary structures.

Now, this basic structural framework is often the same. The same secondary structure elements are packed in the same way because they give this overall protein scaffold. And the loop regions may differ, that is the connectivity or the linkers between the secondary structures may differ. There are sometimes even very wide possibilities even between closely related proteins.

But nevertheless, if they belong to the same family and the active side say of enzymes are similar in a typical family, then for example, if you are looking at ribonucleolytic proteins; all of these have an enzymatic activity related to RNA cleavage. These would be expected to have a similar disposition of their residues around the active site of the proteins. So, they would be homologous in nature.

(Refer Slide Time: 15:09)



For example, if we look at a homology protein like this [refer to slide]. If these are homologous structures we can see that there are some distinct characteristics. For example, if we look at them closely there are some regions that are structurally conserved. If we try and superimpose these structures one on top of the other, we will see that there are variations in the linker regions. Say here and here, where this part is different, and this part is smaller, this part is larger and so on and so forth.

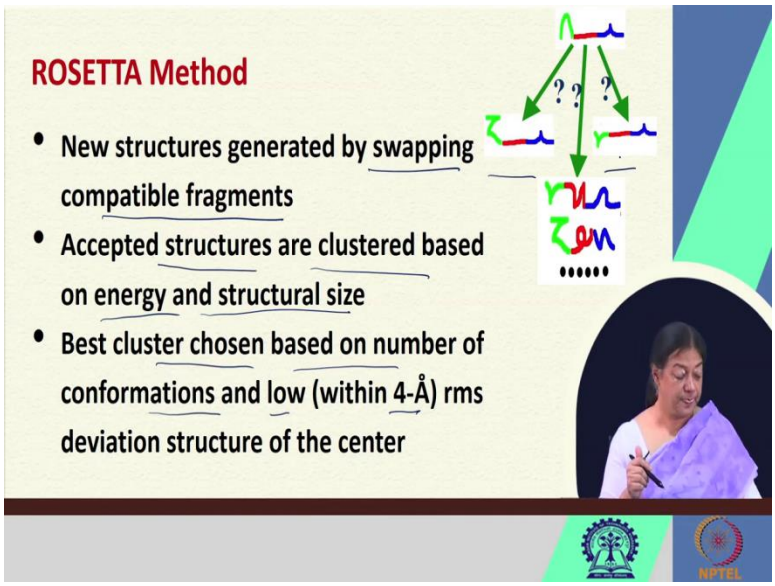
Then again we have this additional sort of a turn here. If we superimpose these structures and we look at a structure based sequence alignment, as it is called, we will see variations around certain regions. But there are the structurally conserved regions, that is the rest of the protein.

So we look at the combination of the structurally conserved regions, with different types of connectors or linkers as we call them and what we can do, is we can place the specific side

chains that we know are in these linker parts, considering that the rest of the molecule or the rest of the polypeptide chain has this structurally conserved region.

When we have these connectivities put in place, we can place the side chains, do what is energy minimization. We will be visiting that in a moment, and come up with a final model that takes care of the structural conserved region, and takes care of the linkers in terms of the side chains that we see here.

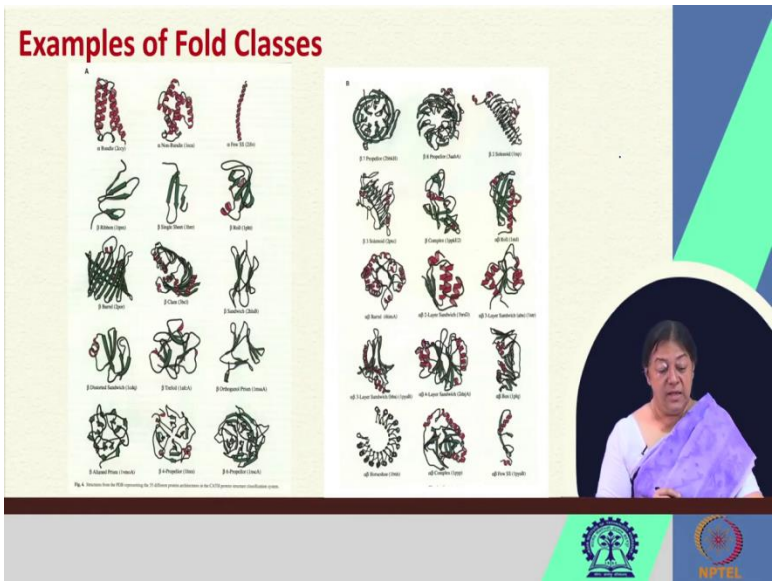
(Refer Slide Time: 17:00)



In a Rosetta method, there are new structures generated by swapping compatible fragments. So the accepted structures are clustered, based on energy and structural size and the best cluster is chosen based on the number of confirmations and low (that is within a 4 angstrom) root mean square deviation structure from the center.

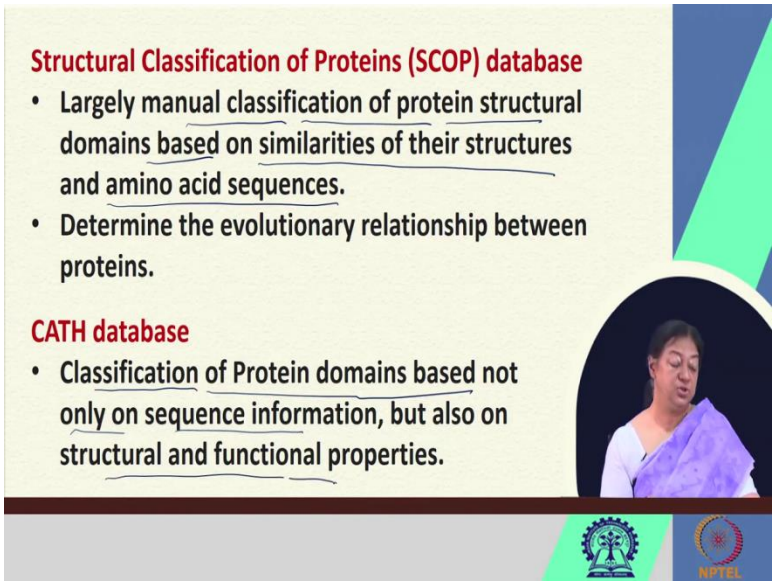
So, there is connectivity of bits and pieces that are put together in a sort of a puzzle to bring us the protein structure.

(Refer Slide Time: 17:33)



If we look at the different examples of fold classes [refer to slide], there are up to a library of folds that has a 1000 possible intermediate structures. So, we would look at whether it belongs to a specific family, whether it is a specific fold and the structure then would adopt a specific fold where we can look at what is called threading and then look at specific scores. It is absolutely a puzzle to find out what this is.

(Refer Slide Time: 18:03)



And if we look at the specific databases for the structural classification, there is the SCOP database, such as the manual classification of protein structural domains, that are based on similarities of their structures and amino acid sequences, which is useful for a structure prediction, knowing that you know the sequence. And you can search for this specific type of

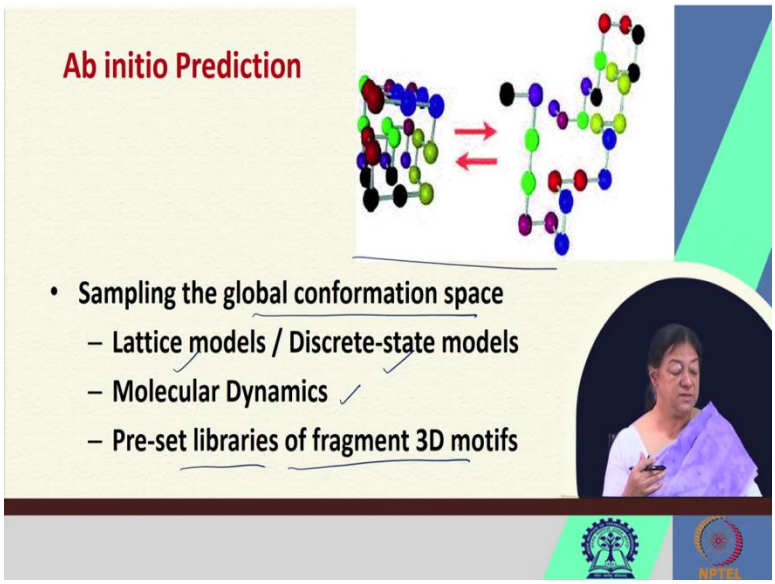
sequence in the protein data bank sequences that are available and then look for similarities in structures in ways like this.

Then we also have the other CATH database that the classification of protein domains based not only on sequence information, but also on structural and functional properties. So, depending on what database we would like to search for, the knowledge that we have is the amino acid sequence. And what do we want? We want to find out the three-dimensional structure.

So there are methodologies that build up the structure from the bits and pieces that we have or we go for homology modeling, saying that we are looking at a sequence that has a homologous sequence, with a structure that is already available in the protein data bank. What is the protein data bank? It is the repository of three-dimensional structures where we have the coordinates of the proteins.

This helps us in identifying the protein structures. And why do we need to know the protein structures? We need to know the protein structures because we want to know their function, as we know that we have a relation between the structure and the functional properties of the proteins.

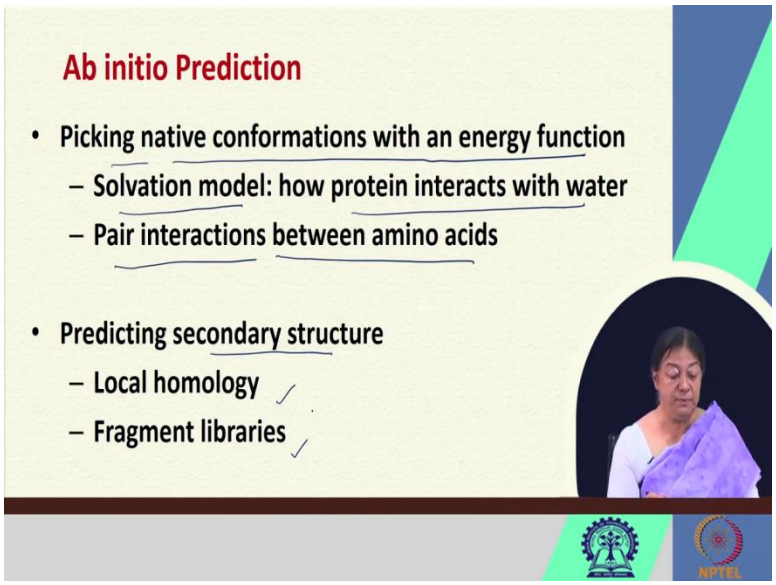
(Refer Slide Time: 19:42)



Now if we look at ab-initio prediction, where we have a protein sequence and we want to know how it folds.

Given that we do not find the similarity and we want to start from scratch, in an understanding that when we are looking at the global conformation space, we have the lattice models, the discrete state models, (let us say we can put all the hydrophobic portions together), we have molecular dynamics that look at the motions of the molecules, we have a preset library of fragments of three-dimensional motifs giving us specific structural aspects.

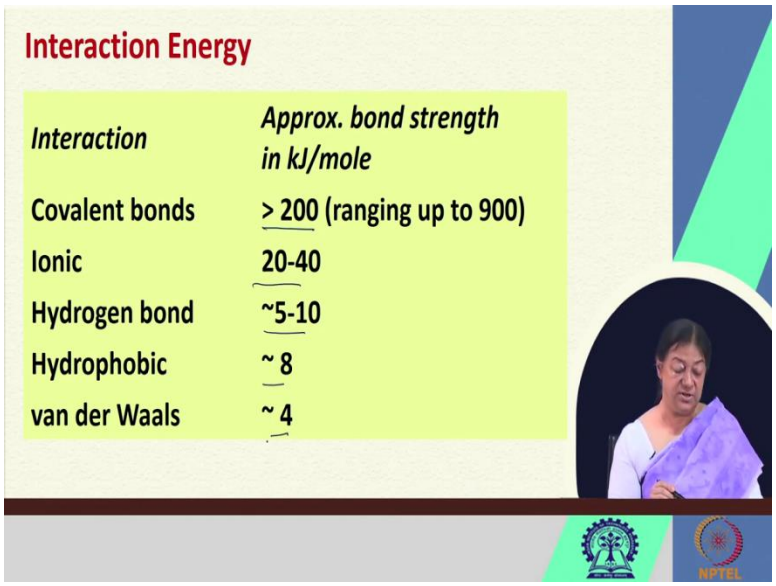
(Refer Slide Time: 20:20)



We can also pick native conformations with an energy function. We can look at solvation models, that is how the protein interacts with water, because that is an important aspect of protein folding. We can look at the pair interactions between the amino acid, the knowledge based information that we have.

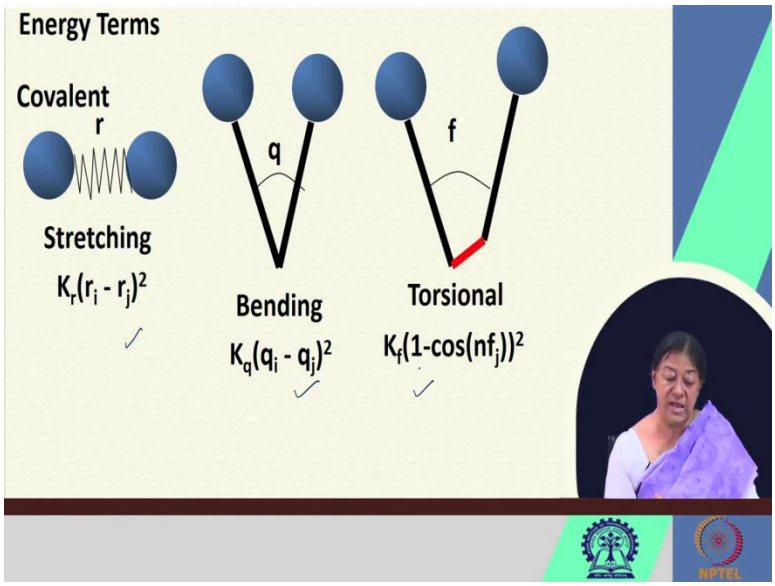
Then we can look for predictions of secondary structure from local homology and from fragment libraries.

(Refer Slide Time: 20:53)



But if we want to look at these ideas, we need to have some information about the interaction energy that we see. The covalent bonds are very strong, then we have the ionic, the hydrogen bond, the hydrophobic and the van der Waals interactions.

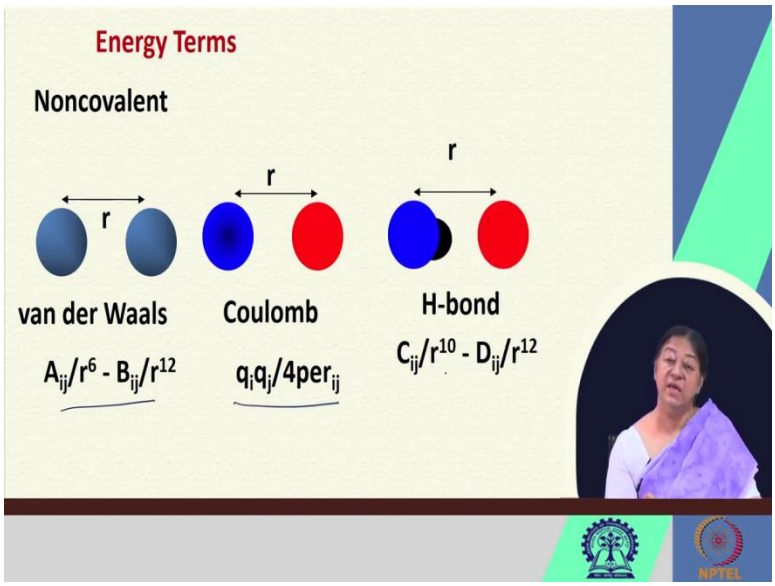
(Refer Slide Time: 21:04)



What are the contributions of the energy? We have covalent bonds, there is stretching also. What are these covalent interactions? If we look at how the atoms in the protein chain, the polypeptide chain are going to come together and we want to look at some energetics of the overall protein structure, there are contributions to the energy terms based on bonds, based on angles and based on torsion.

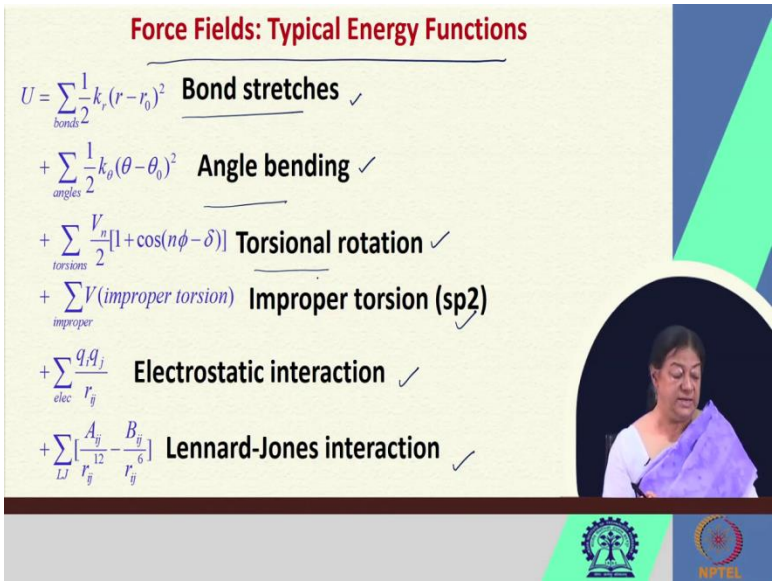
So, each of these in their own way are going to contribute to the overall energy in the covalent fashion.

(Refer Slide Time: 21:43)



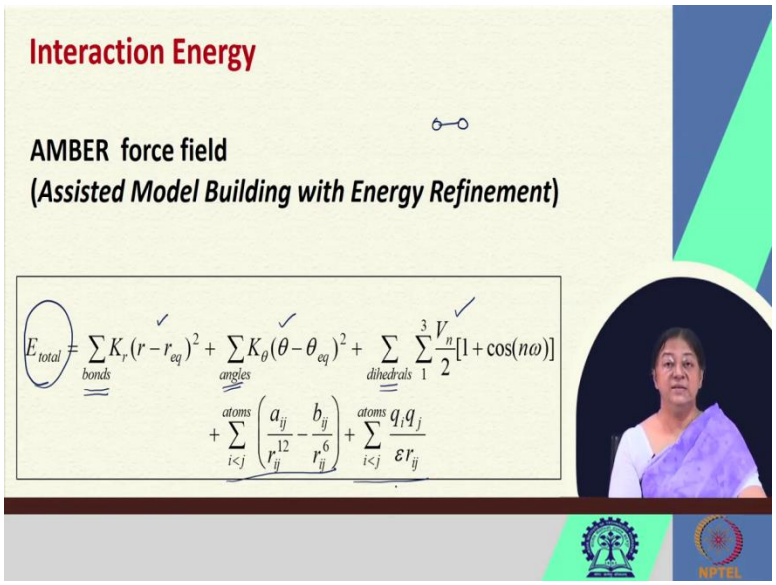
If we look for the non-covalent types, what do we have? We have the van der Waals type that follows a Lennard Jones potential, we have the coulomb type and then we have the hydrogen bond type. All of these are going to contribute to the overall energy function.

(Refer Slide Time: 21:56)



What is this? We have an overall energy function that is given by a specific force field. These this force field tells us that we have can have bond stretching, we can have angle bending, torsional rotation, improper torsion, electrostatic interactions and Lennard Jones interactions. All of this put together is going to give us the overall energy of our molecule. But you realize that when we are looking at all the possible bonds, all the possible angles, all the possible torsions and electrostatic interaction, this makes this calculation of the energy extremely computationally intensive.

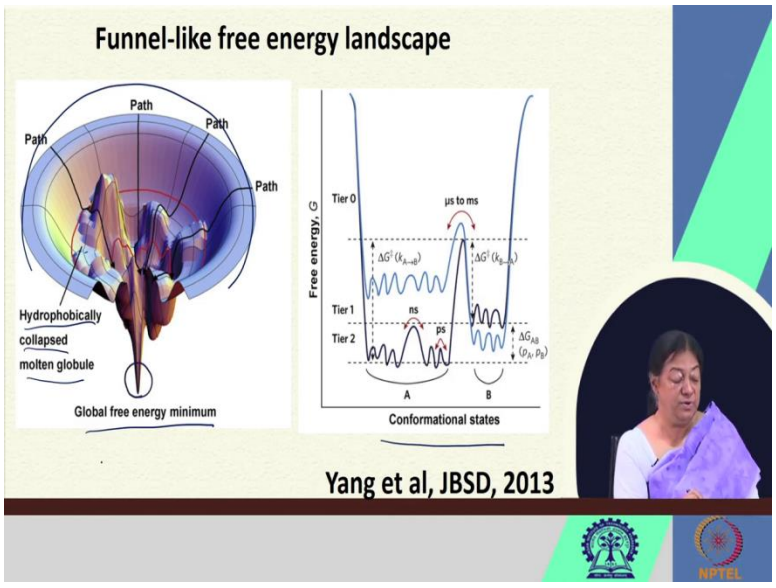
(Refer Slide Time: 22:46)



If we look at the amber force field that is used for proteins, we have the E total that we are calculating and what we learnt is, we have contributions from the covalent possibilities which are the bonds between every atom, the bonds between one atom to the other atoms, the linkages that we know about, the angles, and the dihedrals; these are the covalent possibilities. But if we look at all the van der Waals interactions and all the electrostatic interactions, there are umpteen possibilities where we can look for interactions between different parts of the molecule.

So the idea here is now we have to do an energy minimization. That is going to minimize this energy and tell us that this is a possibility for the global free energy minimum in the structure.

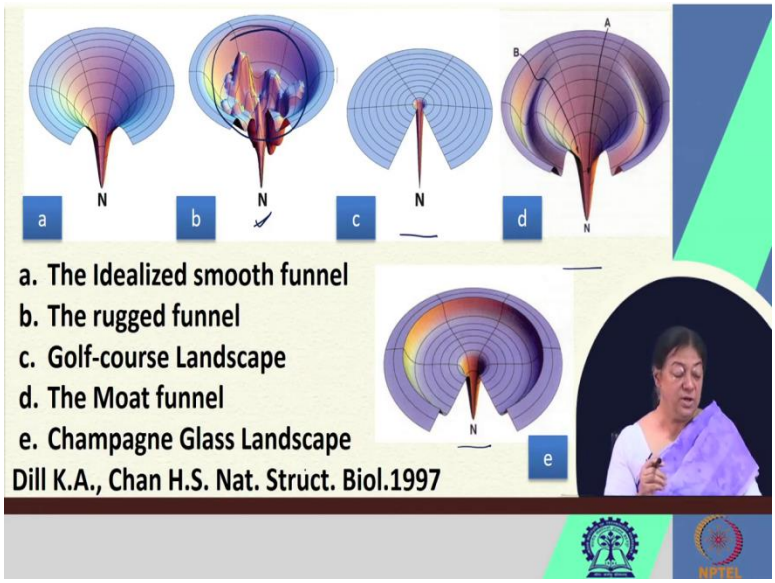
(Refer Slide Time: 23:39)



So we have all these possible structures. And then we have some possibility of a hydrophobically collapsed molten globule, where the hydrophobic parts come together. There are different

conformational states, different energetic possibilities, finally coming to a relatively stable state, that is going to be our global free energy minimum.

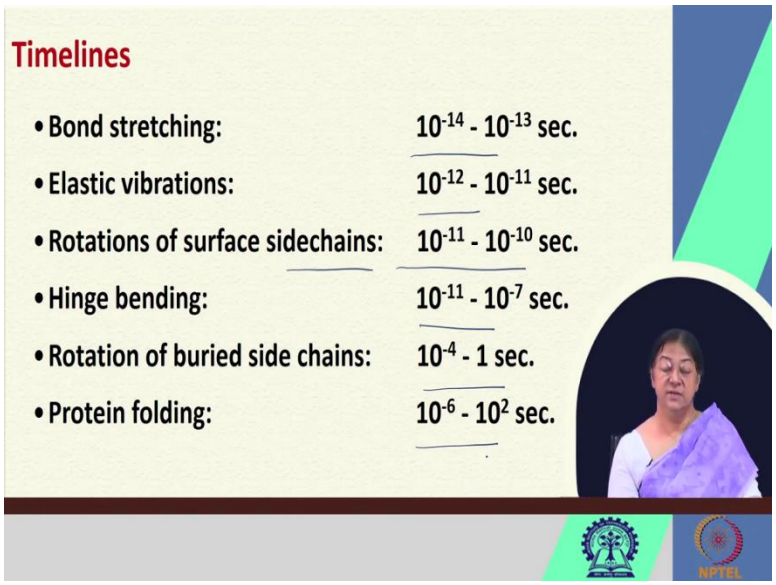
(Refer Slide Time: 24:14)



If we look at [refer to slide] the different types of folding funnels that are known, we have a smooth funnel, we have the rugged funnel. This funnel is what we see where we see different possibilities of the states. We have this golf course type the moat funnel and a champagne glass type funnel.

These are the different types of energetics that we can look at, when we are trying to calculate the overall energy minimum of our structure that we know of has these specific amino acids that are linked to each other in three-dimensional space.

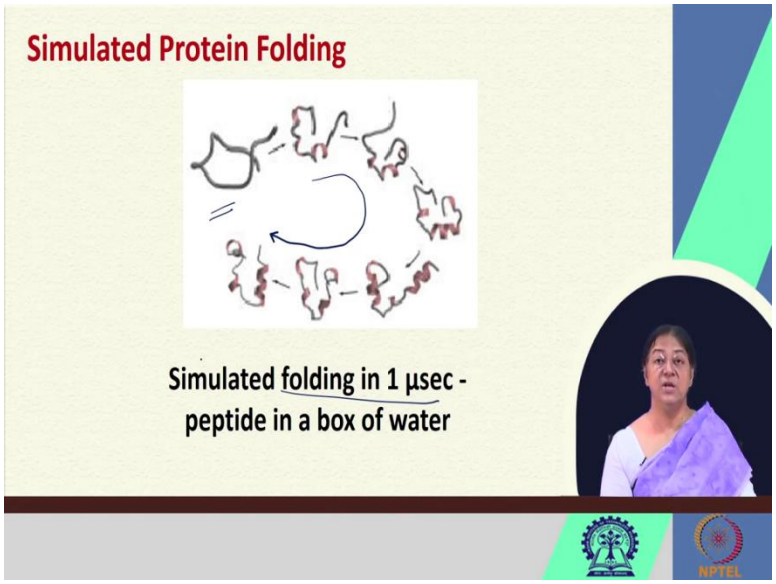
(Refer Slide Time: 24:38)



If we cannot get an experimental observation, we go for computer prediction methods. There are also molecular dynamics methods, where each of these atoms are given a velocity and through a time evolution we look at how the molecule folds.

This time evolution depends upon the specific timelines that we can look at. We look at the bond stretching that follows these less than picosecond types, the vibrations that are the atoms interacting with each other; in the sense the bond stretching that we observe. Then the rotation of the surface side chains, which were the different rotamers that are possible, because we have to look at the possible rotations for favorable interactions. There is hinge bending, rotation of buried side chains and finally protein folding that is going to occur in this time frame.

(Refer Slide Time: 25:49)



So, if we look at the simulated protein folding where we have an unfolded structure that gradually folds into a folding in 1 micro second, in a peptide, in a box of water. The unfolded polypeptide chain is put in the box of water (these are usually called solvent models), where you have an implicit or an explicit solvent module. We will not talk in details here, as it is beyond the scope of this course.

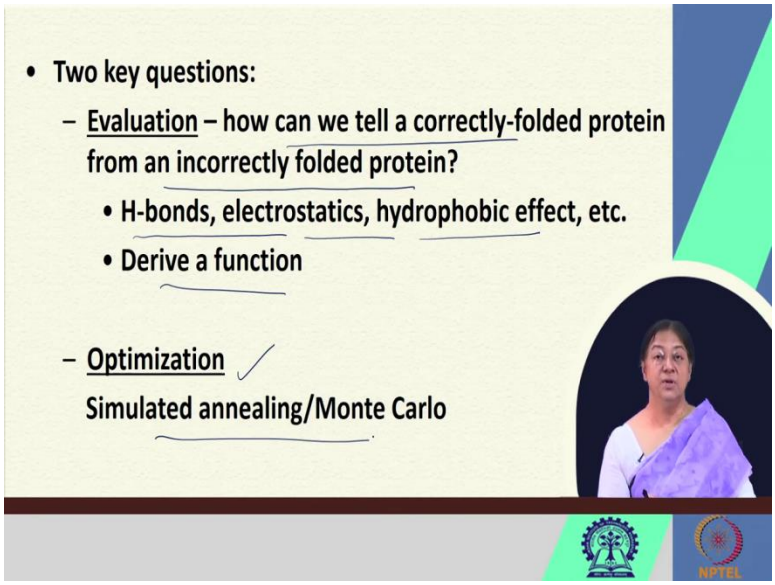
But, nevertheless when we are talking about specific arrangements for the atoms in three-dimensional space, we are looking for methodologies in a type where we have these conformational possibilities and we are gradually trying to get to a folded structure.

So the different methods that we spoke about were, whether we are looking at homology modeling, where we are saying that there may be a similar structure, and we look for the structurally conserved regions and then map the rest of the linkers and so on and so forth, to give us our model.

But again, we look for an energy minimized structure, depending on a specific force field where we are looking at covalent interactions, non-covalent possibilities and trying to get an overall energy value. When we are looking at a simulation like this, we are looking at a dynamic situation, where we are saying we assign velocities for each of these atoms and then we allow to evolve over time where we give some constraints, we know there are some energetics involved and we finally, get to a folded structure.

The whole idea is that we have our polypeptide chain, we want to get to a final folded structure, because that is what the aim is here.

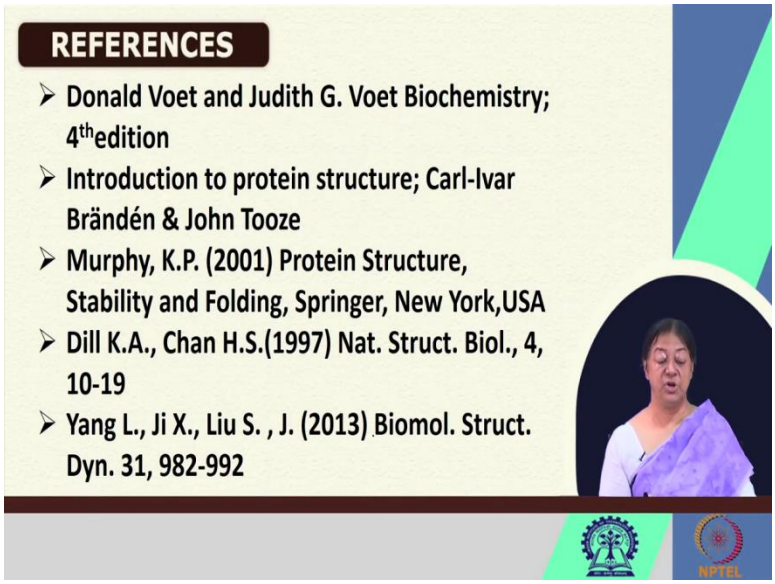
(Refer Slide Time: 27:48)



Because that structure and the stability associated with the structure is going to give us finally the answers to our two key questions. The evaluation is, how can we tell between a correctly folded structure from an incorrectly folded protein. So we have all these hydrogen bonds, electrostatics, hydrophobic effect.

We derive a function, we do optimization, all the simulating annealing and the Monte-Carlo methods that are going to tell us that we have a folded structure, that is going to help us in determination of the function of further characteristics that we would like to know about the protein structure.

(Refer Slide Time: 28:23)



These are the references [refer to slide].

Thank you.