

Physics of Biological Systems
Prof. Mithun Mitra
Department of Physics
Indian Institute of Technology, Bombay

Lecture – 37
Hydrophobic Polar Protein model

(Refer Slide Time: 00:16)



THE PROTEIN FOLDING PROBLEM

SEQUENCE → **Structure**

Structure → **Function**

How does a protein sequence correspond to a particular structure?

A random search in structure space is IMPOSSIBLE!



This is a very difficult problem, what we will do is a, is one of the most simplified models that is out there but, even that should give some sort of insight into what is required for this course ok. So, the basic paradigm of this protein folding is that, you have an amino acid sequence, and that sequence dictates how the protein will fold.

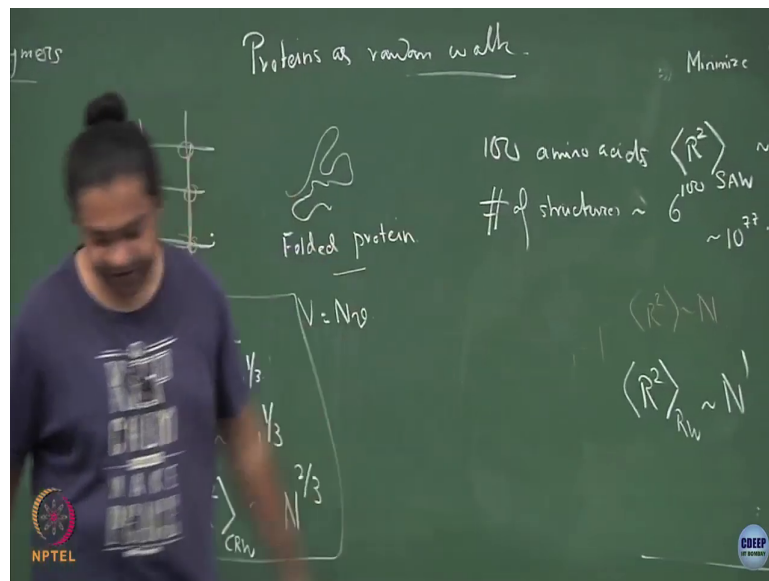
So, that sequence dictates your structure and the structure of the protein dictates your function. So, precisely how you fold dictates how you what function you perform. If you do not fold properly, their miss folding errors can cause a number of diseases ok, for a given; for

a given set of conditions. But, in the sense that, if you change the pH a little bit like these or if you do this post translational modifications like we saw right.

So, you will go to a slightly different native states. So, you can go, you can have a discrete set of conformations, but given a certain condition whether its evacuated or whatever post translation modification is done. So, given a set of conditions, there is an unique ground state and the basic question that we want to answer is that, given a sort of sequence how do I go from that sequence information to a particular structure ok, and you might say that well I will do a sort of random search in whatever possible structures I can obtain.

But, a quick sort of back of the envelope calculation will show you that random search is going to be pretty well impossible. For example, let us say that, let me work with this compact random walk model in 3 dimensions, and let us say I have a protein which has some 100 amino acids; a protein which has some 100 amino acids right.

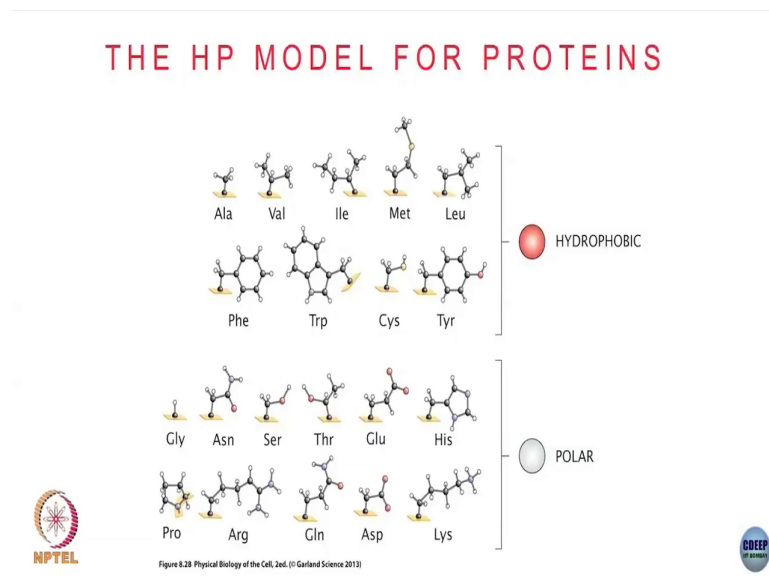
(Refer Slide Time: 02:09)



The number of possible structures; the number of possible structures given a compact random walk like this, would go roughly as some 6 to the power of 100 right. At each step, you have 6 possible links this is an you have a 100 such amino acids. So, that would typically be the number of structures which is roughly some 10 to the power of 77.

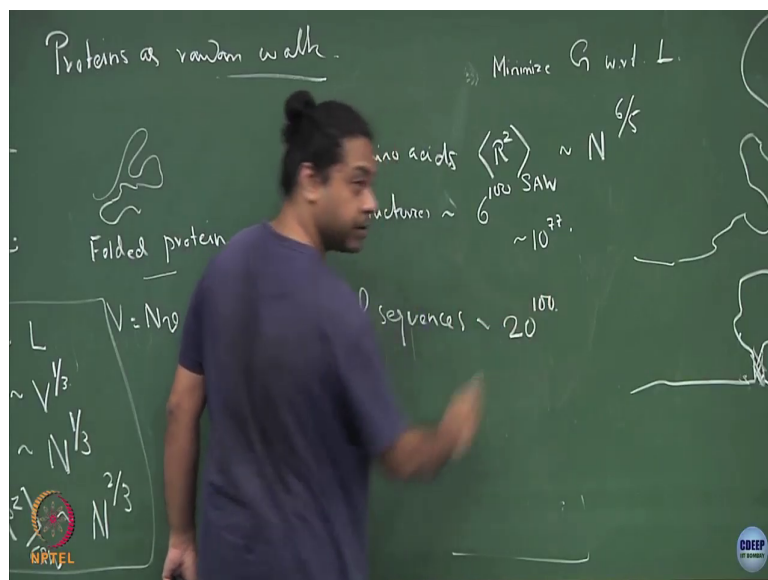
And, then regardless of how fast you search, even if you were to say that every structural search in a femto second and so on; if you wanted to scan this entire space of structures that would roughly take you more than much more than the age of the universities ok. So, it cannot be a folding process, there has to be some sort of principle behind this folding, and we will try to look at it in a very restricted sense.

(Refer Slide Time: 02:56)



So, what this so, what we will do is that, we will get rid of this and remember the number of sequences for a 100 amino acid like this is.

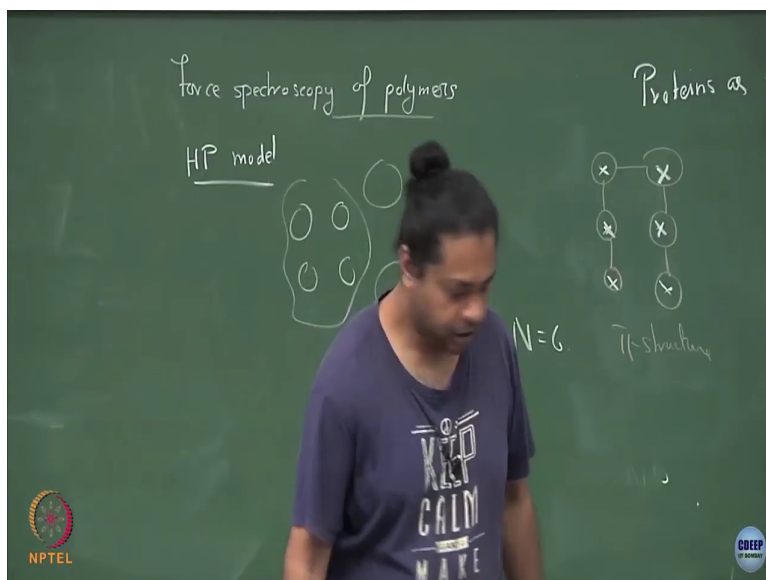
(Refer Slide Time: 03:14)



So, the number of sequences is roughly like 20 to the power of 100 right. You can have 20 possible amino acids. So, for a 100 amino acid protein, the sequence space is like 20 to the power of 100, which is even more than the structure space. So, the first step to do in order to reduce the complexity and starting to think about this problem; what these people did way back in the 1980s and so on, (Refer Time: 03:43) and others was to come up with this way of reducing this complexity. So, what they said is that, yes, I have 20 different amino acids.

But, let me not worry about that, let me say that I know that hydrophobic forces are a very important driver of folding right. We argued that when you have hydrophobic residues, they would want to sort of cluster together and leave the hydrophilic residues on the outside ok.

(Refer Slide Time: 04:09)



So, let me say that, that is one of the important drivers of this protein folding process, and then let me use that in order to reduce the complexity of this sequence space from 20 to 2. So, I will group all amino acids into 2 classes; either they are hydrophobic or they are polar. So, the hydrophobic ones do not like to interact with water, the polar ones like to interact with water.

So, let me see that if I take a model like this, how far can I go in trying to sort of understand how to think about this folding problem. So, this class of models and their derivatives are called the HP models; Hydrophobic Polar models. There is whole sort of literature on this hydrophobic polar model dating for 20-30 years now. But, what we will do is a very simple one.

(Refer Slide Time: 05:13)

A TOY MODEL FOR AN HP PROTEIN

Consider a 6-amino acid protein on a 3x2 lattice

How many possible sequences can we have? $2^6 = 64$

How many possible structures can we have? 3

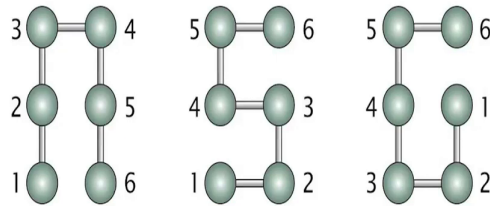


Figure 8.30a Physical Biology of the Cell, 2ed. (© Garland Science 2013)



So let me say, that I consider a 6 amino acid protein ok; so, very very small amino acid. It is a very very small protein, which is not really when a protein itself on a 3 cross 2 lattice. So, I will instead of doing 3D I will do a 2D. So, let me say, this is my lattice and I take a protein which has N equal to 6; and I will work within this HP model ok. What is the number of possible sequences in this sort of a paradigm?

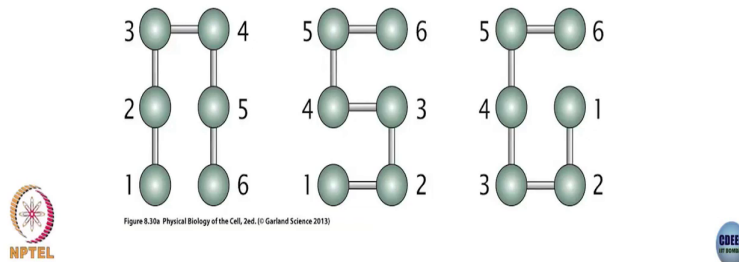
(Refer Slide Time: 05:56)

A TOY MODEL FOR AN HP PROTEIN

Consider a 6-amino acid protein on a 3x2 lattice

How many possible sequences can we have? $2^6 = 64$

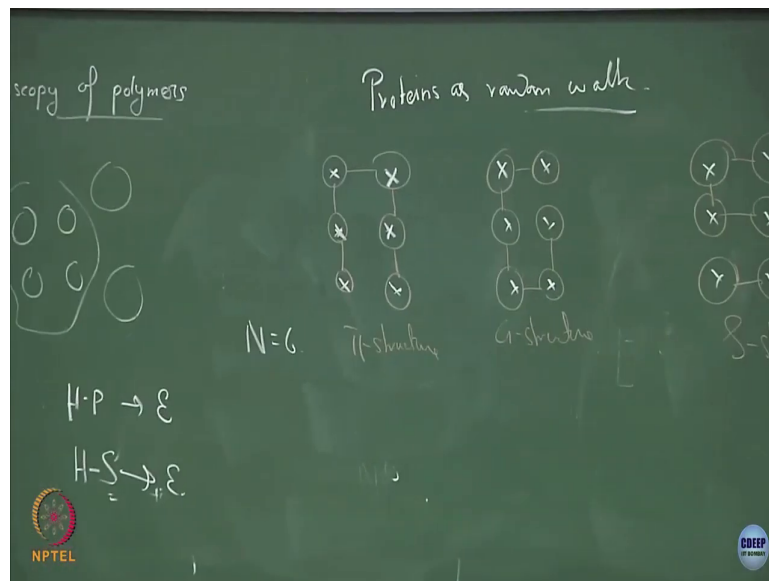
How many possible structures can we have? 3



It is 2 to the power of 6 right. So, how many possible sequences we can have that is, 2 to the power 6 that is 64 ok. So, these are the possible number of sequences that I can have for this very simple, very short protein. What are the number of possible structures that I can have, any one? Given this 3 cross 2 lattice what are the possible structures that I can draw? Possible unique structures that I can draw ?

Remember, this is a compact random walk; which means I have to occupy every lattice site once. So, what are the possible structures that I could have? So, let me draw one. Let us say I draw a protein like this ok, recall this is the pi structure. Say here is the 6 amino acid protein; I filled up all the available lattice sites, I have got a structure like this.

(Refer Slide Time: 07:08)



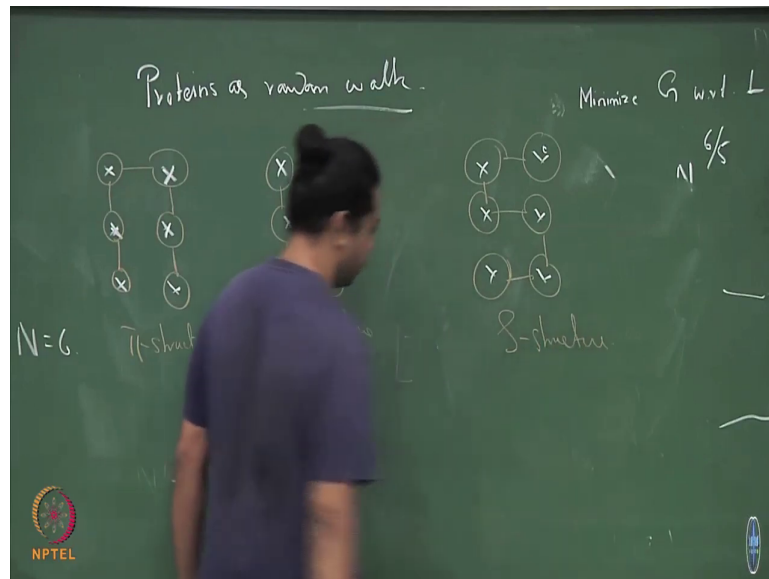
What other structure can I have? So, again I can have a.

Student: (Refer Time: 07:16).

A G right; so, I can have a G; something like this good. So, I will ok; so, I want structures which are not related by translations or rotations or reflections. So, I want unique structures ok. What else? I could have an S right. So, I could have an S structure, and if you think about it hopefully you can convince yourself that these are the only 3 structures that are possible for this sort of a 3 cross 2 lattice; here is my pi, here is my S, here is my G.

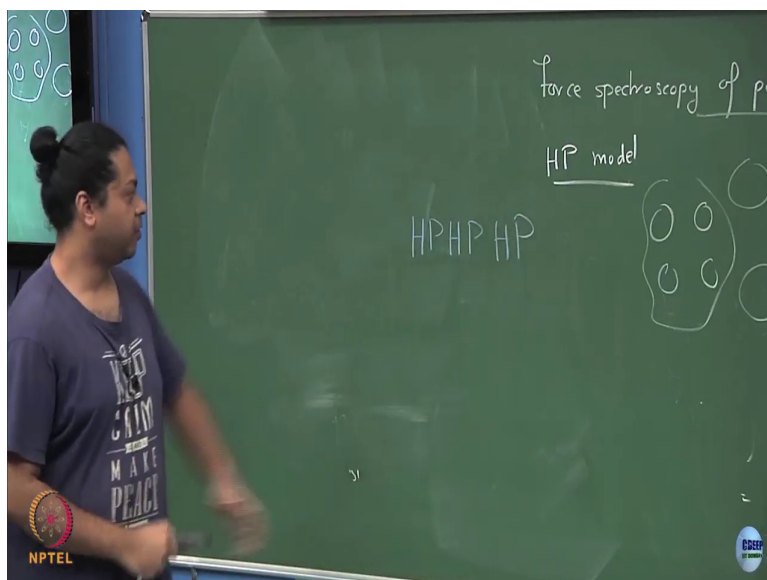
So, a sequence space of around 64, but I have a structure space of around 3 at least for this very simple case ok; so, these 3 possible structures. Now, let me take a particular sequence.

(Refer Slide Time: 08:25)



Let me draw the third structure anyway 1, 2, 3, 4, 5, 6 and then so, this is my S.

(Refer Slide Time: 08:51)



Now, let me take a particular sequence. So, let me take a sequence which is HP HP HP ok. I have 1 hydrophobic, 1 polar, hydrophobic polar, hydrophobic polar; that is one of these possible 64 sequences that I could have, and the idea is that whenever you have a hydrophobic residue in contact with another hydrophobic residue, that is energetically advantageous right.

So, if I had a structure like this, for example: if I have a sequence like this HP HP HP and I put it let me say that, whenever I have bond between a hydrophobic and a polar residue, it cause me some energy epsilon. And, whenever I have a bond between a hydrophobic residue and a solvent, that also cause me energy epsilon, just the simplest case. You could take different energies and so on.

But let me just say that whenever a hydrophobic residue comes in contact with either a polar residue or a solvent molecule; what is solvent? Solvent is all the sites over here, say everything

is surrounded by solvent right. So, all of these are solvent points in all of these structures. So, whenever I have bonds like this, it costs me some energy epsilon.

So, if I take this sequence and I put it onto these 3 structures, could you tell me what is the energy of these 3 structures given this particular sequence? So, I place HP HP HP HP HP HP and then so on. Actually, I think I have the figure in the next slide. So, there is a sequence that I want here is that HP HP HP sequence laid out on these 3 possible structures.

And in fact, even the bonds are drawn so, we might as well count ok. So, this is the pi structure. Here is a hydrophobic residue that has a bond with this polar residue over here and 2 solvents over here. Similarly, this hydrophobic has bonds with 2 polar residues over here. This hydrophobic has bond with 1 polar and 1 solvent over here, sorry this is solvent. So, the total number of energy causing bonds is 1, 2, 3, 4, 5, 6, 7 right.

So, this structure cost in energy 7 epsilon. If you count each of these others 1, 2, 3, 4, 5, 6, 7; this also cause 7 epsilon; this one also 1, 2, 3, 4, 5, 6, 7 right. So, each of these structures have the same energy for this particular sequence of bus. And therefore, in this very toy like model, I would claim that this is actually not a very, this is not a protein like sequence at all in that there is no unique ground state of this sequence.

Again within this HP model, within this compact random walk model, a sequence each and the 6 amino acid protein, this HP HP HP sequence does not have a unique ground state; which means it is not a very protein like sequence ok. On the other hand, I can choose a different sequence let us say, PHP PHP; there is another of those 64 sequences I could choose. And, then again calculate which of these structures would be optimal or if there is at all an optimal structure for this sequence PHP PHP.

So, here is the spy structure, see the advantage here is that, I have placed 2 hydrophobic next to each other which means that, I now have only 2 energy costing bonds with the solvent here and here; in this S versus this G, I have 1, 2, 3, 4 here; 1, 2, 3, 4 here. So, there is a clear winner, this spy structure is the most stable given this sequence. This is an energy 2 epsilon, the others of an energy of 4 epsilon. Therefore, this has an unique ground state, then I might in

this toy models sort of claim that a sequence like this corresponds to some a protein like sequence in that if I let it fold, it will fold into this unique pi structure.

You can do this for all these sort of sequences this 64 you could ask that well; which structure out of these 3 is the most common or the most designable in that sense. So, before we do that, you can also look at what is the probability of this state versus probability of this state given that you have these 3 possible states and that is just e to the power of minus beta $2 e$ by the partition function and that will look something like this.

So, the probability of it folding into the correct conformation as a function of epsilon or as a function of sorry the function of temperature in that sense, let us epsilon is whatever property, it is a function of temperature would look something like this. And, you would see similar curves if you did real protein folding experiments which is not to say that this model is correct, but at least it has some grain of truth in it ok.

Now, going back to this question, if I ask if I look at all these possible sequences the 64 possible sequences and I ask that which structure was the most common, you would find that this pi structure has the largest number of sequences that fold into this. So, there are 9 sequences out of this 64 that will fold into this pi, that has this pi structure as the unit ground state. There are 6 that has this S structure is the unique ground state and there are three that has this G structure is the unique ground state.

The rest do not have a unique ground state at all, like that HP HP HP these sort of things and you can sort of construct it by saying that for example, you can come so, here the sequence is PH PH or P does not matter and then H again and then H. So, that is how to read that sequence. You can sort of constructed by thinking about what pairs are there in each structure that are not there in the others for example, here I have this 2, 5, which do not come into contact in either this S or in this G.

So, what I want to do is that, I want to place hydrophobic residues on this 2 and 5. So, that is what I do, I place hydrophobic residues on this 2 and 5. And then, I play with the residues at these other locations in order to get the positive sequence. Similarly here, I would place

something in 3 and 6 over here is so, I placed hydrophobic residues in 3 and 6 and then play with the others.

Does this have, does this sort of in a this very toy like, very handy, even approach does it have any value? What people have done is that, they have designed sequences for example, it is known that a common secondary structure in proteins is alpha helices right. Alpha helices and beta sheets; and in alpha helices, it is known that you have hydrophobic residues roughly around every 3 or 4, after every 3 or 4 residues.

So, if I design a structure like that; so, I have my library of which amino acids fall under H, and which amino acids fall under polar, and I design a sequence completely a priori such that I have a hydrophobic residue for every 3 or 4 base pairs. And I could ask that, whether that does fold into a alpha helices or not.

You people have done these experiments and they have shown; so for example, here is a purely constructed sequence, you could replace an H by any of these hydrophobic residues. You could replace a P by any of these polar residues and people have shown that often when you construct a sequence like this, they will indeed fold into alpha helices. Not only that they are also sort of functional in that some of these design sequences have shown enzymatic activity as well ok.

So, again to reiterate this is extremely toy model have reduced the complexity of the problem a lot, but even so, what this says is that it is indeed true that these hydrophobic polar force, these hydrophobic forces which try to shield hydrophobic residues from contact with solvent are indeed a very important determiner of this ultimate folded structure of the protein. This is not the full answer but, at least it should form an important ingredient in whatever the full answer is which is why these HP models have met with a reasonable degree of success, for a well chosen definition of success ok.

I think I will stop a little early today, I do not think I have much more to say I say much more and biopolymers, but I think we will move on to a new topic of crowding and how crowding effects biological assemblies and reactions so on, from the next class ok. I can upload some,

there are very nice papers on this HP model and up to how far you can push a model like this.
If you are interested, I can upload on modules some of the papers on this HP sort of model ok.

I will see you on Friday.