

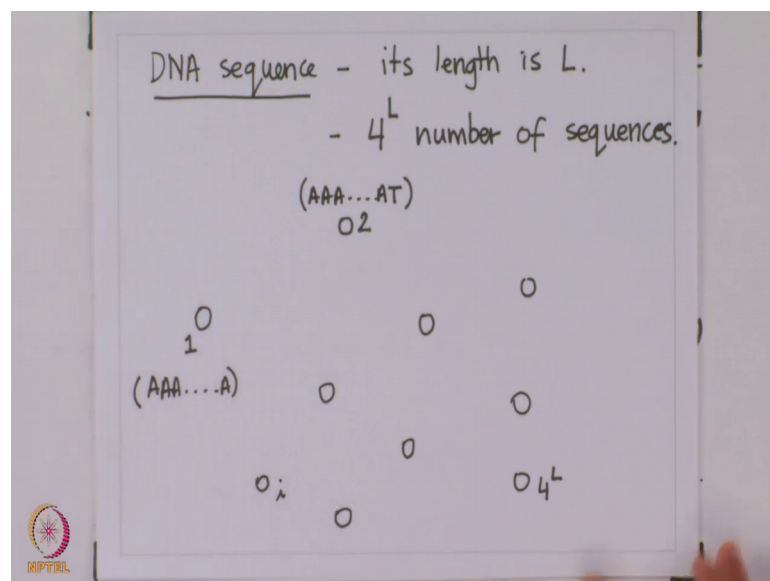
Introduction to Evolutionary Dynamics
Prof. Supreet Saini
Department of Chemical Engineering
Indian Institute of Technology, Bombay

Lecture – 13
Sequence Space as Networks

Hi, and let us continue our discussion on sequence spaces and fitness landscapes. So, what we had done last time was introduced the fact that given a sequence A DNA sequence of length L there are in all of 4 to the power L number of sequences which have the same length l , but different nucleotide sequence. Similarly if you are talking of a protein which has 100 of amino acids in it there are 20 to the power 100 polypeptide chains each one of them equal to length 100 and length here refers to the number of amino acids which go towards forming the polypeptide chain.

But the constituents of that 100 amino acid chain are different in each one of the 20 to the power 100 amino acid changes and we had stopped our discussion by representing we had just started towards representing these sequences 4 to the power L sequences or 20 to the power L sequences of amino acids in a graphical sense. So, let us continue our discussion towards that, and we will come to the definition of what is referred to as the fitness landscape. So, we will start our discussion by the example of DNA.

(Refer Slide Time: 01:39)

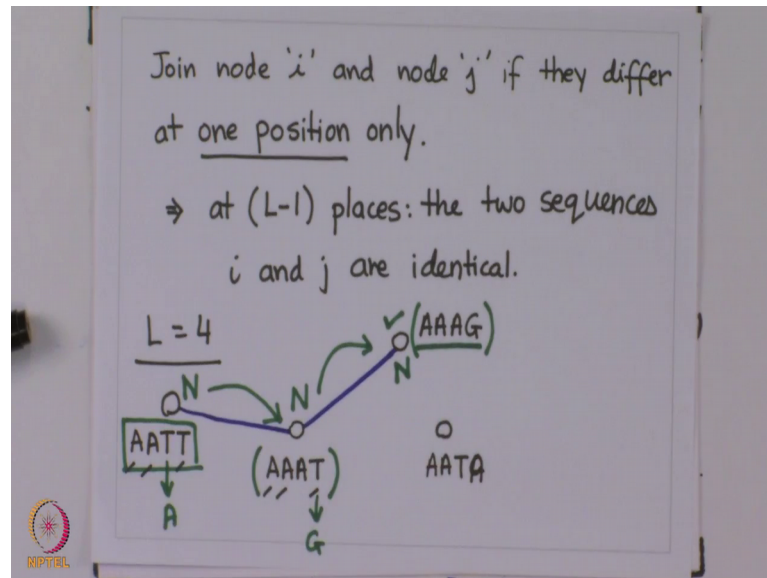


So, we have a DNA sequence, and its length is L . And we have already done that if we were to define a sequence of length L there are 4^L number of sequences.

Now, what we are going to do with these 4^L sequences is think of this as a 2 dimensional plane, and do a graphical representation of these 4^L sequences where each node represents one of the 4^L sequences that we are talking about. So, let us say node number one represents that sequence of length L which is all As this is node number 2 this is that sequence which is all a s, but the last nucleotide is T and so on and so forth. We keep doing this till we populate this plane with these nodes and the total number of nodes in this plane will; obviously, be equal to 4^L because every distinct genotype or the DNA sequence that we are talking about will be represented as a node here.

So, at the end of the whole exercise we will have 4^L nodes on this plane of paper. Let us draw some of them, let say this is the 4^L nodes that we are interested in, and this is the last one which is node number 4^L . This is node number i and so on and so forth. So, we have populated this plane with these nodes. Now what we are going to do is we are going to join node i and j when we write that down on a separate sheet of paper the rules are what we want to define a structure and we want to define the rules about how we are going to go towards defining that structure. The first exercise that we have to do towards explaining fitness landscape which is a structure we are interested in represent all possible DNA sequences in this case where we are talking of DNA in the case of proteins, it will be a similar way similar approach towards defining that fitness landscape.

(Refer Slide Time: 04:29)



So, in DNA sequences first we represent all possible DNA sequences and each sequence is represented by a single node. So, we do that and we have all 4 to the power L nodes here. Next what we are going to do is join node i and node j . So, this represents node i represents the DNA sequence corresponding to that particular node. Because remember each node corresponds to a distinct DNA sequence e which is of length k . Now we are going to join node i and node j if they differ at one position only which implies that at L minus 1 place the 2 sequences i and j are identical. Let me make that clear again. Let say if L is equal to 4; that means, I am only interested in understanding sequences of length 4, that is all I am interested in then the total number of sequences that I am going to end up with is 4 to the power 4, which is 16 into 16 which is 256.

So, I will have 256 nodes drawn on the plane of my paper. Now each of these seq nodes represents a particular sequence and I join 2 nodes if they differ at one nucleotide position only, and the remaining 3 are the same. For instance if this node refers to AAAT and this one refers to AAAG; that means, the first 3 nucleotide positions here and the first 3 here are the same, these 2 are distinct. So, these 2 sequences sequence i and sequence j only differ at one nucleotide position which is one position only. And at the remaining L minus 1, L which is 3 here the 2 sequences are identical and if that is the case then these 2 sequences are joined by a node. That is; take another example if you have a sequence called AATA which is represented by this node.

Now, what we want to do here is say that the question is whether these 2 nodes should be connected or not in one sense these are identical in the sense that they have 3 A's and 1 T. This one also has 3 A's and 1 T, but these will not be connected because we are interested in the specific order of the nucleotides that we are talking about in a particular sequence here. Of the 2 sequences that we are looking at right now, they are only identical at to nucleotide positions and the remaining 2 nucleotide positions are different in them. If we again look at them carefully these 2 are identical at the first 2 positions A A and A A after that the third this sequence has an a this sequence has a T that is a difference this sequence has of T th 4th position this one has A A th 4th position.

Hence these 2 differ at not 1 not 0, but 2 positions and they are only identical at L minus 2 places which is 2 hence we will not join these 2 nodes with the h here. You could also have differences in the middle which is you could have AATT this is the node that corresponds to AATT. And now if we are talking about these 2 sequences they are identical at first second and 4th position. However, they differ at third positions which only satisfy; which again satisfies this rule that they only differ at one position only, the other 3 are identical and hence these 2 nodes will be connected by an H and so on and so forth. You keep building that for all 4 to the power L nodes that you have in the original scenario that we started with, and if L is the genome of an organism we are talking of very large number of nodes.

So, again we connect 2 nodes if they only differ by why one nucleotide position. What is the significance of connecting them if they have only differ by 1 nucleotide position? We want to do that because via this linkage this gives us an example or an idea about how a sequence might jump from one particular sequence to another particular sequence. For instance, if I have a population such that it is genotype is this AATT. All n individuals belong to this particular genotype. Now what might happen is that one of these individuals might acquire a mutation, and that mutation is going to be a single nucleotide which is called up point mutation. And during that mutation this T might get changed to an A which means the mutant individual has jumped from this sequence to this sequence on my graph.

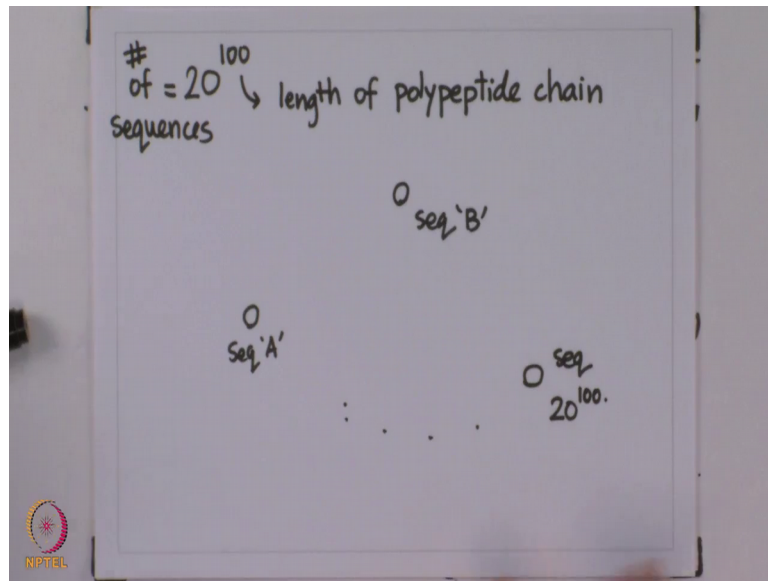
Now, depending on what is the growth rate whether this beneficial help the individual or not let say imagine that the beneficial that had happened was beneficial in nature that the mutation that happened was beneficial in nature in and that resulted in this mutant

growing faster than the rest of the population, than that mutant starts we go back to the selection and the mutation rules that we have just studied when we can easily see that should that mutation have been beneficial in nature. The mutant that has a reason at this point will start eliminating the rest of the population and eventually the population moves to this point.

And now again that same story repeats itself suppose we have a population of size n , occupying this genotypic space all n have the same genotype and now should a mutation happen where this A gets changed to T, we have one individual by which has come back to the original position original genotype which was there several generations back, but now we know that the fitness of this genotype is more than fitness of this genotype. And hence, that mutant individual is no longer going to be able to outcompete the parent genotype from it from which it mutated from and hence this genotype survives, but what happens if now this T changes to a G now what; that means, is we have a mutant which is of this genotype. And if this genotype grows faster than this genotype then the mutant which has come at this place is going to be able to outcompete everybody here and then the entire population moves towards this node.

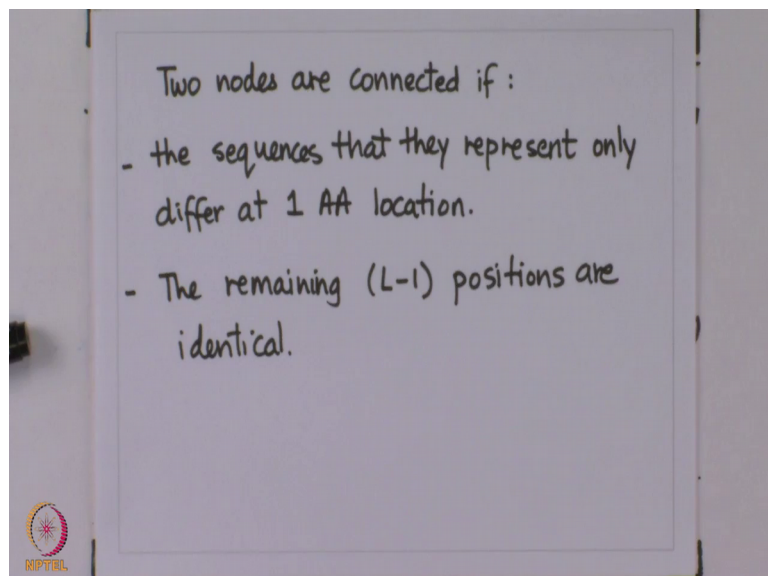
So, you have movement of populations taking place on this graph where each sequence is represented by a node. And nodes are connected if you can go from one sequence to another via single point mutation. In our case node this this and this node are not connected because you cannot have a single point mutation occur to this sequence and automatically reach this this sequence. You need to go through an intermediate sequence and hence approach that that is the first thing that we want to understand while we go towards understanding fitness landscapes and the concept is similar for proteins.

(Refer Slide Time: 12:48)



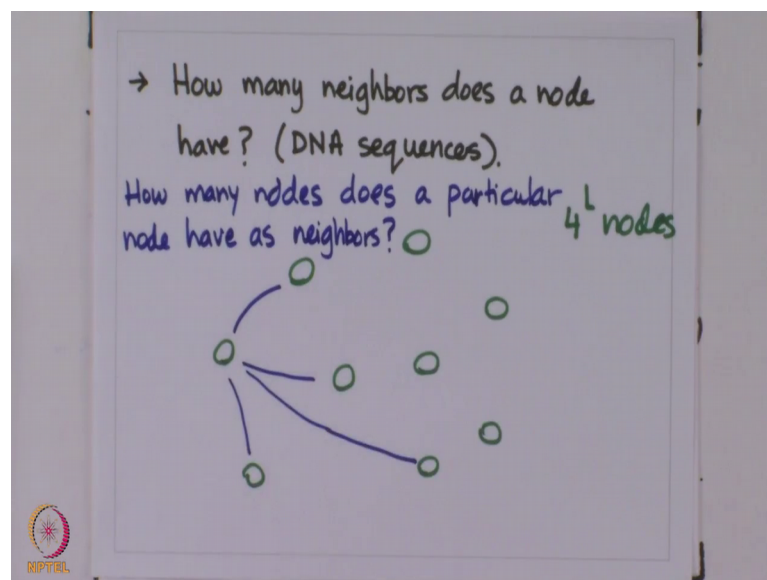
And now for proteins we are again going to look the example that we had started last time we have a protein of length 100. So, the total number of nodes which is going to happen is 20 to the power 100 where 100 is the length of polypeptide chain. And each one of these this is number of sequences. And each one of these sequences is going to be represented by a node. This is sequence A this is sequence B and so on and so forth going all the way up to sequence number 20 to the power 100.

(Refer Slide Time: 13:46)



And the rule for connecting nodes remains the same as the one in DNA that when we are talking of protein sequence spaces 2 nodes are connected if the sequences that they represent only differ at one amino acid at a particular position, at one only different one amino acid location. If the length of this polypeptide chain is L then the remaining L minus 1 positions are identical in the 2 sequences, to the point of this whole exercise being that to try and understand that whether you are dealing with DNA sequences or polypeptide chains what you can represent them as these nodes and edges where each node represents a sequence, and 2 nodes are connected by an edge if via a single mutation in nucleotide or amino acid you can jump from one node to the other.

(Refer Slide Time: 15:34)

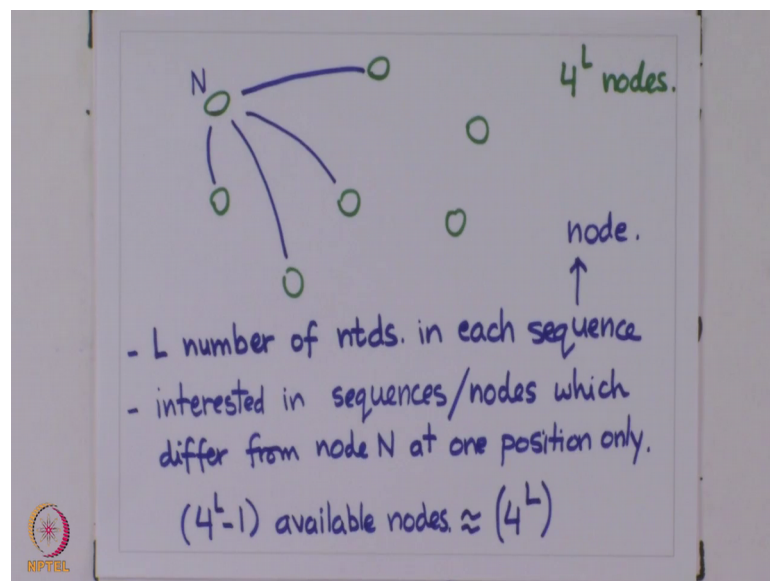


So, we have this. So, irrespective of whether we are talking DNA or proteins we have disconnected edges and nodes diagram now. So, the first question that I would like you to think about is that; how many neighbors does a node have? And that is first thing in terms of DNA sequences. So, the question is that we have this nodes and edges diagram we have a bunch of nodes 4 to the power L nodes, and not all nodes are going to be connected to every other node we have 4 to the power L nodes. And any node is going to be connected to a subset of these 4 to the power L nodes. A node is not going to be connected to every other node in this graph, but only going to be connected to a subset of them and the question is; how many neighbors does a node have or how many nodes does a particular node have as neighbors?

And, remember the rules that i and j are connected if they only differ at one nucleotide position. So, what we are interested in is that out of the 4 to the power L sequences, how many sequences can a particular sequence go to via acquisition of just one nucleotide position. So, so if I represent a node I present a particular DNA sequence of length L should a mutation happen a single point mutation happen in me how many of the other 4 to the power L sequences can I generate and hence I will be connected to each one of these other sequences, that I can tell myself into if I were to acquire one point new mutation in my sequence.

So, what I would like here is for you to pause the video for a few minutes try and work this out by yourself. We will work through the answer after that. So, the way to think about this problem is that it is re draw our example.

(Refer Slide Time: 18:14)



These are our 4 to the power L nodes this is 4 to the power L nodes and let say this is the node that I am interested in, and I am interested in computing the number of neighbors of this particular node. How do I find that out? 4 to the power L comes from the fact that there are L numbers of sequences L number of nucleotides in each of these sequences in each sequence and each sequence here is represented by a node.

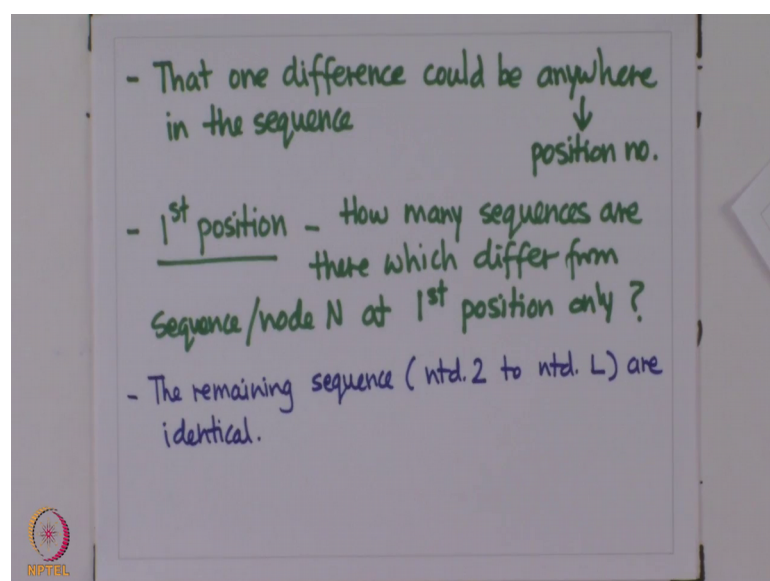
Now, if that is the case and I am only interested in talking about single point mutants, nodes are connected if they only differ from one another at one particular position and are and are identical at the other L minus 1 position. So, if that is the case that one

difference. So, we are interested in sequences, let us give the node that we are interested in some names of this is node n we are interested in sequences which are represented by nodes. So, nodes and sequences are going to be used interchangeably in this discussion. We are interested in sequences or nodes which differ from node n , the sequence corresponding to node n at one position only. Those are the nodes that we are interested in and of the total available nodes we have 4 to the power L minus 1 available nodes, because one of the nodes is the node and itself. So, we subtract that.

However because L is such a large number for all practical purposes this is just equal to 4 to the power L . So, we just are going to keep talking as if the total number of available nodes is 4 to the power L and not 4 to the power L minus 1 . So, that is what we are interested in we are interested in all those sequences or nodes which differ from the sequence defined by node n at one position only, and are identical at n minus 1 positions. So, how do we talk about this that difference that we have to talk about?

So, we are interested in sequences or nodes which differ from node n at one position only that difference could be anywhere in the sequence we have not specified that that difference could be at the first position and nucleotide number 2 to L could be identical that difference could be at the second position the first nucleotide could be identical and nucleotide number 3 to nucleotide number L could be identical and so on and so forth.

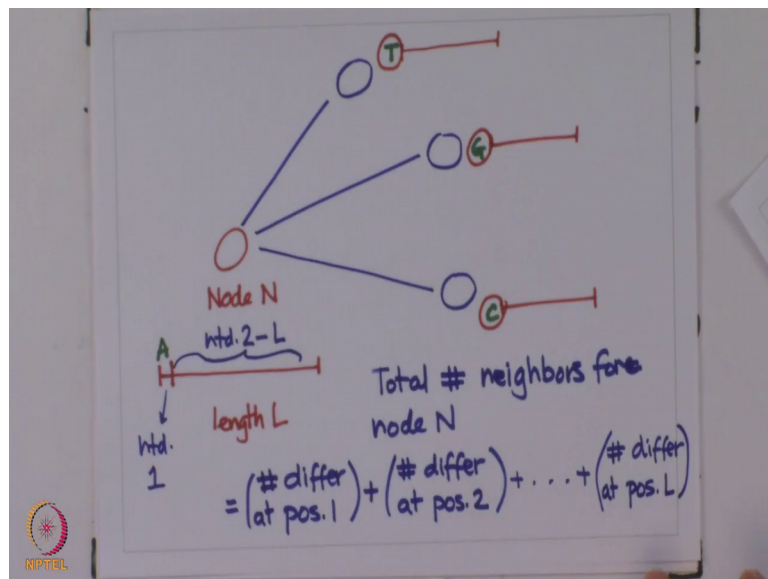
(Refer Slide Time: 21:40)



So, that one difference could be anywhere in the sequence. And anywhere by anywhere I mean the position number of the nucleotide. If that difference is happening at the first position, then my question is how many sequences are there which differ from sequence or node n at first position only. So, the difference between this node and this node is one nucleotide position, but that one difference could be anywhere that could be at position 1 that would be at position 2 going all the way up to position L, but it could only be at one position only. So, we are going to try and do this systematically and try and identify the number of nodes which are different from the sequence corresponding to node n at position 1 only. The remaining positions position 2 to position 2 to position L the remaining positions sequence from nucleotide number 2 to nucleotide number L are identical.

So, this is the question that we are interested in under these constraints take a 30 second pause from the video and try and answer this question. The way to think about this is I have my node n.

(Refer Slide Time: 24:01)



And it is sequence is of length L, and I am going to divide this into 2 where this is nucleotide number 1, the position of this nucleotide is 1 and this chain represents nucleotide 2 to L if that is the case then I am interested in how many neighbors does this node have which only differ at this nucleotide position and have the identical nucleotide sequence from nucleotide number 2 to L. And the answer to that would be 3. And the 3


will be because I am not free for the neighboring nodes if the difference is at nucleotide position 1 the neighboring nodes do not have any choice when it comes to nucleotide position 2 to L they have to take on the same sequence which node n that we are talking about has because this node and this node only differ at one nucleotide position.

However, there free to take a different nucleotide at position 1. So, imagine that the nucleotide at position 1 in my node n was a then one of the possible neighbors could be T at the first position and the same identical nucleotide 2 to L in at the for the remaining sequence or it could be a G or it could be a C. So, the number of neighbors that node n has which differ at only the first position of the sequence r equal to 3. And hence the total number of neighbors the total number of neighbors for node n will be number of neighbors which differ at position 1 plus number of neighbors which differ at position 2, which can be analyzed in an identical way as to what we have done just now and going all the way up to number of nodes which differ at position L.

(Refer Slide Time: 26:58)

$$\begin{aligned} &= 3 + 3 + \dots + 3 \\ &= 3L. \end{aligned}$$

$0 \rightarrow 3L \text{ nodes.}$



And these numbers are just equal to we have just seen that this number is equal to 3. Similarly, it can be shown that this number will also be equal to 3 and going all the way up to 3 for every position where the difference between node n and it is neighbors exist. And these are L number of 3 is added. So, the answer becomes 3 L. So, of the total 4 to the power L nodes and node is connected to 3 to the power L nodes. As an exercise you should try and do a similar exercise for a protein peptide change for a protein chain of for

a protein of length L , construct all possible polypeptide chains which have an equivalent length L and then try and decide for that of the 22 the bar L nodes that you have in your graph, how many nodes is one particular node connected to why are the same rules that we have been defining. So far we will continue this discussion in the next class onwards.

Thank you.