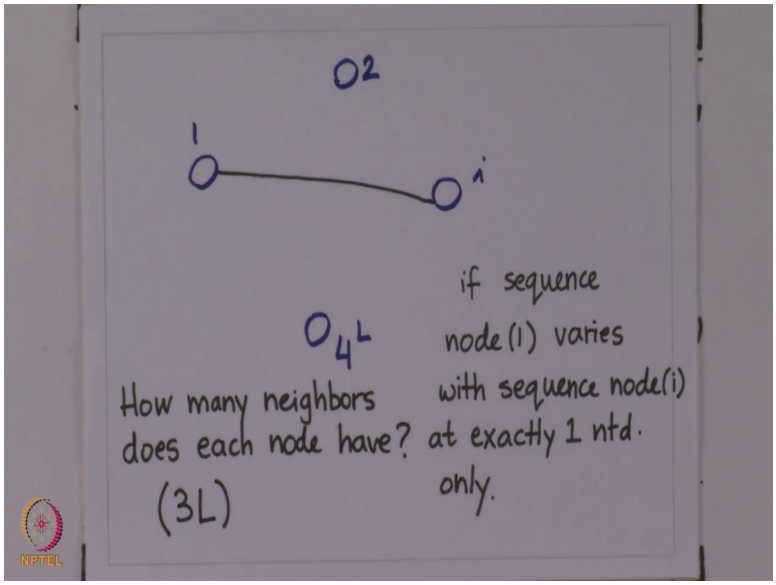
Introduction to Evolutionary Dynamics
Prof. Supreet Saini
Department of Chemical Engineering
Indian Institute of Technology, Bombay

Lecture – 14
Sequence Space to Fitness Landscape

Hi everyone. Let us continue our discussion from the last time where we were talking about sequence spaces and fitness landscapes. We will continue from where we had exactly left of last time we are considering a sequence space associated with DNA of length L and we had calculated that there will be 4 to the power L nodes each representing a particular sequence of length L, I am your interested in calculating the number of neighbors that each of these 4 to the power L nodes has and I never been defined as that sequence which varies from one particular sequence at one nucleotide only and it is identical at the other L minus 1 positions.

(Refer Slide Time: 01:03)

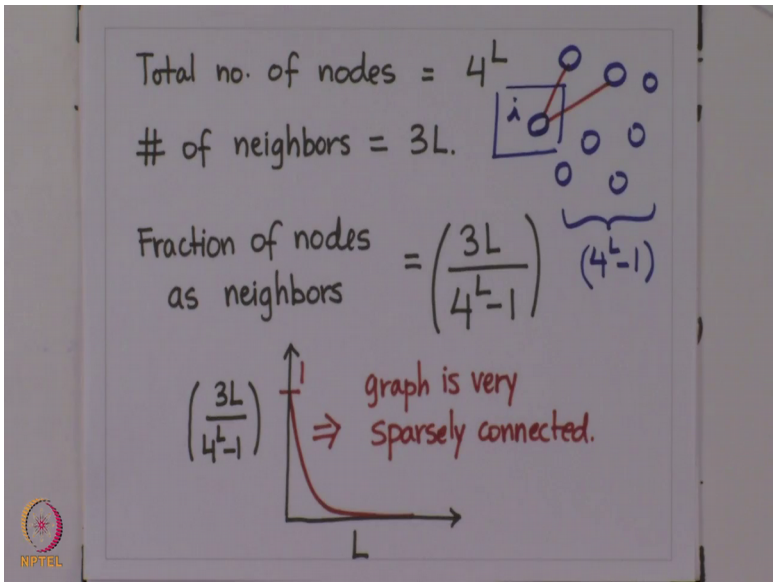


So, just to recap we want to represent my our sequences as nodes, this is node number 1, 2, node number i and eventually going all the way up to node number 4 to the power n. And we are connecting node 1 and node i, if sequence associated with node 1 varies with sequence associated with node i at exactly one nucleotide only. If that is the case that is when we are we are connecting node 1 and node i on all graph and this gives me connection between various nodes associated with this DNA sequence of length L and

then we had moved on to talk about how many nodes are how many neighbors does each node on this graph have and the answer to that was 3 to the power L because for every position in this DNA sequence of length L the variation between a node and it is neighbors could be at any one of those L positions.

If the variation between the neighbor and the node itself occurs is occurring at position number one of the sequence, then it could have 3 different nodes suppose the node that I am talking about has nucleotide A on the first position the node no node that varies on the first position only could have a t could have a g or could have a C on position number once hence sequences that very only at position number one or 3 in number similarly if I am talking about sequences that vary only at the ith position in this DNA sequence of length L, I am going to have 3 seq I am going to have 3 neighbors associated with that and when I sum them or sum all these neighbors up from nucleotide position number one to position number L, I get 3 times L and that gives me the relationship that every node on this graph has 3 L neighbors, right.

(Refer Slide Time: 03:45)



So, what the tells me is that the total number of nodes that corresponds to the total number of sequences which are possible of length L is 4 to the power L and to and each node has number of neighbors equal to 3 L. So, now, what I want you to think about is that how densely is this graph connected with each other given a number of nodes and edges how many nodes of the total number of nodes that are available in this graph what

fraction of nodes is every single node connected to. So, the total number of nodes is 4 to the power l, but every node is connected to 3 L other nodes; that means, the to the fraction of nodes as neighbors.

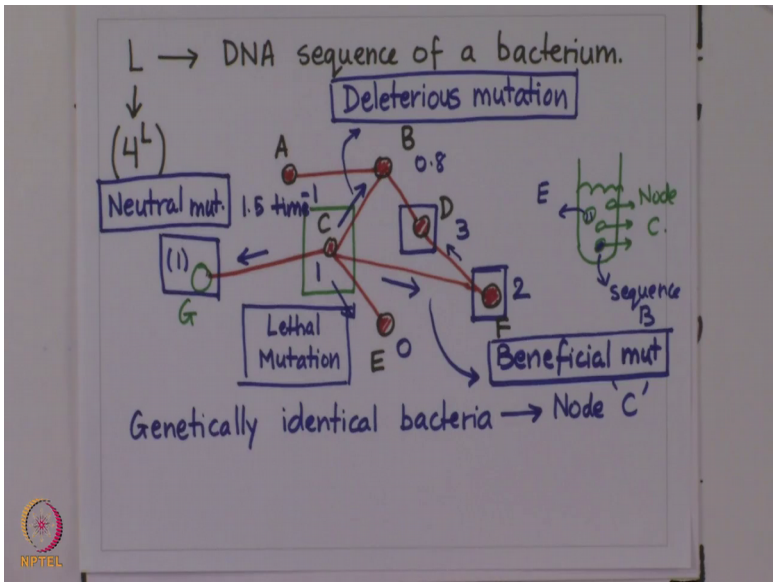
So, of all the nodes that exist in this graph what fraction of nodes are in neighbor of any particular node ith and that will be 3 L which is the actual number of neighbors divided by total number of nodes 4 to the power L and we can subtract 1, because one is that no itself. So, if I am what we are doing here is that if I am talking about node I there are lots of other nodes in this network and the total number of nodes if I leave aside the ith node the total number of nodes is 4 to the power L minus 1, because I am not counting that node itself and out of these 4 to the power L minus 1 this node is only connected to 3 L number of nodes.

So, the fraction of the nodes that every single node is connected to is given by 3 L divided by 4 to the power L minus 1. And now what I would like you to do is perhaps pause the video for maybe 30 seconds or so and then try and plot the graph and see what you what you expect this behavior to be as the sequence of the DNA that we are talking about increases. So, x axis represents increasing length of the DNA sequence that we are talking about. The y axis is what is the fraction of nodes the as neighbors that every single node in this graph has. So, on y axis I want you to plot 3 L divided by 4 to the power L minus 1. So, perhaps you can just pause the video for 30 seconds or and thing about this problem for 30 seconds or. So, and come up with the very qualitative estimate of what you would have what you would expect intuitively this relationship should look like all right. So, what we can see here is that when L is equal to 1; that means, we are only talking of DNA sequence of length 1 this is 3 divided by 3. So, this is one and very rapidly what you should expect and you can work out the numbers for L equal to 2 L equal to 3.

Some of the smaller DNA sequences are you can be worked out by hand what you would node is that it decreases very rapidly and approaches 0. So, what this tells us that our graph is very sparsely connected for any reasonable for any reasonable length l; that means, the total of the total number of nodes which are present in the sequence every single node is connected to a only a very small fraction of the nodes other than itself in the total number of graph. So, this graph the sequence space is associated with this DNA is a very sparsely connected graph all right. So, now, what we are going to do is imagine

that this DNA sequence, that we are talking about which is a sequence of length L corresponds to the DNA sequence of a organism and for the sake of simplicity let say bacteria.

(Refer Slide Time: 08:12)



So, this L corresponds to DNA sequence of bacterium; obviously, if length of the DNA polymer that we are talking about is l; that means, we are talking about a graph which is 4 to the power L nodes, may be let us just draw a small section of this graph and let say these are the these are 6 nodes on this graph let us number them call them A, B, C, D, E, F and again consistent with what we have been defining. So, for each one of these nodes represents a particular sequence. Now not all are going to be connected with all graphs. So, they are going to establish these connections and the connection would mean that you can go from one node to another by acquisition of a single point mutation.

And if that is the case let us define these relationships to be of this sort. Let this be the graph that we are talking about. So, you can see connections between some of the nodes, but not connection between all, but all possible connections do not exist on this graph. Because L is a DNA sequence of bacteria every sequence is going to have a corresponding growth rate associated with it, if I had a bacterium who sequence whose DNA sequence corresponded to that of the sequence which is at node 1 that bacterium would have a growth rate associated with it. Should there be a point mutation and the

sequence from a goes to sequence at B this bacterium which is carrying sequence B would have a growth rate associated with the sequence corresponding to that of node B.

So, let us write down the growth rates associated with each of these sequences and I am just going to makeup numbers. So, as to illustrate the characteristics associated with these fitness graphs, these are not representative examples, but I am just making up numbers of growth rates which could perhaps exist in a graph suggest this. So, let us as write down the growth rates associated with each particular sequence, that is each particular node that we have in our graph. So, let say sequence a grows at 1.5 and the units for the growth rate will be time inverse we not specifying units be could be may be 0.8 D is 3 F is 2 C is 1 and E is 0. So, these numbers represent the growth rates associated with the sequence that each node is representing.

Now, let us imagine that we have a sequence we start of our experiment with this with genetically identical bacteria. And all bacteria have sequence which corresponds to node C. So, what; that means, is I start of my experiment in a test tube and I add a few bacteria to this tube and all of them have the exact same sequence as node C. So, genetically all individuals are identical which means this is where I am starting my experiment from, now what is going to happen as this experiment place out in future is that these will keep on this bacteria will keep on dividing and eventually one of them will acquire point mutation, for the sake of simplicity for the time being we are only talking about point mutations we are not talking of deletions duplications etcetera or acquisition of DNA horizontally we are only talking of single point mutations.

One of these bacteria upon division is going to acquire point mutation the let see what happens what is going to happen here when that happens, because mutations can happen randomly and in this graph the 3 point mutations which are possible are C going to B C going to F and C going to E what happens if one of the bacterium in this culture acquires a mutation and goes to sequence E for instance. Now the growth rate associated with sequence E is 0, which means this mutation that has occurred here the result and bacteria does not grow at all it just it never divides which means that the mutation that has happened is what is called a lethal mutation; that means, the bacteria which is carrying this particular genotype would not be able to survive hence it is growth rate is equal to 0. So, one of this this mutant is not going to survive and it is going to get eliminated by selection and the population still remains at node C.

Then you have may have another mutation occurring and one of the individuals in the population could go to sequence B. What has happened here is that there while the rest of the population is growing at growth rate equal to 1, this one mutant which has sequence corresponding to that of node B is growing at a growth rate 0.8 which is less than 1.

So, the growth rate or fitness of this bacterium is less than the parents genotype that it came from and when there is been a decrease in fitness or growth rate this mutation will set to be a deleterious mutation. So, you could have lethal mutation you could have deleterious mutation. What could also have happened the third possibility as shown here is that mutation from C to F might happen. And this results in increase of fitness from 1 to 2. And when this is that to happen this type of a mutation is referred to as a beneficial mutation. And now when there is a mutant which belongs to genotype F that mutant grows faster as compared to the original genotype that the rest of the population is located at, enhance because of selection that mutant starts eliminating the individuals of the parent genotype and the population eventually moves from point C to point F, the relative frequency of individuals which belong to genotypes C decreases. And the relative frequency of the individuals which belong to genotype F increases and eventually all and L in n individuals come to sequence F.

As and when we are here the whole story repeats itself from F you could go back to see in the graph that we have been given from node F you could have a mutation and go back to C sequence C, but that entails a decrease in fitness hence this is a deleterious mutation and will be eliminated by selection. So, this this does not work, but you could have a mutation going from F to going to node D which entails an increase and fitness from value of 2 to value of 3 and then you will have movement of population approach d. So, when we are talking of mutations you could have these 3 type of mutations, there is another type of mutation that could that you could have happened in the process which is we draw another node here, just let say C going to node g and the fitness corresponding to node g is just 1.

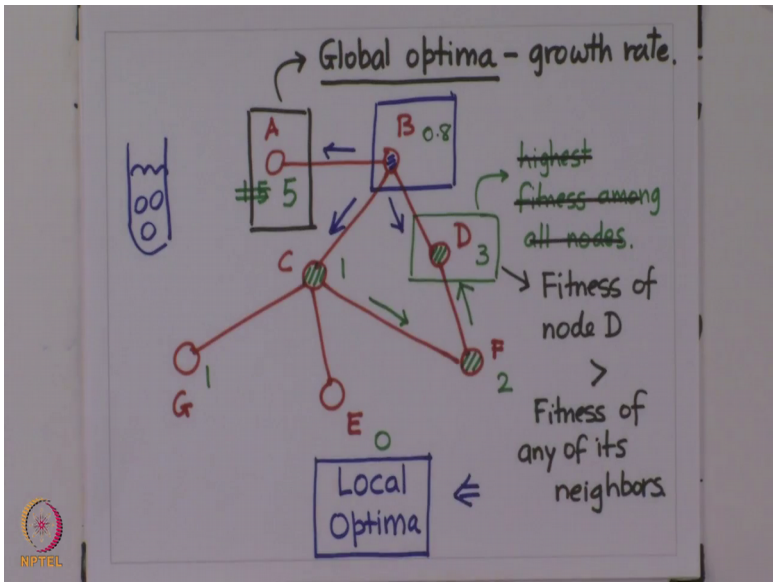
So, should this mutation happen you have an individual in the population which is also growing at the same rate as the parent's genotype, but the genotypically this mutant is distinct from the parent phenotype that it came from. So, these are genotypically non identical, but phenotypically identical and when you have a mutation that has happened which does not lead to a change in the fitness associated with the growth of the organism

this type of a mutation is said to be a neutral mutation. So, again you have this graph and you are going to have movement of populations on this graph depending upon the fitness values associated with particular node when you have movement from a node to another node which leads to an increase in fitness; that means, selection is going to act and is going to eliminate the individuals which belong to the parents genotype and your going to have population more number of individuals present in the new genotype which is which is growing at a rate higher than the original genotype.

On the other hand, if the mutation that has occurred is deleterious in nature or lethal in nature than again selection is going to act and eliminate those mutations. Neutral mutations which are mutations which lead to individuals which are genotypically distinct, but phenotypically identical and the sense that there growing at the same rate have a have a very different implication on evolutionary dynamics there something that will come back to later in the course. For the time being I want you to remember I want you to sort of realize the significance of these 4 type of mutations that can happen and how in this sequence space and the graph like structure that we have defined here populations are going to move along channels where fitness is increasing.

All those paths where fitness is decreasing are not going to be allowed because of selection. I want to emphasis a couple of other points related to this graph. So, slide is quite (Refer Time: 19:21). So, let me just re draw quickly this graph again.

(Refer Slide Time: 19:27)



So, we have node A B C and the corresponding fitness members are 0.8. 3 2 0 1 1 and 1.5. So, what we realized is that if are starting population is node C. Then what is going to happen is that population is going to move towards node F first and from F once all individuals selection will dictate that individuals belonging to genotypes C get eliminated and the mutant which is at sequence F goes faster than those at sequence C and hence population will move towards F. Once you have large number of individuals here you wait for a mutational event. So, that one genotype one individual with genotype corresponding to D arises and hence you have population moving to genotype t again why action of selection taking place here. So, from C this is the direction that the population will move towards and eventually end up at node D starting from C.

We can see that this ensures that the population ends up at a node which has the highest fitness among all nodes. And which make sense because natural selection would mean that you are moving towards sequences which grow faster and survive better than any other sequences that might be competing with them in the environment, but what happens let say if then fitness of node A was not 1.5, but the fitness of nodes a was let say 5. Now in this graph that we are talking about what going to happen is that there are few things that play out one thing that we can clearly node is that D is no longer the highest fitness among not all nodes. If this changes this is no longer 2 and now what we have that node A is the highest fitness and for that reason this is called the global optimal in terms of growth rate associated with bacteria.

So, this is the global optimal value of growth that individuals which are carrying one of the sequence is defined by these nodes can hope to reach, but we realize that F you are starting from genotype C to get to a we have to take a step down in fitness and then take another step up in fitness going up to value of 5, but this step down in fitness is not going to be facilitated why a natural selection. So, this in a strict (Refer Time: 23:06) is not allowed enhance population does not quite reach A, but we will take these 2 steps up and end up at node D. What is a characteristic of node D is that the fitness associated with node D is greater than fitness of any of it is neighbors? So, D has 2 neighbors here B and F and the fitness corresponding to node D is more than that of fitness corresponding to node B or node F.

Hence this result holds for node D and whenever that holds we said that that particular node is a local optimum, and local optima will have consequences in terms of how

populations move on fitness landscapes something that we will come to a little bit later on right. So, we have global optima and we have local optima in on this graph all right. Another curious perspective that is associated with this graph is what happens if the starting genotype of our population was not nodes C in this graph, but was actually node B what happens if it was node B that we were starting with now node B, remembered mutations are random events mutations take place randomly. There is no directionality associated with where and how mutations are going to take place.

So, starting from node B we do not know a priority if the first mutation that is going to happen in this test tube where I have genetically identical cells of genotype B growing, I do not know a priory if the first mutation that is going to arise in this cube is going to be a mutation towards sequence A or mutation towards sequence C or mutation towards sequence D these are all presumably equally likely, but this has important implication in terms of where the population ends up if the first mutant that arises here corresponds to sequence A; that means, the fitness of that mutant has gone up to 5 as compared to fitness of 0.8 which was that of the original genotype which is the big increasing fitness and hence selection will at and the population moves towards 5, and the population has reached the global optima it cannot to better than growth rate of 5 at any of the nodes hence the population stays there.

However, if the first mutant that arises was this mutant C. Then the population fitness has increased from point a to one selection acts and on the entire population moves towards the sequence corresponding to node C. And from there on the story repeats itself from C it moves to D from F and from F it moves to eventually D and stays there because it is stuck in a local optimum. Or it could have done that the first mutation is from node B to node D and in that case it is immediately reach this local optimum and the population stays there. So, depending on which was the first mutation that happened in the environment the dynamics of evolution of these populations are going to vary and in that sense this this evolutionary dynamics of population growing in this test tube is being played out by chance which decides which is the first mutation that happens in a test tube and that eventually decides the movement of the population on the fitness landscape structure that we are talking about.

So, we will we will stop this lecture here and we will continue our discussion on fitness landscapes in the next class.

Thank you.