Evolutionary Dynamics
Supreet Saini
Chemical Engineering
Indian Institute of Technology Bombay
Week 05
Lecture 22

Hi, welcome back, everyone. Let us continue our discussion on fitness landscapes. So, so far, we have seen two concepts. One is the concept of sequence space, and the second one is the concept of Hamming distance.

So, from the context of sequence space, we saw that for any sequence of length L, the number of possibilities of DNA sequences, unique DNA sequences, is 4 to the power of L. Hamming distance is defined as HIJ, which is the distance between sequences I and J, and is simply equal to the number of nucleotide positions where sequence I and J are unequal, are not the same. So, we saw that if sequence I is AA and sequence J is AT, then the Hamming distance between these two sequences is simply equal to 1. Because at the first position, they are identical; at the second position, they are not identical, so this is not true.

So, the number of positions where they are not the same is 1, and that is the value of Hamming distance IJ. So, having made these nodes, we represented each sequence by these nodes. The next step we will do is connect nodes I and J if the Hamming distance IJ is exactly equal to 1. So, in the DNA sequence of length 2, what we have is 16 nodes. What we need to identify is what we need to identify here is every pair which is one Hamming distance away from each other. If that is the case, then we join the nodes with an edge. So, let's do this, but let's first make these nodes. Let's make these nodes. So, what you should see is that A and AT, the Hamming distance between these two nodes is 1 because they are identical at the first position A, but not identical at the second position T.

So these two have to be joined with an edge. Same with AA and AG because they are identical at the first position but not identical at the second position. So these two have to be joined as well. And the same logic applies to AC. Let's continue and find other partners for this node AA.

AA will also be connected to TA because these two are identical at the second position but not identical at the first position. So these two have to be joined with each other as well. AA will also be connected to GA for the same reason: identical at the second position, non-identical at the first position. So these two are joined as well, and similarly, AA will be connected to CA. And you will

see that AA should not be connected with the remaining nine nodes because for each one of these nine nodes, the Hamming distance between AA and them is two.

And our rule is that we will only connect two nodes if the Hamming distance between them is one. So the number of partners of each node in this case is simply equal to 6. By the same token, we can see the number of partners of another node, AT. AA will be a partner, but that has already been connected. AT will be a partner with AG, and AT will also be a partner with AC.

Similarly, AT will be partnered with TT and with GT and with CT. So, again the number of partners for AT as a node will also be 6. So, every node will have 6 partners. and the way we get that is because length is equal to two so if i have any sequence let's say aa and i am looking at hamming distance one partners i can only mutate one position i can't mutate both the positions because if i mutate both positions then hamming distance becomes two so for hamming distance one i can only mutate one of the two positions Now, I could mutate the first position, but I could mutate it in three different ways.

I could change this A to G, T or C. So, that gives me three nodes which are neighbors of AA by only hamming distance 1. Similarly, I could mutate the second node also in three different ways. I could mutate it to G, T and C which is three nodes. So, 3 plus 3 is equal to 6 nodes. And we can generalize this that if we have a sequence of length L, if we have a sequence of length L, then the number of nodes

In the sequence space is equal to four to the power L that we have already seen. But in this. In this map that we have. How many how many partners will each node have? And it's easy to figure that out because I could when I'm looking for partners, each of which is at a hamming distance one, that hamming distance one could be because of dissimilarity at position number one, which could happen in three different ways.

Or that Hamming distance could be 1 because of dissimilarity at position number 2, which could be in three different ways, and so on and so forth. So, every position offers me a chance to make a mutation in three different ways. And hence, since I am only allowed one mutation, the total number of neighboring nodes, which are connected nodes, is going to be 3 times L. Three options for every one of the L positions that this sequence has. And this is the reason why, in the previous case, we got the answer 3 times 2, which was 6 neighbors for every position.

So that gives me some idea of what to connect and what not to, and I will end up with a structure that looks like this. Here we have only drawn completely for these two nodes, AA and AT, but the

entire network is going to be much messier than this. So, what do we do next? Now there is only one thing left. In this diagram...

We are going to have to draw that diagram again. Node number 1, node number 2, going all the way to node number 4 to the power L. We are going to do a hypothetical experiment. And our hypothetical experiment is as follows. That for a sequence of length L, I have 4 to the power L options, unique sequences.

$$\#\ edges\ of\ each\ node = 3L$$

I will list them all here. So let's say this is sequence number one, which is represented by node one. This is sequence two, sequence three, sequence I, sequence J, going all the way to sequence four to the power L. I will do a thought experiment now. The thought experiment is, imagine that I have E. coli, but it doesn't have DNA. And obviously that's not going to be possible, but this is the thought experiment.

In this E. coli, I will add DNA, which is corresponding to sequence which is represented here. by node 1. So let's say this is the sequence space map and connections are made where L is 5 into 10 power 6, which is E. coli genome. So each one of these sequences is of length 5 into 10 power 6. And in my thought experiment, I take E. coli's cell membrane and everything that comes except for its DNA.

$$Total\ \#\ nodes = 4^L$$

In place of DNA, I will take the sequence associated by node 1 and put it inside this E. coli and add this to a flask where I have growth media. And in this growth media, I will add this E. coli whose DNA is represented by the DNA sequence corresponding to node 1 in the sequence space map. And I will measure the growth rate here, which will be a proxy for fitness. So I will do this experiment and I will record against node 1 the fitness associated with the bacteria that was carrying the DNA sequence represented by that particular node. In my thought experiment, I will do this again, but this time the DNA that will be carried by E. coli will be the DNA sequence represented by node 2 and hence I will get the fitness measure associated with the sequence associated with node 2 and I will write this as F2.

And I will keep on doing this till I am done with all of these sequences of 4 to the power L. So eventually giving me F 4 to the power L. And we say that this is a thought experiment because of the sheer numbers involved in this. If L is 5 into 10 power 6, there is no way we can do this experiment for 4 to the power 5 into 10 power 6. But anyway, in this thought experiment, E. coli carrying DNA represented by nodes is added to growth media and the growth rate is measured. So

we have a number corresponding to every one of these nodes that we have drawn here. And in the final step to get to fitness landscape, what we do... So now we have to imagine this sequence space as...

nodes in this plane so sequence let's do this on the next slide so i have this sequence space nodes in this particular plane going all the way to 4 to the power l and let's say 1 and 2 are connected 3 and 4 are connected I and J are connected and so on and so forth. What these connections mean is that these two sequences are separated from each other by only one hamming distance. And now remember, corresponding to sequence 1, I have a fitness proxy growth rate, which is F1.

Corresponding to sequence 2, I have fitness proxy corresponding to F2, and so on and so forth. Going all the way to 4 to the power L, I have a fitness proxy, which is F4 to the power L. So now what I'm going to do is that this landscape exists in this flat plane of the screen. I will go to a particular node which is node 1 now. I will see what is the fitness value F1 and I will lift this node up from the screen by a distance which is proportional to its fitness value which is F1.

So let's say that node 1 has its fitness units. Let's write some numbers here. Let's say F1 was simply equal to 1, and F2 was simply equal to 2. So what I will do is add another axis to the screen, which is perpendicular to the screen. Now, I will go to node 1 and lift it perpendicular to the screen by a distance equal to 1 unit.

Similarly, I will go to node 2 and lift it perpendicular to the screen, but by a distance twice as much as how much I lifted node 1, and so on and so forth. I will do this for every node in this sequence space. Again, this is akin to adding another axis perpendicular to the screen and then going to each node and lifting it by a distance equal to the fitness of the organism carrying that particular DNA sequence. If I do this, I hope what you can visualize is that we will get a fitness landscape structure with ups and downs.

There will be many ups and downs associated with this. The nodes corresponding to sequences with very high fitness will be the peaks. Some nodes will have fitness equal to zero. They will remain on the screen. They will be the troughs, and so on and so forth.

And this three-dimensional structure is called a fitness landscape, where the plane is occupied by all the sequences, and the perpendicular axis is the fitness corresponding to each one of these sequences that we are talking about. I hope that the idea of a fitness landscape makes sense. So, viewed from the top, a fitness landscape could look like this. So, if these two axes are the plane of sequence spaces. What you should be viewing this as is like a map-reading exercise, where these regions represent heights.

So, this is the highest peak, which means this sequence—corresponding to this particular node—is the fittest one. And as I move away from this, I'm moving toward lower fitness values. Let's also look at it another possible way. Let us say that we have these sequences on one axis only. AA, AT, AG, and AC.

And let us say this is fitness. Now, imagine that this node had this particular fitness, AT was slightly higher, AG was significantly higher, and AC was a really low fitness or a really bad genotype. So, the structure we get is because A and AT are separated by a Hamming distance of 1, and these are all connected to each other. Because the distance between them is exactly 1. And this is the three-dimensional structure that we have now.

If there is a population at AT, let's imagine a population of bacteria present at node AT. One member of this population could acquire a mutation and move to AC. This movement from there to there leads to a decrease in fitness. The transition that happened here was AT going to AC. So, the mutation was T going to C. This mutation has led to a decrease in fitness, and hence, it is a deleterious mutation.

As a result, when one of the offspring of these individuals lands here, natural selection will remove it. Natural selection weeds out deleterious variants. The same story will happen if one of the individuals acquires a mutation such that an offspring with genotype AA is born. In that case also, this leads to a decrease in fitness. This is also a deleterious mutation, and hence, natural selection will remove the AA individual from the population.

However, there will be Eventually, one of the individuals will land at this particular genotype. Because it has a distance of 1, this mutation—a beneficial mutation—will occur, resulting in an individual with genotype AG. So, from AT, the population moves to AG. Now, because this has resulted in an increase in fitness, this is clearly a beneficial mutation.

And natural selection selects this particular variant. And as a result of this, the progeny of this individual are going to outcompete the progeny of the ancestral population, which will be outcompeted by AG genotype. And hence, the entire population, given sufficient time, the entire population transitions from being at AT to being at AG. That transition takes place. What is important to recognize here is that in this view of evolution of populations here,

Populations, any variations that are of lower fitness, that are variations that are obtained via acquisition of deleterious mutations are movement down on this mountain structure that we have made and natural selection gets rid of them. And only variations which lead to an upward movement on this landscape are beneficial mutations and hence are selected by natural selection. As a result,

this adaptive trajectory of the population is always going to be in an upward direction towards increase in fitness. And this is the metaphor of fitness landscape and the geometric interpretation of fitness landscape that is used to visualize evolutionary processes in this context. So next video onwards, we are going to look at actual concrete examples from last few years.

We'll discuss three specific examples from the last few years, which have experimentally tested these ideas and study how populations move on these landscapes. And we'll continue that in the next lecture. Thank you.