Evolutionary Dynamics
Supreet Saini
Chemical Engineering
Indian Institute of Technology Bombay
Week 05
Lecture 23

Hi, welcome back everyone to the next lecture in the video. So, what we have discussed so far is that fitness landscapes are these geometric interpretations of the process of evolution. Where adaptation to an environment or increase in fitness in an environment is viewed as this uphill climb on this mountain on which the height represents the fitness of a population. And we learned to do this by drawing these sequence spaces, then connecting nodes which represent sequences that are only separated by one Hamming distance, and then adding the third dimension which represents the height or the fitness of that particular genotype. Now, we'll soon move toward a discussion of three examples of fitness landscapes from experimental systems.

All three are examples from bacteria from the last few years. But before we do that, we need to understand one or two more concepts. So the first concept that we need to understand is that when we suppose in the sequence space, we had these three nodes. When we were joining them, we defined the rule of joining two nodes that only nodes which are separated from each other by one Hamming distance will be connected. So AA and AT will be connected, and AT and TT will be connected, but AA and TT will not be connected because the Hamming distance between AA and TT is equal to two.

And only nodes whose Hamming distance is one are connected with each other. So the first question that we want to understand is, why did we only connect sequences which are separated by one Hamming distance? That's the first question that we'll address in this video. To understand that, let's imagine a fitness landscape. Let's draw a fitness landscape where this sequence is, let's say this is the genotypic space.

This is AA. And then AT is further up in the fitness scale. This is AT, and TT is among these three the best possible genotype, which confers the highest fitness. So, in this representation, the X-axis represents the different genotypes available to us, whereas the Y-axis represents the fitness of each of these genotypes. So fitness, and as we've seen, the most commonly used proxy for fitness is simply the growth rate of the bacteria with that sequence. Now, one question that may arise among you is: how can bacteria only have sequences of length two?

The way to think about this is to imagine three bacteria with their genomes, and the genomes are identical everywhere. So all three are identical bacteria except for two locations. This particular one has sequence AA, this one has AT, and the third bacteria has TT. As far as the genome scale is concerned, the length of this DNA might be anything. But because, in this entire population with three individuals, the genomic sequence is identical everywhere except these three locations, we only consider the places in the genome where sequences differ.

So, as a result, this represents three different bacteria, and we assume—the assumption in this graph—that in the rest of the places, their DNA sequences are identical. So, in that context, let us imagine that our starting population is at this location. And The starting population is at AA and has a size N, meaning there are N individuals in this population. Now, mutations will arise, moving from AA to AT.

can only happen via acquisition of mutation. This event can only happen if during the process of division, one of the progenies acquired a mutation and it became AT instead of AA. This is the only process via which an AT individual can arise, which means the relative quantity that we should be interested in is the mutation rate. which is represented by mu, which is the rate at which the DNA polymerase makes errors as an individual divides. So, the units of mu are errors per cell per generation.

And in E. coli, the value of mu is 10 to the power minus 3, which means that In 1000 divisions, there will be one error that takes place during the process of replication. So, if we have divisions taking place like this, in 999 of them, these two will be identical to the parent that they came from, whereas in only one of them will there be a mutation and which brings about genetic diversity in a population. So, per division, this is the probability or the chance that error is going to take place.

Hence, the number of mutations that take place in a generation is simply N mu. Now, from the context of the fitness landscape that we drew, the three genotypes that we are looking at are AA, AT and TT and their fitnesses are in this particular order. So, what we are saying is that starting with this population here, if our starting population is at the location AA, for this population to move up on this fitness landscape, this is fitness, For this population to move up on the fitness landscape, what needs to happen is that one of these individuals needs to acquire a mutation and come to this particular genotype.

And then the fitness is higher, and then this individual's progeny are going to outcompete the progeny of the population here, and the entire population shifts to the AT genotype. But we saw that in every replication that is happening at this place in this population, the chance that an error—any error—will happen is 10 to the power of minus 3. This change, which in the context of genome

replication, So remember that if this is the genome of the organism, then there is only this location where there is AA. The chance that during the process of replication, any error will happen,

which could be anywhere on this genome, is equal to 10 to the power of minus 3. The chance that the error that takes place happens at this particular location that we are interested in is obviously much, much lower than this. So, the chance of movement of the population from AA to AT via this division, which leads to the production of this particular progeny, is much, much smaller than this. So, this in one sense is a very fortuitous event to happen to the population—that one of the progenies changes from the AA genotype to AT. Albeit, they are only separated by a Hamming distance equal to 1.

You just need one mutation to move from one to the other, one to the other. But in all likelihood, errors are going to take place at other locations, and it is going to be extremely lucky that this A changes to T. So, that is a change that will happen. But now, let us say that the probability that AA changes to AT is this mutation will happen—let's say—is of the order of 10 to the power of minus 9. That means only in one billion cell divisions will there be one individual who has a change from AA to AT.

So that's a very fortuitous event to have happened that I get one individual of this kind. And it is this fortuitous event that is reflected in the sequence space when I am connecting AA and AT with a node. So, I have three nodes here. These are the three nodes and I am connecting AA and AT here. with an edge and AT and TT with an edge but not AA and TT.

And as we have just seen that this movement from AA to AT happens with a probability of 10 to the power 9. Mutation rate is 10 to the power minus 3 but we want a specific change not just any change. Hence, this has to be divided by 10 to the power 6 which is the genome size of E. coli and hence we get the chance that this transition happens is 10 to the power minus 9 which means in every billion divisions one division will lead to generation of this particular progeny but this is possible and this is possible via acquisition of just one mutation hence we connect this so the connection here this edge represents This edge represents that an AA individual can acquire mutation to become AT.

that is the meaning of two nodes being connected by an edge, that you can acquire one mutation and become the other genotype. So, that is the reason for connecting any two nodes, that you acquire one mutation, become the other one. Now, we want to understand why are we not connecting AA and TT? So to understand that, if we look at this E. coli cartoon again and we see that we have this genome and at a particular location in the genome I have AA. And the change that I am seeking is this AA should become TT.

I am seeking AA. a change of 2 hamming distance because these two sequences are separated by 2 hamming distance. the chances of these two occurring together, the chance of one occurring is 10 to the power minus 9 as we have just seen, but the chance of the second one occurring is also 10 to the power minus 9 and the chance of both occurring in the same cell is 10 to the power into 10 to the power minus 9 which is 10 to the power minus 18. The chances, so what this analysis shows us is that a chance that an AA individual is going to divide and the progeny will have a TT genotype, that chance is 10 to the power 9 times less likely as compared to this particular chance. So because this is such a rare event, even bacterial populations will be of the order of 10 to the power 10, 10 to the power 11 in lab settings.

So the chance of an event like this happening are exceedingly rare even in a lab population. And as a result, what we are saying is that for a population, it's not possible for this AA population to directly jump to TT via acquisition in one division process only. it has to happen this way. If this population has to move to TT, it has to first acquire this mutation AA to AT and then when the population has shifted, then the population size will grow here because this individual will divide and only then one of the individuals will acquire a mutation and then we will have a new individual at genotype TT. So AA to TT transition can only happen in two steps and not one.

And the Hamming distance associated with this is HIJ for the first mutational event. is 1 and for the second mutational event, HIJ which is again 1. So, you have this two-step process which leads to transition of a population from AA to TT and these probability numbers show us that these transitions are exceedingly unlikely to happen in just one step. So this is the reason that we don't connect nodes which are separated by more than one Hamming distance. The second idea that we need to explore before we go to empirical evidence of fitness landscapes is the idea what is referred to as epistasis.

To understand this, we will start with a case where there is no epistasis. Let's assume that we have a DNA sequence of length 3. And at each of these three places, let's call this position 1, position 2, and position 3. So, at each of these three positions, we can have either A, T, G, or C. These are the only options available for a DNA sequence.

Now, we are going to analyze this sequence in the following way. We are going to assume that the presence of any nucleotide at a particular location makes a fitness contribution. For instance, if at position 1, A is present, it makes a contribution to the fitness, which is, let us say, 0.2. However, if a G is present at position 1, its fitness contribution is 0.4. So, clearly at position 1, G is more preferred compared to A, and so on.

If we look at the correlation between the fitness and the DNA sequence through this prism, what we see is that A at 1 gives 0.2, G at 1 gives 0.4, let us say T gives 0.1, and C gives 0.3. At position 2, it's 0.4, 0.4, 0.4, 0.6. And let's say at position 3, this is 0.5, 0.1, 0.2, and 0.3. Let's imagine that these are the contributions of every nucleotide at every position in this DNA sequence, which is of length 3. Now, let's imagine that we have a bacterium whose starting sequence is given by TAA.

That means the starting sequence is TAA. TAA. And this population of bacteria which is carrying the sequence is evolving in this environment. So as we can see, suppose a change like the following takes place. That TAA, we have an individual whose sequence becomes TAT.

So the parent fitness was T at position 1 gives fitness of 0.1, A at position 2 gives fitness of 0.4 and A at position 3 gives fitness of 0.5. So the sum total fitness of this individual is simply equal to 1 which is 1 plus 0.4 plus 0.5. If an individual with sequence TAT is born then its fitness is 0.1 plus 0.4 plus 0.1 which is just 0.6 which is lesser than the ancestral one. So this individual is going to be eliminated by selection. This is eliminated by selection.

So what would be a change that would not be eliminated by selection in a case such as this? And what you should be seeing is that at position 3, A confers the highest fitness. If I look at position 3, A's contribution to fitness is highest and I already have an A. So at position 3, it's not really possible for me to improve. However, if I am looking at position 2, then if I have a transition from A to C, then the fitness contribution at position 2 becomes 0.6 instead of 0.4. So I can increase my fitness then.

And in position 1, T actually confers the least possible fitness of all the options available to me. So any change away from T would be a good change. In that context, suppose the following mutation took place: I get GAA. That means the new fitness will be 0.4 plus 0.4 plus 0.5, which equals 1.3. So that's a big increase from what I started with.

And that means the population will move from TAA to GAA. And now I see that even in position 1, I have the best possible fitness I could have had. Now the only other change that will increase fitness is GCA. And the fitness in this case will be 0.4 plus 0.6 plus 0.5, which equals 1.5. And that's the maximum fitness one can have under this definition when you have G at position 1, C at position 2, and A at position 3.

And that is what we end up with. Let's try another starting sequence, which is, let's say, CAC. So my starting sequence is CA and then C. This could go to, let's say, GAC. That would lead to going from 0.3 to 0.4, which would be an increase in fitness of 0.1. So that move is permitted.

Then the next move could be GCC, which is this move from A to C, which is an increase of fitness contribution from the second position. This increase in fitness contribution is of 0.2, a movement from 0.4 to 0.6. And lastly, I have C at the third position, and this could move to GCA, which is an increase of another 0.2 units in fitness, an increase from 0.3 to 0.5. What this means is, if we look at this—and as an exercise, you can try starting from other positions and keep increasing the fitness up until a point where you can't increase fitness anymore. What you will note is that, irrespective of what the starting point is, irrespective of the starting point, you will always end up at GCA as the final sequence.

So it doesn't matter on this landscape what the sequence is that you are starting with. You are only moving uphill, and those are the only moves that are allowed to you, and you keep moving up until you can't increase in fitness anymore. And irrespective of whatever your starting position was—you could start with AAA or AAT. Any sequence, you just acquire one mutation at a time. And that move is permitted as long as fitness is increasing.

And you keep doing that until you can't increase in fitness anymore. And you will always end up at the same sequence, GCA. So what has this told us? This has told us that if the fitness contribution of each locus, or of each position (also called locus), is independent of what is present at other loci, then

The final fittest sequence is unique. You can start from anywhere. The starting point could be anything. Starting from any point. You end up at this unique sequence, which in the example that we just saw was GCA.

But this is only true when the following condition is met: that the fitness contribution of each locus is independent of what is present at other loci. If we go back to the example that we just discussed, the fitness contribution of 1 is only dependent on what nucleotide is present at that particular location. So this A here makes a fitness contribution of 0.2, independent of whether it's a T, G, or C. It does not care what is present at position 2 and position 3. Its fitness contribution is independent of what is present.

It does not care what is present at positions 2 and 3, and its contribution is only dependent on the nucleotide present at that particular location. So if this condition is met, then we are led to this conclusion: the final sequence is unique, and starting from any point, we will end up at this unique final sequence with the highest fitness level, which in the example that we just saw was GCA. However, that is not true generally, and this phenomenon of the fitness contribution of a particular location being dependent on what is present elsewhere is called the phenomenon of epistasis. So,

the fitness contribution of A here is a function of what is present at position 2 and position 3. And this is the phenomenon of epistasis that we will continue in the next video.

Thank you.