Evolutionary Dynamics
Supreet Saini
Chemical Engineering
Indian Institute of Technology Bombay
Week 06
Lecture 28

Hi, welcome back to the next video. Before we discuss the final example of landscapes that was published recently, let us look at one important limitation associated with these two studies that we have examined. In both these studies, the underlying logic was the same: we have a DNA sequence, and there were five different locations on this DNA. In the first one, these five happen to be within one gene, and in the second one, they happen to be spread over five different genes. But the point that we are going to make is independent of that aspect.

In both these studies, these five locations acquired a mutation. So, let us imagine that again for simplicity—let us imagine that the nucleotide was present prior to mutations. So, this is the ancestor prior to mutation. The nucleotide present at these five locations was, let's say, A. Subsequent to mutation, in the first case, through the isolation of an E. coli with a higher MIC in nature, or in the second case, through this evolution experiment, when I sequenced the evolved strain, I saw that this nucleotide had changed. Now, without any loss of generality, let's just assume that the change happened to C.

At each of these locations. So there was an initial nucleotide present at each of these five locations, and there was an evolved sequence which was post-mutation. So in this case, using these Using these ancestral and evolved strains, we studied evolution as A transitioned to C in each one of these, and we called these mutations A, B, C, D, E, and we saw that going from ancestor to evolved, there were 120 paths. However, we know that

That DNA is not made up of two nucleotides, the ancestral and evolved. It's actually made up of four nucleotides. So what was missing here is that we did not take into account G and T at all. So these were completely omitted from these studies. Again, I'm writing A and C; these nucleotides could be different, but the point is that at every one of these five locations, we only worried about the two nucleotides that happened to be present before and after the mutation and did not worry about the other two.

As a result of this, When I am talking about the number of genotypes that I can construct. So, let us imagine that I want to construct a genotype by choosing between these two genotypes. So, in the rest of the places, the genotypes are identical. When it comes to the first position, I can either choose an A or choose a C. That means I have two options.

Of which nucleotide to choose when I am trying to make a list of all possible genotypes from these two sequences. When it comes to position number one, I can choose either A or C. When it comes to position number two, it is again two, and so on and so forth. So, for all five positions, I have two possible choices to make. I can either choose the nucleotide A, which was there prior to the existence of the mutation, or I can choose C, which is the nucleotide present after the mutation.

What that means is, in such a scenario, the total number of genotypes which is going to be 2 × 2 × 2 × 2 × 2, or simply 2 to the power of 5, that is equal to 32. So, this is an adaptive path. This study involves 32 different genotypes, and depending on the combination in which these mutations occur, I can find out the trajectories associated with all 120 paths. However, we know that this is not a complete picture associated with these five locations.

In reality, at each of these five positions, I have four choices, not two. Not only do I have a choice of A or C, but I also have a choice of G or T. Which means the total number of paths available, or the total number of genotypes possible with variations at these five sites, is not 2 to the power of 5, but actually 4 to the power of 5. Which is much, much larger compared to 2 to the power of 5. So, thus we run into some sort of problem, and this is the challenge here.

which has faced experimental determination of fitness landscapes, that 4 to the power of 5 is a very large number. And this is not even one gene. This is actually only five positions in the genome of an organism. If I wanted to build a landscape of an entire gene, Remember yesterday, in the previous video, we did some, we sketched some numbers, and we said that for a gene which starts with ATG and ends with TAA, this length is typically 10 to the power of 3 nucleotides.

At each of these nucleotides, I have four choices. That means the total number of DNA sequences of length 1000 is going to be 4 to the power of 1000, which is an astronomical number—impossible to construct in a lab and then test to construct the entire fitness landscape. So more recently, the effort has been that instead of worrying about the complete sequence of the gene, we will do the following. We have ATG to TAA. We will not worry about the entire sequence of a gene, but we will only worry about some parts associated with the gene, which we know are important for function.

When we define a gene that is eventually transcribed and translated into a protein, and this protein, let us say, performs some catalytic function of converting X to Y, then we know that We know in many cases that there are certain amino acids in the protein which are very important for catalytic function. In addition, there are other amino acids present far from where catalysis occurs, which are less important. So we have this understanding for several proteins, and our effort has been to understand how amino acid residues define the catalytic properties of a particular protein. Hence, one approach has been to only focus on sites that are important.

Important sites. And in this way, we limit the number of possible DNA sequences because we can't do it for a thousand nucleotides, but let's say there are only six important sites. As drawn in this cartoon, we can make 4 to the power of 6 DNA sequences and test the catalytic performance of each one of these 4 to the power of 6 sequences. 4 to the power of 6 is still a very large number, but it is now possible to test this quantity of DNA sequences and their corresponding fitness.

So that is the approach that the third paper has taken, which was published last year in Science by Wagner and co-workers. And that is what we wish to discuss in this video. The gene that they chose is called folA. It encodes an enzyme. So let's say this is the entire length of the gene.

This is the ATG start and this is the TAA stop. This, of course, undergoes transcription and translation. And then this folds, and the resulting protein is called dihydrofolate reductase. It is not important to remember the name, but this is only for completeness. So what they do—what is known from previous studies of DHFR—is that there is a part of the gene somewhere near the start of the gene.

This is a nine-base-pair region, which is important for function. And the function is to catalyze a reaction that eventually leads to the synthesis of tetrahydrofolate, which is an important metabolite in the cell. This nine-base-pair region encodes three amino acids. The sequence that is present—the nine-base-pair sequence in the actual gene—is given as follows. GCC, GAT, CTC.

GCC in the standard genetic code encodes for alanine, which is represented by A. GAT encodes for aspartate. which is represented by D, and CTC is code for leucine, which is L. This is amino acid number 26, this is amino acid number 27 and this is amino acid number 28 in the coding sequence associated with this. That means prior to this region there are 25 amino acids which means and each one of these amino acids is encoded by 3 nucleotides which means that there are 75 bases before we get to this catalytically important site. So I have a region of DNA which is 9 base pair long.

What is also known previously that mutations in this region in this region confer resistance to an antibiotic called trimethoprim. So this nine base pair region is also important from the context of resistance to an antibiotic. Using all of this background, the strategy that Wagner and coworkers adopt is that in this nine base pair region, They synthesize every possible sequence of length 9.

Every possible DNA sequence of length 9. What that means is that there are going to be 4 to the power 9 DNA sequences that have to be constructed. This will start with something like all A's and go all the way and maybe end with all G's. And in between, there are all sequences, all possible sequences of length nine. So you construct this, so let's say we have

Let's say we have a bacteria which is carrying sequence number 1. We have another bacteria which is carrying sequence number 2. And so on and so forth, going all the way to bacteria which is carrying sequence number 4 to the power L. this 4 to the power 9 is a number which is a little bit above 2 lakhs. So, you should sort of try and understand the change in scale that has happened from the two examples that were around 2010 to now where these numbers have changed from 32 genotypes to over 2 lakhs genotypes.

Next, we sort of pool them together. We have a flask where all of these variants are added in equal number. So, the starting number for each one of these 4 to the power L are same. So, equal number of each of the 4 to the power L variants. L here, of course, is nine.

So there are equal numbers of each, and because we want to test the fitness of these in the context of an antibiotic, this media contains trimethoprim. And we let growth take place in this flask at 37 degrees Celsius, shaking. And what is going to happen is that some variants are better adapted at surviving. Some of these variants of the 4 to the power L variants that I have thrown in this flask, some of these variants are better adapted at surviving in these conditions in the presence of an antibiotic, whereas some of these are going to be really bad at surviving. So as growth takes place, as this experiment moves forward, the relative frequencies of these four to the power L variants is going to change as we move in time.

Imagine that this individual is really good in these conditions, whereas this individual can barely divide. So as the experiment moves forward, What I might see in a few hours time is that now the flask contains, now the flask contains the following. That I have three individuals of the green type, whereas the red one hasn't been able to divide, there is only one type of individual of the red type. So what started between the green and the red, the experiment started with a ratio of 1 is to 1.

It has now become in a few hours 3 is to 1. And using this approach, and this is of course only for two genotypes, I can do this for every genotype that exists in this flask. And hence, I can get a map of relative frequency of each genotype with respect to another one. So, for the green and the red that we are saying, their relative frequencies have changed. And this allows me to get their relative fitness in the environment that we are talking about.

This quantification of 4 to the power L genotypes after a few hours happens by the following procedure. That at the start of experiment, if I sequence the population, I will get equal number of individuals of each kind. Of each kind. However, as the experiment goes forward, because differential growth has taken place, some genotypes have grown more and the others have grown less. Their relative frequencies are no longer one.

And hence, now if I deep-sequence this population, I will get an unequal number of individuals of each kind. And this allows me to get estimates of the fitness of each variant, which allowed starting from an equal frequency to this unequal frequency in a given amount of time. So using those calculations—using this approach—I get fitness. For every one of the 4 to the power of 9 genotypes. And not just that, I also know that in these 9 genotypes—in these 4 to the power of 9 genotypes—let us say there is genotype 1, and let us say this is genotype i, this is genotype j. Let us say the only difference between them is just this one mutation.

They are the same at all of the other 8 positions. That means the Hamming distance between i and j is equal to 1, which means that using this fitness data and using the Hamming distance between each pair of sequences, I can draw a fitness landscape. I can draw a fitness landscape which is complete in nine different positions in the genome. So what did they see? Remember the basic goal of deciphering movement on landscapes starts from the following question.

That is a landscape like this. Let me remove this line. Is the landscape like this, which is single-peaked, which means irrespective of where we start from, the populations are going to move towards the same peak, and hence evolution is very predictable? Or is the structure of the landscape like this? And so on and so forth.

This is a multi-peaked structure now, where your evolutionary fate as a population is highly contingent on the starting point—and not just the starting point. You could be present at a starting point such as this, where fate is going to be determined by whether you're moving in this direction, where the first mutation happens towards the left or towards the right. So that is a question they answer using this landscape, which is complete in nine possible positions within the framework of this gene called FOLA. Now, remember, if we compare the lessons we learned from the first example—which had five mutations in the first gene—we saw that the landscape was extremely, the landscape was highly dictated by sign epistasis, which means there are going to be lots of peaks in this landscape. In fact, we saw that of the 120 pathways—120 paths to transition from zero mutations to five—only 18 were permissible, and 102 of them had a local peak. So the first question we ask is: how many peaks are there in this landscape?

And what they're, data shows in the published paper is that this landscape has 514 peaks, which means the population could get stuck in this peak or that peak, and so on—and there are 514 peaks like this, which means Such a high number of peaks tells us that evolution is highly unpredictable. However, they found one interesting property associated with this particular structure. Let's take a look at that.

Let's draw this a little bit bigger. One peak, more peaks, more peaks, another peak, and so on and so forth. There are 514 of them. So, If this is the landscape I'm talking about, then there are high peaks and there are low peaks.

The total number of peaks is 514 high peaks are fewer in number low peaks are more in number which suggests to us that because they are more in number, that for a population starting out on this landscape, it's very likely that it is going to get trapped in a low peak instead of going to a high peak, which means adaptive solutions to a population are more likely to be bad as compared to good because there are much more number of smaller peaks as compared to a larger peak. However, what they find that that's not the case. And the reason for that is that the way this landscape is structured, that this peak, for instance, this is one low peak, is only going to be accessible to sequences which are present in this region.

Any sequence which is present in this region is going to go to this particular peak. So the fate of all these sequences is this peak. Let's say the number of sequences here is the basin of attraction for this particular peak. Similarly, Each one of these peaks, local peak, has a basin of attraction such that if any sequence started out from that basin, it's going to end up on that particular peak.

What should become apparent from this figure is that larger peaks have a much larger basin associated with them as compared to smaller peaks. As a result of which, that even though there are fewer in number, the total fraction of landscape which is the basin of attraction for high peaks is close to 80%. Which means despite there being fewer number of high fitness peaks, 80% of the starting sequences are going to end up at these high peaks because these higher peaks have a much larger basin of attraction as compared to all these large number of small peaks even put together. That was the main result associated with this paper.

So what we have seen through these three examples is that landscapes are difficult to construct experimentally. They often show unpredictable results. Their manifestations are different depending on where the mutations occur. Within a gene, mutations work differently compared to between genes, epistasis, and so on. And now, today we have landscapes of many proteins available to us, and they show interesting results such as this.

A few examples of which we have tried to glance at through these discussions of three papers. We will continue with another aspect of evolution from the next video onwards. Thank you.