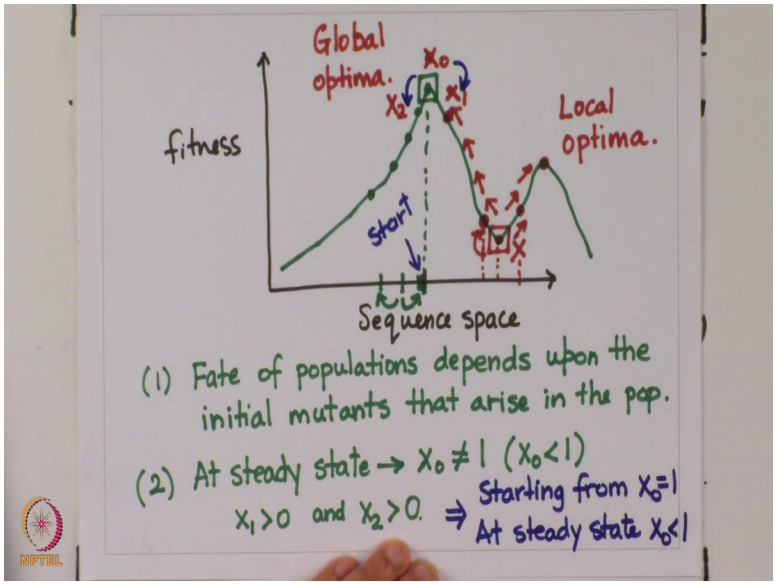
Introduction to Evolutionary Dynamics
Prof. Supreet Saini
Department of Chemical Engineering
Indian Institute of Technology, Bombay

Lecture - 17
Obtaining fitness landscapes experimentally

(Refer Slide Time: 00:36)



Hi everyone, and welcome to the next lecture of the course evolutionary dynamics. We will continue our discussion on fitness landscapes from where we had left off last time. And we were discussing the concept associated with what are called quasi species. So, just to recap if we have a landscape in one-dimensional which you know, so the x-axis here represents the sequence space, and the y-axis is the fitness corresponding to a particular sequence on the x-axis. Then our landscape may look something like this. And now what we are saying here is that mutations can only happen between neighbors. So, these two sequences are neighbor. If this one were to acquire a mutation it could either they come this sequence or its neighboring sequence on the other side and the corresponding fitness are represented by the values on the y-axis here. And this of course, is the sequence which represents maximum fitness. This is f max, which is given by this particular sequence on the x-axis.

And now we what we had looked at last time is if we were to start our evolutionary experiment by the entire population at this particular sequence then the fate of this

evolutionary experiment is going to be decided by in which direction does the first mutation happen. If the first mutation happens along this direction, so then you have an individual at this point corresponding to this sequence, whose fitness is higher than the original genotype that we were talking about that we had started from. And this particular individual belonging to this genotype then via selection eliminates the other individuals in the population which are at a lower fitness and corresponding to this particular sequence.

And hence the entire population moves up the hill and you have a greater number of individuals belonging to this particular genome, and this gets eliminated via selection. And the same story keeps repeating itself and eventually the population reaches the peak of the fitness landscape, and we saw that when mutations are happening in an environment where DNA replication is not error free eventually there is an equilibrium established between frequencies of populations belonging to. So, you have X naught equal to frequency of individuals which belong to this particular genotype X 1 is the number of individuals belonging to this particular genotype and X 2 are the individuals belonging to this particular genotype. And we saw that at steady state X 1 and X 2 are not going to be equal to 0, because of finite mutation rates that are taking place X naught.

On the other hand, if the initial mutation that we started from happened in this particular direction then you have an individual, which belongs to this particular genotype. Again at a higher fitness level than the original genotype from which we started the experiment and the population that keeps moving in this direction and get trapped in what was described as a local optimum. And this one in the sequence corresponding to x naught here is what was called the global optimum, so that is the first aspect associated that we want to highlight here that fate of population on when they are moving on fitness landscapes depends up on the initial mutations or initial mutants that arise in the population.

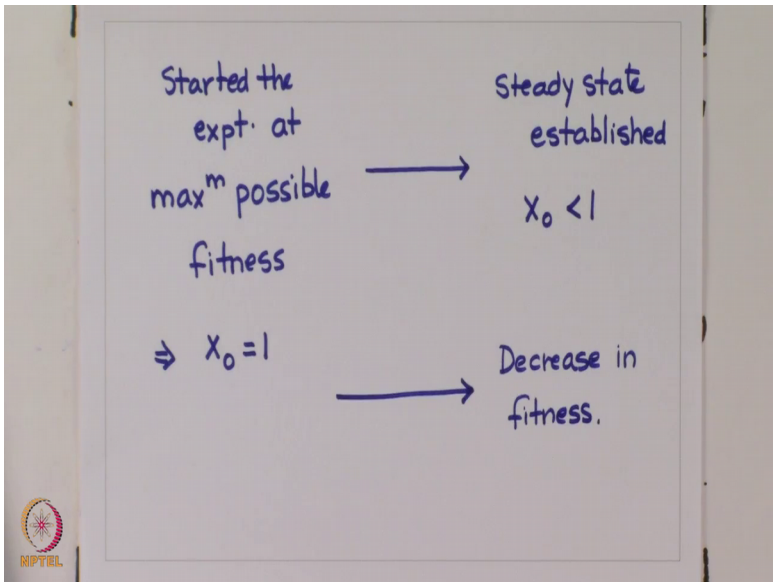
And depending on this, the population could gather either get trapped in a local optima or be able to reach global optima depending on which was that direction that the population embarked up on mode. And the second consequence of this is that at steady state. And this we are talking about the case when mutations are allowed. So, at steady state X naught is not equal to 1, which is the fraction of individuals which belong to the

genotype corresponding to the highest fitness is not equal to 1, X naught is going to be a number which is less than one which means that X 1 and X 2 are greater than 0.

This has an interesting consequence. Imagine you are doing an experiment where the starting population is has a genotype which corresponds to the maximum fitness. So, all your cells are actually present at this particular sequence. Suppose, we were to start our experiment at this particular genotype here, which means the starting population has the maximum possible fitness that it could have had. But because mutations are happening eventually steady state is going to get established in this system, and you are going to have mutations take place and at steady state X naught is not going to be equal to 1, and you will have positive and finite values of X 1 and X 2.

What that implies is that starting from X naught equal to in such an experiment when you start from X naught equal to 1, at steady state, it can be easily visualized here that X naught is less than 1. What that implies in the context of fitness of the organism is that we are starting with all individuals at maximum possible fitness X corresponding to sequence corresponds the sequence which corresponds to the maximum fitness. However, because mutations are happening at steady state there is going to be a dynamic equilibrium where the fraction of individuals belonging to a particular genotype is going to get distributed around the peak. And then the fraction of individuals which correspond to the sequence which gives the organism its maximum fitness will be less than 1.

(Refer Slide Time: 07:19)

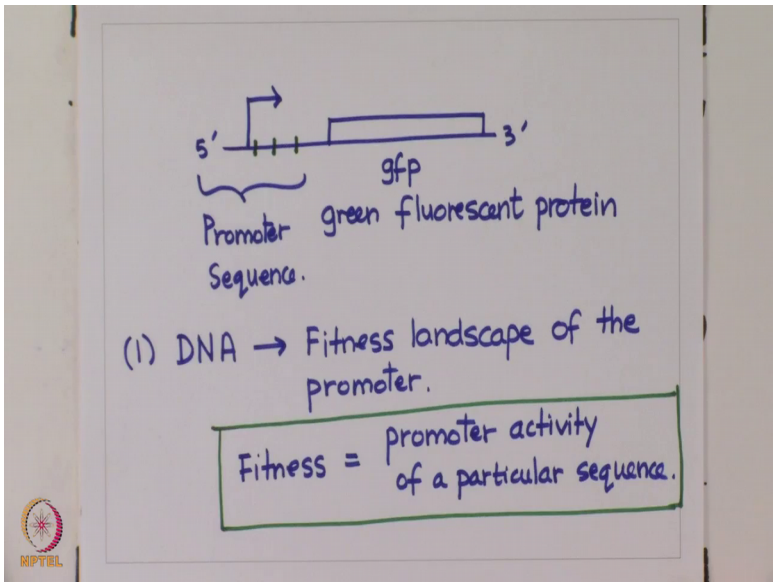


And hence from the start of experiment, so we started the experiment at maximum possible fitness which implies that X naught is equal to 1, which if you look at the previous slide here all the individuals is of the population were carrying this particular genotype. So, the average fitness of the entire culture that I started with was maximum, but as steady state is established, I do this experiment steady state is established, because of mutations taking place in the system, X naught at steady state is less than 1. What that implies is as I moved in this direction from start of experiment to the steady state there is a decrease in fitness.

So, this is an interesting example where what we are showing is that natural selection does not always lead to an increase in fitness associated with the environment that we are doing an experiment in. But you could imagine scenarios where you can imagine scenarios such as this, where actually even you allow the experiment to run its course and allow mutation selection and reproduction to act in concern, you could actually get in to situations where there is a decrease in fitness taking place with time such as this. So, these are just a couple of interesting ways interesting results associated with properties of fitness landscapes.

What we are going to do next is understand that how do actually, how do people actually study fitness landscapes experimentally. And we are going to talk in terms of both DNA and proteins and we are going to take a very simple example associate a very simple example to just get an idea of how can we think of designing x an experiment which would tell us more about the structure of fitness landscapes associated. And the model that we are going to use is going to be a fluorescent protein called green fluorescent protein. And the property of this protein is that when it folds properly and when it is in the right sort of an environment its fluorescence and it gives out green light and hence it appears green and hence the name green fluorescent protein. So, let us define our system in which we will be looking to define experiments which help us visualize these fitness landscapes experimentally.

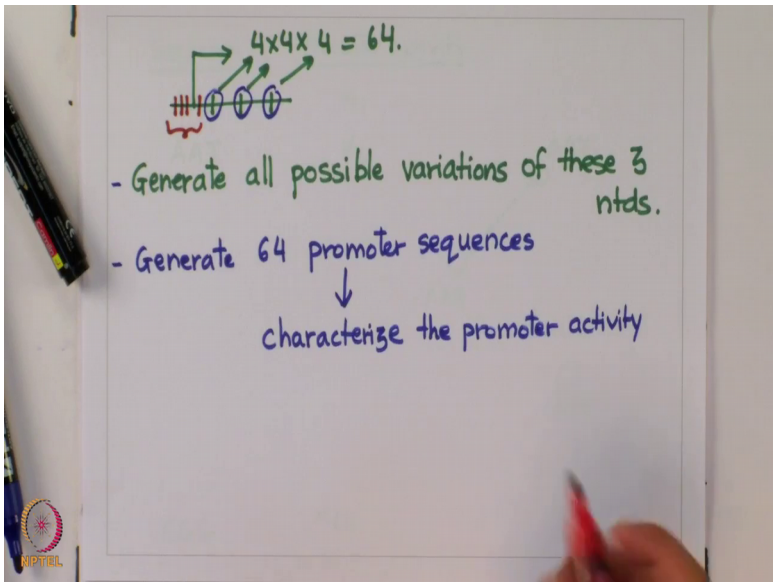
(Refer Slide Time: 10:20)



And the system that we are going to talk about is let us say we have a gene. And this gene is gfp which stands for green fluorescent protein. And there is a promoter sequence here this is the this is 5 prime, 3 prime. This is the promoter sequence where RNA polymerase comes and initiates transcription. So, when we are talking the fitness landscapes, we will use this as our model system, and we will define fitness landscapes associated with both DNA and protein. So, first when we are talking of fitness landscapes associated with DNA, what we are going to be looking at is the fitness landscape associated with the protein of the promoter.

So, I am interested in this bit of the DNA sequence and I want to understand the fitness landscape associated with the promoter. And the way I am going to define fitness here is just equal to the promoter activity that a particular sequence enables. So, my fitness is if I take a DNA sequence, how strongly does RNA polymerase start the process of transcription from that particular promoter is how I am going to define fitness. And let us say in the promoter region of interest I am interested in studying the variation of the promoter activity as I vary this nucleotide, this nucleotide and this nucleotide. Suppose I start off by understanding how does the promoter activity how efficiently or how does the rate at which when RNA polymerase comes and binds this DNA sequence, how does that rate of RNA polymerase starting the process of transcription change as I vary the nucleotides present at these three particular positions.

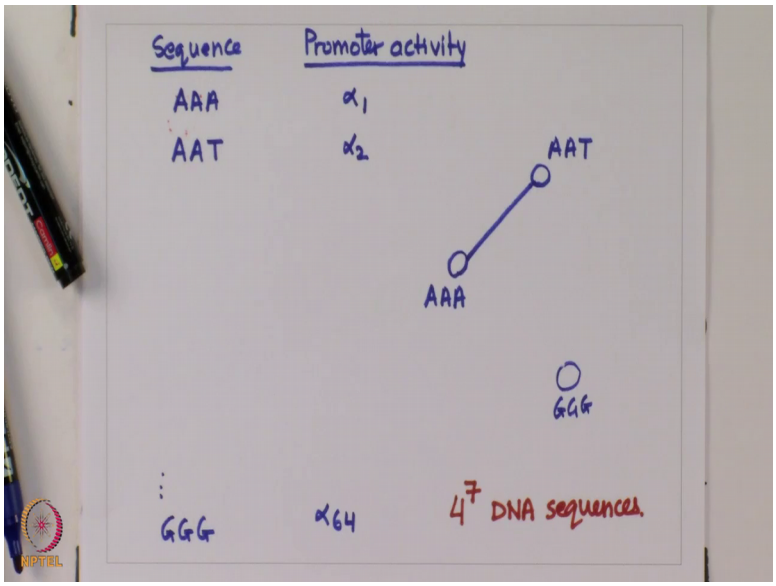
(Refer Slide Time: 13:32)



So, how do we look at that experimentally what we are going to do is we have this DNA sequence and we have these three nucleotides of interest to us. Now, what we are going to do is generate all possible variations of these three nucleotides. And this is just a simple counting problem, the number of possible variations that we can generate is just going to be four options for this one times four options for this one times four options for this one, we have 64 possible sequences which we can generate, but where we vary these three nucleotides, and we keep all other nucleotides constant. We do not change any of the other nucleotides these three are the only ones, which we are interested in studying how do they impact the promoter activity associated with this string this bit of DNA polymer.

So, we what we do is we generate these 64. So, we synthesize or we generate these 64 variants, 64 promoter sequences, and then perform experiments either in vivo or in vitro to characterize the promoter activity corresponding to each of the 64 promoter sequences. And depending on how you are framing your experiment this can be done in vitro or in vivo. Then at the end of this exercise what you will get is a table of promoter activity.

(Refer Slide Time: 15:36)



We can perform these 64 transcriptional experiments and corresponding to each variant of the promoter sequence, we get a corresponding promoter activity number. And we can summarize our results in a tabular form where we have sequence and promoter activity. And remember when I say sequence we are only talking about the three nucleotides which we are wearing here, keeping all the other sequences in the DNA polymer of interest of the promoter region constant.

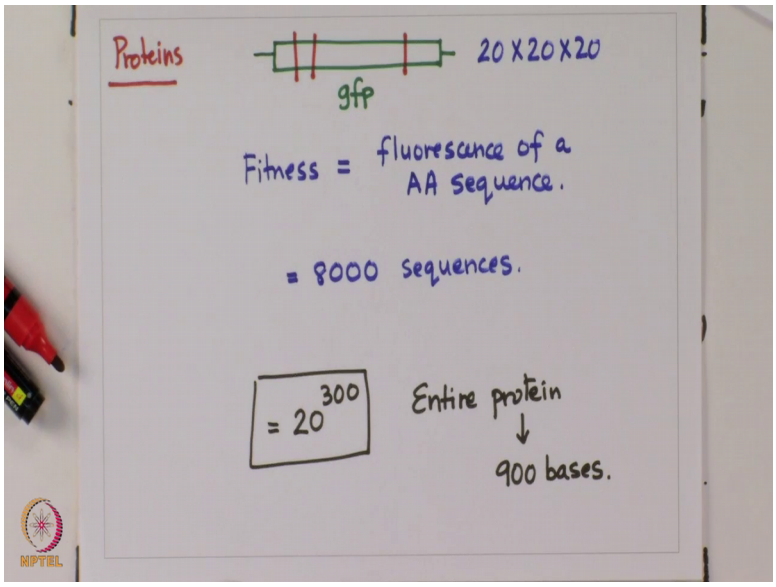
So, what is the first sequence is AAA promoter activity might be measured in sum. So, you have some number alpha, then you have the next sequence AAT, you have its called this alpha 1 called this alpha 2 and so on and so forth and you end up with the 64th one which is say GGG alpha 64. If you remember how we define fitness landscapes each sequence is going to be are node, so this is AAA, this is AAT. And these two are going to be connected, because I can move from one to the other by just changing one nucleotide and keep the other two constant here. So, these two are connected and so on and so forth, we are going to have 64 such nodes and appropriate connections, and this gives me the entire fitness landscape associated with the promoter activity of a particular DNA sequence. And I am only looking at three bases, three nucleotide bases from the entire DNA sequence which I am calling to the promoter here.

The problem with this is. So, this looks like a pretty good way to characterize the fitness landscape associated with the promoter activity, but the problem associated with this is

the scale of the problem the sheer number of experiments that you have to do to able to get complete information about the fitness landscape associated with this promoter for instance. For example, if I decide that I want to characterize the fitness landscape of this promoter not by looking at just these three nucleotides, but there are lots of other nucleotides which I want to study as well.

So, let us say there are these four nucleotides here that I also that I want to study. Now, suddenly from a manageable 64, now I am looking at creating all possible variations from not just these three nucleotides, but seven nucleotides and that immediately changes the scale of the problem. And now experimentally I am going to have to generate 4 to the power 7 DNA sequences that is a fairly large number to synthesize. But even with this number I am only looking at 7 nucleotides in the entire promoter region associated with that DNA polymer. Whereas, if we are talking bacteria promoter sequences can easily be up to a few hundred bases in bacteria which would mean that I am going to have to synthesize 4 to the power a 100 DNA sequences just to be able to get a fitness landscape associated with one promoter in a simple bacterium. So, that is the problem that is associated with finding these fitness landscapes associated with DNA.

(Refer Slide Time: 19:16)



How does all of this play out when I am talking about fitness landscapes associated with the protein. So, we will go back to the framework that we had started with. And remember that the gene of interest that we are looking at is called the green fluorescent

protein or gfp. And now I am interested in a similar exercise that I am interested, so this is for proteins node. And now I am interested that what role does this, this and this residue have in controlling the property associated with this protein. And the way I am going to define this is that the corresponding fitness now is given by the fluorescence of a sequence. This is an (Refer Time: 20:03) sequence.

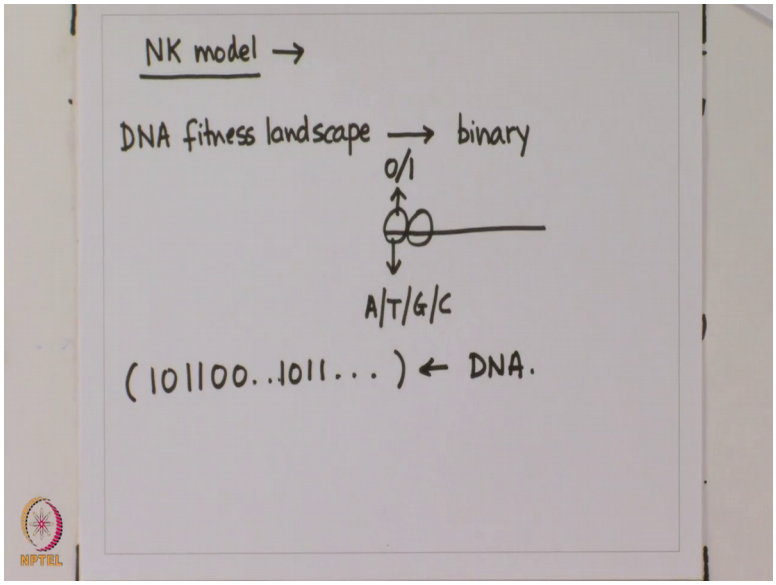
So, because I am changing residues here, the resultant protein that I am going to get could be more or less fluorescent than the original I mean amino acid sequence that I started with. And hence I am relating that fluorescence from a particular sequence equal to fitness of that particular amino acid sequence. So, if I want to study these three residues, and find out the corresponding fitness landscape then I am going to have to generate a number of variant. And that number of variants is going to be now 20 into 20 into 20, because I have choice of 20 amino acids for each of the residues and that comes out to about not about exactly 8000 sequences. Again that means, I have to synthesize lot of protein sequences just to get an estimate of the fitness landscape associated with three residues in a gfp protein.

Suppose, I was to try and attempt to get a fitness landscape associated with the entire protein then we are going to have to change every residue at every position in the entire sequence of that protein. And gfp is a molecule which has about 300 amino acid sequences so that means, if I want to find out the fitness landscape associated with the entire sequence of the protein the number of sequences that I have to generate and then analyze is going to be equal to 20 to the power 300. So, this is for the entire protein and this is a protein which we are saying is 900 bases which is approximately what g f p's, g f p's slightly shorter than that.

So, what this tells us is that finding these sequences creating these variants experimentally and finding the corresponding fitness landscape associated with those variations is not a trivial exercise. And it is not trivial, one the major reason is because of the sheer scale of the problem involved, and the second reason is that you may not have such a quantifiable as a such as fluorescence very easily doable enzymatic assay that might tell you an idea about fitness of a particular protein sequence. So, that is hence determining fitness landscapes experimentally is a difficult exercise. But there are attempts to understand them mathematically when people have been trying to create frameworks in which to analyze fitness landscapes. And we are going to discuss a very

simple and a very famous mathematical framework, which people have been using since the 80s to define and visualize fitness landscapes in their heads.

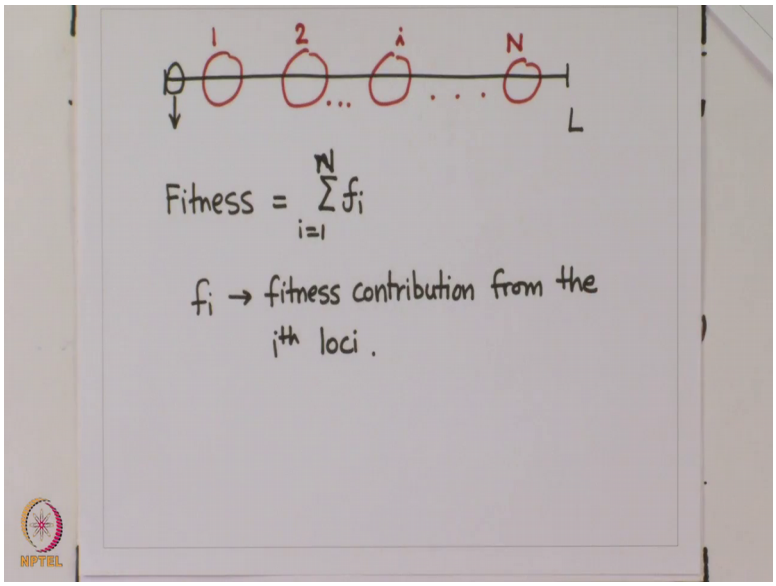
(Refer Slide Time: 23:25)



And this model is called NK model which was proposed in the 1980s by somebody called Stuart Kauffman. The first thing that this model assumes is that we are going to these are DNA well one fitness landscapes. And the first thing that it assumes is that DNA is going to be treated as a binary molecule. So, DNA is binary in nature and what that means, is that at every position on the string. So, if this is the position on the string, whereas we know that on an actual DNA string this position could be occupied by any of the four nucleotides A, T, G, C the big assumption that this model makes is that according to this, this could only take a value of 0 or 1. So, we are going to change the representation of DNA entirely and represent DNA as a binary string.

So, a DNA string for instance could be 1 0 1 1 0 0 1 0 1 1 anything that is a DNA string in this one binary representation. Then now if you are interested in fitness landscape associated with a stretch of DNA, what this model will assume is that the entire fitness associated with a string of length L of DNA is equivalent to sum of contributions of fitness from n different loci are which are present on this sequence.

(Refer Slide Time: 25:18)



So, let me make that clear what this model says is that if you have a DNA string of length L. Now what we are interested is we are interested in is finding the fitness of this particular sequence. And the model what the n k model will say that the fitness is just a submission of fitness contributions from a loci, i going from 1 to n we make that capital N. What the n k model will say is that the fitness associated with the entire DNA length DNA polymer of length L is just equal to summation of f i, where f i is the fitness contribution from the i th loci.

So, you could you could visualize this a certain number of ways you could visualize it one way that n is equal to L. So, every nucleotide on this sequence has a fitness contribution. And what you are doing is here is that saying that the fitness associated with this length of DNA is equal to fitness contribution of every single nucleotide. And you add them up this is just a simple addition you add fitness contribution of first nucleotide f 1, fitness contribution second and so on and so forth go up to L and that gives you the fitness of this entire chain of DNA polymer. Or you could think that there are actually n number of another way to think about this is that there are n number of genes which are present on this stretch of DNA. And the fitness of this stretch of the entire length of DNA is just the summation of fitness contributions from each of the n loci and each loci now corresponds to a particular gene which is present on this strand of DNA.

So, you add the relative contributions from each gene and that gives you the total fitness associated with the length of this DNA polymer. So, this is a very crude estimate on how to visualize fitness landscapes and the whole point of this exercise n k model is not so much to find out the exact nature of fitness landscapes, but try and understand the generalities of the behaviors associated with fitness landscapes. And see if we can capture those properties in a mathematical framework that can be analyzed for other purposes then. So, will we will discuss NK models in some detail in the next lecture. We stop this one for now.

Thank you.