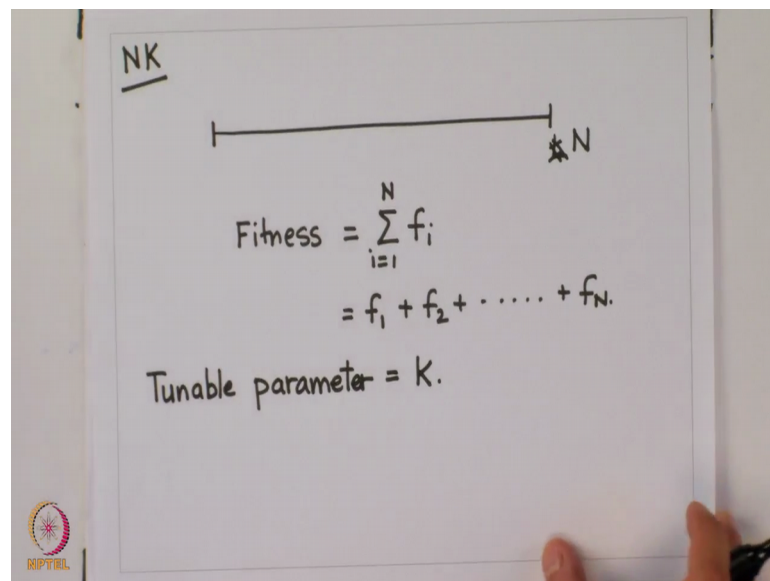


Introduction to Evolutionary Dynamics
Prof. Supreet Saini
Department of Chemical Engineering
Indian Institute of Technology, Bombay

Lecture - 18
NK model of fitness landscape

Hi, let us continue our discussion on NK models of fitness landscapes. So, will pick up from where we had left off in the last lecture, that an NK fitness landscape can be estimated first of all as a binary string of 0s and 1s. And then if this is a DNA of length let me not call it L but N then fitness of this of this DNA sequence can be defined as relative contribution; as the summation of relative contributions from each of the nucleotides associated with this sequence.

(Refer Slide Time: 00:30)



NK

_____ N

$$\text{Fitness} = \sum_{i=1}^N f_i$$
$$= f_1 + f_2 + \dots + f_N$$

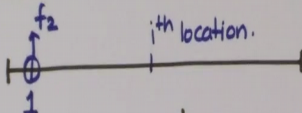
Tunable parameter = K.

So, the total fitness is just f_1 plus f_2 going all the way up to f_N . So, this is pretty straight forward, but what this model allows you to do is that in addition to this it gives you a tunable parameter which is called K. So, remember these are called NK models, this is where the N comes from and the tunable parameter is called K.

And we will try and understand that what does this represent, and what this represents is something which is experimentally observed very frequently in fitness landscapes which is called Epistasis.

(Refer Slide Time: 01:54)

Epistasis :




f_2

i^{th} location.

$$\text{Fitness} = \sum_{i=1}^L f_i$$
$$= f_1 + f_2 + \dots + f_N.$$

i^{th} location $\rightarrow 0 \quad \alpha$

$\rightarrow 1 \quad \beta \quad \alpha \neq \beta.$



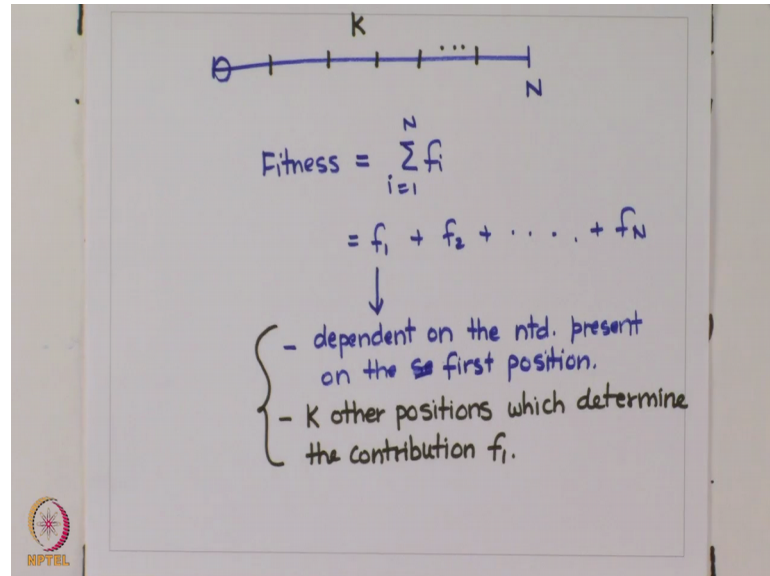
Very often when you are experimentally trying to determine a DNA or protein fitness landscapes you will notice that, so N K model will say that fitness for a particular sequence is the sum of f_i , i going from 1 to L and this will play out to f_1 plus f_2 going all the way up to f of N . But what you will find is that the contribution that f_2 makes will be often dependent on. So, let us say this is the second position, this nucleotide makes a contribution of f_2 , but the contribution that f_2 makes will often be dependent on what is present at the i -th location.

If the i -th location or the i -th position on DNA in our framework is a 0, f_2 might be a certain value, but if the i -th position is a 1 f_2 will take a different value. So, for instance let us say the relative contribution of f_2 towards the fitness of this DNA string is dependent on what is present on the i -th location. If the i -th location is a 0 then the contribution of f_2 , contribution of the second nucleotide here which let us say is equal to 1 then the contribution of this 1 is α , but if the i -th location has a 1 then the contribution of that same 1 present on the second position is a β and α is not equal to β .

This phenomena that the fitness contribution of a particular nucleotide or loci on the DNA is dependent on what is present elsewhere on the DNA is called Epistasis and it is a widely observed phenomena when it comes to fitness landscapes. And that leads us to the

tunable parameter K which is associated with NK models and what K allows us to say what K represents is.

(Refer Slide Time: 04:28)



So, again for sake of clarity let us redraw this is length N and fitness is equal to $\sum_{i=1}^N f_i$, i going from 1 to N , this is equal to $f_1 + f_2 + \dots + f_N$. What the tunable parameter K represents is that f_1 will obviously, the relative contribution to the total fitness that comes from the first nucleotide is called f_1 and f_1 will obviously, be dependent on the nucleotide i will call this a nucleotide even though we are dealing with binary sequences 0 or 1.

f_1 will depend, will obviously, depend on the nucleotide present on the second position sorry f_1 is the first position, so on the first position, but in addition to that what the parameter K represents is that it allows you to incorporate K other positions which determine the contribution f_1 , contribution f_1 . So, what this says is that the contribution that comes from the first nucleotide is not just dependent on what nucleotide is present on the first position, but is also dependent on K other positions elsewhere on this piece of DNA. So, these are K other nucleotides else located these could be located anywhere on the piece of DNA that we are talking about. Together these K plus the nucleotide whose contributions we are talking about here these all these put together determine what is going to be the quantity f_1 .

So, the relative contribution of the first nucleotide is not just dependent on which nucleotide is present on the first position, but also dependent on these K other nucleotides which are present elsewhere on the chromosome or the string of DNA that you are talking about. That and you can tune K and incorporate epistasis because now this is our first step towards understanding how do we incorporate epistasis in our fitness landscapes. An epistasis as I had mentioned it is a widely observed phenomena in all fitness landscapes that we observed. What this allows us to do is make the fitness contribution from the first nucleotide, contingent up on what is what are the nucleotides present at other locations on the piece of DNA and hence incorporate epistasis in to our fitness landscape models. What happens when I choose K to be equal to 0? Think about the special case for a little bit.

We are saying that the fitness contribution of a particular nucleotide is dependent on what nucleotide is present at that location that I am talking about. So, if I am talking about contribution of the first nucleotide it; obviously, depends on what is that first nucleotide is it a 0 or is it a 1, plus it is also dependent on K other nucleotides which could be present elsewhere on the chromosome. Now what if K is equal to 0? If K is equal to 0, what I am saying is that the fitness contribution of the first nucleotide is dependent on itself and no other positions on the chromosomes. So, this is only dependent on itself no epistasis.

(Refer Slide Time: 08:46)

$K=0$
 ↓
 No epistasis.

$\text{Fitness} = \sum_{i=1}^N f_i$
 $= f_1 + f_2 + \dots + f_L$
 ↓
 only dependent on the ntd. at position 1.

$f_0 = 1$
 $f_1 = 2$

DNA seq. = 10110
 $= 2 + 1 + 2 + 2 + 1 = 8$

sequences → 2^5

NIPTEL

So, K equal to 0 is a special case which says that again that same picture the contribution of the first nucleotide fitness is equal to $\sum_i f_i$, i going from 1 to N equals f_1 plus f_2 so on and so forth plus f_L and this is only dependent on the nucleotide at position 1. So, that is what is represented by K equal to 0. If that is the case then how do I determine the fitness of a sequence of a DNA sequence, what I define is that the fitness contribution of a 0 is let say 1 and fitness contribution of a 1 is let us say 2. These are the contributions that I choose for my DNA sequence and then if I have a DNA sequence which is equal to 1 0 1 1 0, I have this relationship that fitness is equal to contribution of each of these positions. So, fitness of 1 since K is equal to 0, fitness of this 1 in the DNA sequence is only dependent on its own nucleotide and on none of the other nucleotides present on the DNA because K is equal to 0.

Hence I know that the fitness contribution of a 1 is 2 then I have a 0 and because K is equal to 0 the fitness contribution of 0 depends only on itself and fitness contribution of a 0 is defined to be 1 and then another 2 and another 2 and another 1. So, fitness of this DNA sequence it will be 8, in an N K modeled with these fitness contributions and with a K equal to 0. So, that is how you would define the fitness for each of the, for every possible nucleotide sequence of length 5 here which will be the number of sequences here will be 2 to the power 5 and not 4 to the power 5 because we are treating DNA as a binary polymer.

So, you find the fitness corresponding to each of the this is 32, each of the 32 sequences and then connect represent do the same exercise that we have been talking about, draw the nodes, draw the edges, connect them if they can be changed in to one another via single mutation and understand the structure associated with the fitness landscape that you get. So, this K is equal to 0 is a fitness landscape which you get with no epistasis. What happens if we incorporate epistasis?

(Refer Slide Time: 12:15)

$K=1$

\Rightarrow - that nucleotide (itself)
- one more ntd. located elsewhere.

$\text{Fitness} = \sum f_i = f_1 + f_2 + \dots + f_N$

I.

$f_{00} = 0$
 $f_{01} = 1$
 $f_{10} = 1$
 $f_{11} = 2$

Seq = 10110
 $F = 1 + 1 + 2 + 1 + 1 = 6$

II.

(ntds. 1&2) ntds 2 & 3

NPTEL

So, let us again choose K equal to 1, what that implies that the fitness contribution of a nucleotide depends on that nucleotide which is itself and also one more position, one more nucleotide located elsewhere ok.

So, in this case the fitness is equal to $\sum f_i$ which is f_1 plus f_2 and so on and so forth going up to f of N , if N is the length of the DNA that I am interested in. But now f_1 is not just dependent on what is the nucleotide that is present on the first position, that is present on the first position on this string, but is also dependent on one more DNA nucleotide elsewhere on the chromosome.

So, if I have a DNA string such as this, this is position number 1; for sake of simplicity I will define that the f_1 contribution is dependent on what is present on position number one and that second position elsewhere on the chromosome which the fitness contribution of this is decided by is just the neighboring nucleotide. So, f_1 is dependent on what is present as nucleotide 1 and what is present as nucleotide 2. So, f_1 is a function of nucleotides 1 and 2.

Similarly, f_2 which is the relative contribution to fitness from the second nucleotide will be dependent on nucleotides 2 and 3 its neighbors. So, these 2 go together to determine the fitness contribution from this base, these 2 go together to determine the fitness contribution from this nucleotide and so on and so forth. So, we need rules to define this since fitness contribution of each is dependent on 2 nucleotides now, I am going to have

to define that the fitness contribution whenever you have a 0 0, is let us say 0 whenever you have a 0 1 its 1, whenever you have a 1 0 its 1 and whenever you have a 1 1 you have 2. If that is my definition and the sequence that I am interested in is 1 0 1 1 0 then the fitness of this sequence can be defined as when I look at the fitness contribution of this first nucleotide one that is not solely decided by itself because K is equal to 1 that is decided by this 1 and the neighboring nucleotide 1 0 and contribution of 1 0 is equal to 1. When I look at the contribution from the second nucleotide that is a 0, that is decided by this 0 and the neighboring nucleotide which is 1 and contribution of a 0 1 is 1.

Similarly, what contribution for this third nucleotide will be decided by this 1 1 which is 2, for the fourth one will be one decided by the sequence 1 0 which is 1 and for the last one since this is a 0 and no other nucleotide follows this one we go back to the first one and decide that the contribution of this 0 will be decided by this 0 1 sequence and that is equal to according to our definition equal to 1, this is equal to 6 now. So, in this way you can define K equal to 1 and then come up with nucleotide and therefore, define the fitness associated with this particular sequence having chosen K equal to 1 and incorporated epistasis in to your fitness landscape.

This again, this the length of the sequence that we have been looking at is equal to 5 and hence just as in the previous case. In fact, this is the same sequence just as in the previous case they are going to be 32 sequences of length L equal to 5 and then what you do here is use this relationship, these assignments of relative contributions of f 0 0 1 0 0 1 and 1 1 to find the corresponding fitnesses of all 32 sequences and then draw the edges and nodes the same way that we have been talking about and get the corresponding fitness landscape when there is epistasis incorporated in to the model. And you can go to higher complexities you can increase epistasis in the model by choosing higher values of K .

So, the maximum possible epistasis that you can incorporate here is when contribution of one nucleotide depends on every other nucleotide on the DNA strand that you are looking at, if that is the case then every single nucleotides fitness contribution is dependent on the entire DNA sequence and that is the maximum amount of epistasis that you can introduce in your model. We have somewhat randomly chosen we have 2 aspects according to the definition that we have talked are chosen somewhat randomly.

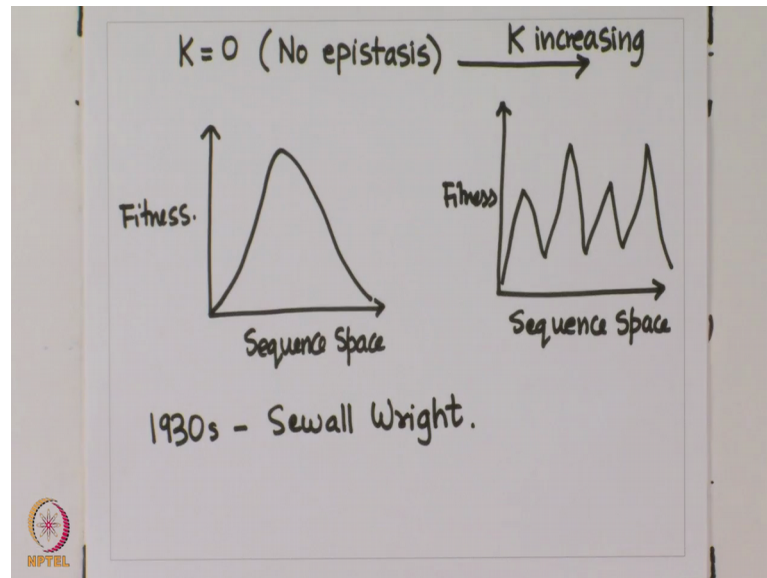
The first one if we look here on our slide for K equal to one is these fitness definitions these are somewhat randomly chosen. So, this is aspect one and the second one is that when I say when I choose K equal to 1 what I am trying to say is that fitness contribution of the first nucleotide depends on itself and any one of the other N minus 1 nucleotides. Somewhat randomly I have chosen that one other nucleotide to be the next one for every nucleotide in my system that. So, this choice of this one other nucleotide which determines the fitness of a particular nucleotide is the second aspect which is somewhat randomly chosen in our definition here.

But the point that I want to emphasize here that these definitions the NK models do not claim to define fitness landscape of a particular protein or a DNA sequence in its precise details. But what they do allow us to do is visualize and analyze the properties of fitness landscape in general by giving us these mathematical structures which allow us to define these fitness landscapes mathematically and do further analysis on. What they also help us with is they give us this tunable parameter K and for instance if we know experimentally that one particular system has a large degree of epistasis according associated with it then we can try and capture the properties associated with the fitness landscape of that particular experiment by choosing NK model with a somewhat higher value of K .

On the other hand if we know the experimentally that this system does not seem to have very much epistasis then we can use the same NK model, but choose the K value to be 0 or 1 and try and capture the properties associated with the fitness landscape of the system that we are trying to study experimentally. And this is important because as we have seen trying to determine fitness landscapes experimentally is a very very complex task which is often beyond the (Refer Time: 20:33) of experimental constraints.

One more thing about these about these landscapes that we want to discuss the NK model is that what sort of fitness landscapes do they give us.

(Refer Slide Time: 20:50)



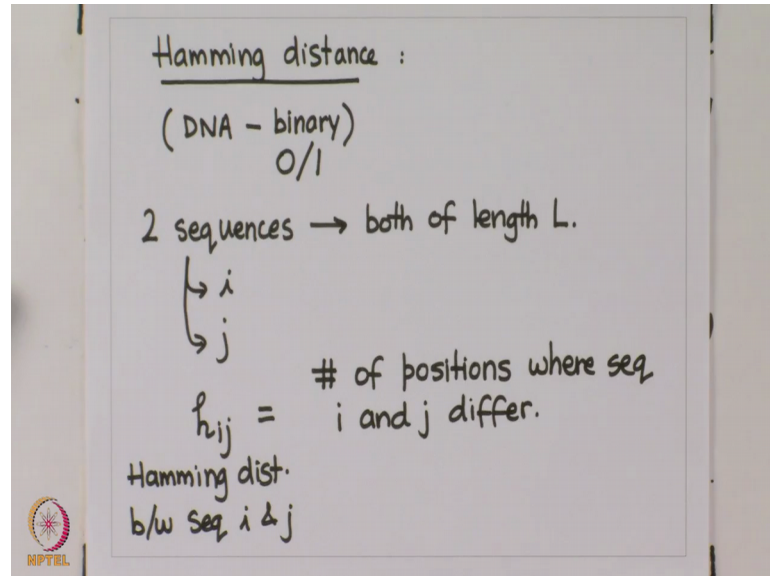
So, we have when we have K equal to 0 which represents the case of no epistasis, the fitness landscape that we will get will be very smooth one, single peaked and a smooth fitness landscape. So, this is sequence space, this is fitness. As I increase K , as we saw from the slides in the 2 lectures back we could have smooth fitness landscapes or we could have rugged fitness landscapes and as we increase K , as we introduced epistasis in to our model we get in to these rugged fitness structures where we have multiple peaks and this will be the fitness landscape associated with increased value of K .

I want to emphasize that fitness landscapes are high dimensional representations and you cannot sort of plot them on a piece of paper. So, these are just cartoon representations of sequence spaces in sequence spaces in one dimension in reality these are L dimensional objects which are impossible to represent on paper. So, this sort of concluding this idea of fitness landscapes and epistasis, these idea of fitness landscape was first proposed in the 1930s by someone called Sewall Wright and he has proposed this as a mathematical concept and its really in the last few decades that people have that experimentation has got to such a degree of sophistication, that allows us to capture the structure of these fitness landscapes experimentally.

Let us continue our discussion on these binary sequences because these help clarify ideas about fitness landscapes and give us some very useful insights in to what could happen

when populations are moving on these landscapes. And one of the quantities that we are going to make use of is what is called a Hamming Distance.

(Refer Slide Time: 23:33)



Again we are going to be treating DNA as a binary string. So, DNA is just a sequence of 0s and 1s, and if we have 2 sequences both of length L let us call them sequence i and sequence j , the hamming distance between these 2 sequences is represented by a variable called h_{ij} . So, this is hamming distance between sequence i and j is equal to number of positions where sequence i and j differ from each other.

So, if you are talking of a sequence of length L all you are interested in when you are computing hamming distance between 2 sequences. So, by definition hamming distance exists between 2 distinct species you do not have a hamming distance for one sequence hamming distance when you trying to quantify this hamming distance between these 2 sequences they should be of equal length L then what you do, all you do is compare base by base every position in these 2 sequences and count the number of instances where the base is incorporated in these sequences are distinct. So, let us help get clarified with the help of an example.

(Refer Slide Time: 25:55)

$i = 101010$
 $j = 110010$
 $h_{ij} = 1 + 1 = 2.$
 $\underline{=}$
Prob. that i was replicating \rightarrow sequence j ?
 $\underline{=}$ Mutation rate = u (Prob that a base is copied incorrectly)
Length = L
 $h_{ij} = \alpha.$

So, let us choose sequence i to be 1 0 1 0 1 0 and let us choose sequence j to be equal to 1 1 0 0 1 0 and the hamming distance between these 2 sequences h_{ij} is just equal to the number of positions where the 2 sequences have different nucleotides incorporated in them. So, all we do regarding that is first position is the same so this does not count, this second position is distinct so hamming distance one, third position is distinct hamming distance one fourth, fifth, sixth positions are identical. So, the hamming distance between these 2 positions, these 2 sequences i and j is equal to 2. That is the definition of hamming distance. Again I want to emphasize that hamming distance is only defined for sequences of equal length and you need 2 sequences to be able to compare, hence these 2 2 subscripts here represent the 2 distinct sequences of equal length that we are trying to compare and compute the hamming distance for.

Now, when there is replication happening what is the probability that sequence i was replicating and because there is a finite mutation rate associated with the process of replication the progeny that was produced was actually sequence j . What is the probability that this happens given that the mutation rate is equal to u which is probability that a base is copied incorrectly is u and the length of the sequence is L and the hamming distance between i and j let us say is equal to α .

If these things are given to you that the probability when the sequence is being replicated makes a mistake at a particular base is u , length of the 2 sequences that we are trying to

compare is L and the hamming distance between these 2 sequences i and j is α , what is the probability that when i replicates what you end up as progeny is actually sequence j .

I would suggest that you try and work this yourself and we will pick the next lecture of starting with this calculation itself. Let us stop this lecture for now.

Thank you.