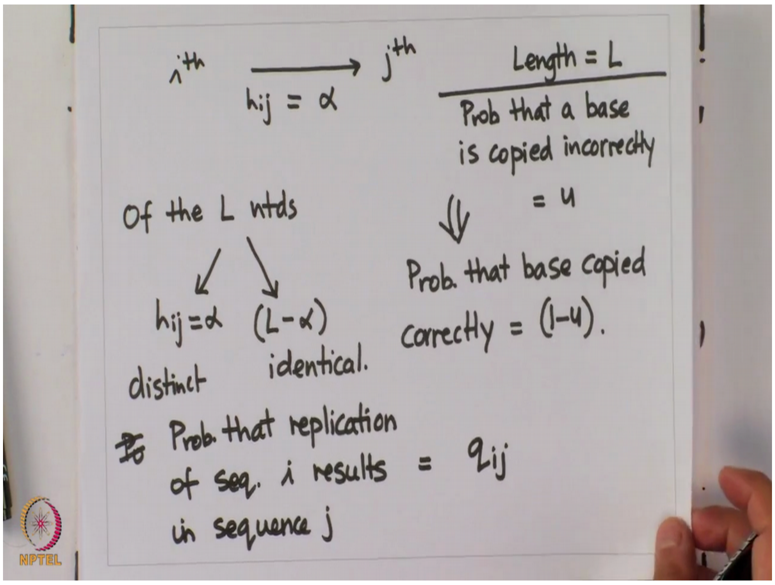
Introduction to Evolutionary Dynamics
Prof. Supreet Saini
Department of Chemical Engineering
Indian Institute of Technology, Bombay

Lecture - 19
Modelling evolution on fitness landscapes – 1

Hi, and let us continue our discussion on hamming distances in binary DNA sequences. And where we had stopped in the last lecture was this calculation that we are interested in given a sequence i, which is replicating and the replication system has an error rate u associated with it. The length of the sequence is given as L, what is the chance that when this i-th sequence replicates; the resultant sequence is another given sequence j whose hamming distance with the original sequence is h ij equal to alpha was what we had ended up with last time.

(Refer Slide Time: 01:02)



$i^{th} \xrightarrow[h_{ij} = \alpha]{} j^{th}$

Length = L

Prob that a base is copied incorrectly $= u$

$\Downarrow$

Prob. that base copied correctly $= (1-u)$.

Of the L ntds

$h_{ij} = \alpha$ distinct

$(L-\alpha)$ identical.

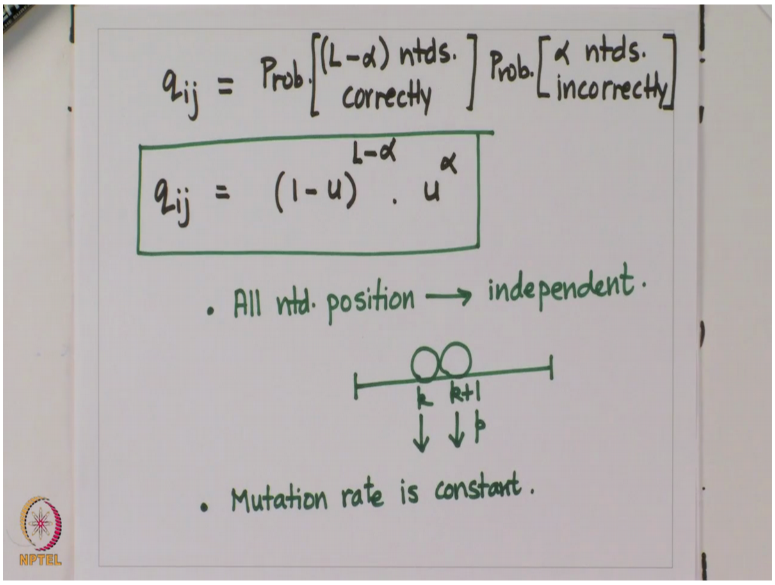
Prob. that replication of seq. i results in sequence j $= q_{ij}$

So, we have the i-th sequence replicating giving the j-th sequence, the hamming distance between these 2 is equal to alpha, the length of these 2 DNA sequences; binary DNA sequences is L and the probability that a base is copied incorrectly is given as u which means; what this implies is that probability that base copied correctly is equal to 1 minus u or because when a base is copied; either it is copied correctly or incorrectly. Hence, this is just 1 minus u.

So, this probability what this means is that of the L of the L nucleotides in these 2 sequences h ij are distinct these are distinct and L minus alpha are identical we do not know the precise location of where the differences between the i-th and the j-th sequence lie we do not know which are the basis which are which are the same in nucleotide sequence and which are the exact alpha positions where the 2 sequences are dissimilar, but we do know their numbers. And we do know that if i is being replicated to j. If i is being replicated and the progeny is of sequence j then the replicating machinery has to make errors at very precise locations. So, as you introduce the same errors which would convert sequence i in to sequence j.

So, we do know that of the total number of L nucleotides that are present in these 2 sequences L minus alpha have to be copied correctly let us call this probability; probability that replication of sequence i results in sequence j, let us call this probability q ij. So, we are going to try and estimate and develop an expression for this a quantity q ij.

(Refer Slide Time: 04:16)



$$q_{ij} = \text{Prob.}\begin{bmatrix}(L-\alpha)\text{ ntds.}\\ \text{correctly}\end{bmatrix}\ \text{Prob.}\begin{bmatrix}\alpha\text{ ntds.}\\ \text{incorrectly}\end{bmatrix}$$

$$\boxed{q_{ij} = (1-u)^{L-\alpha} \cdot u^{\alpha}}$$

- All ntd. position → independent.

- Mutation rate is constant.

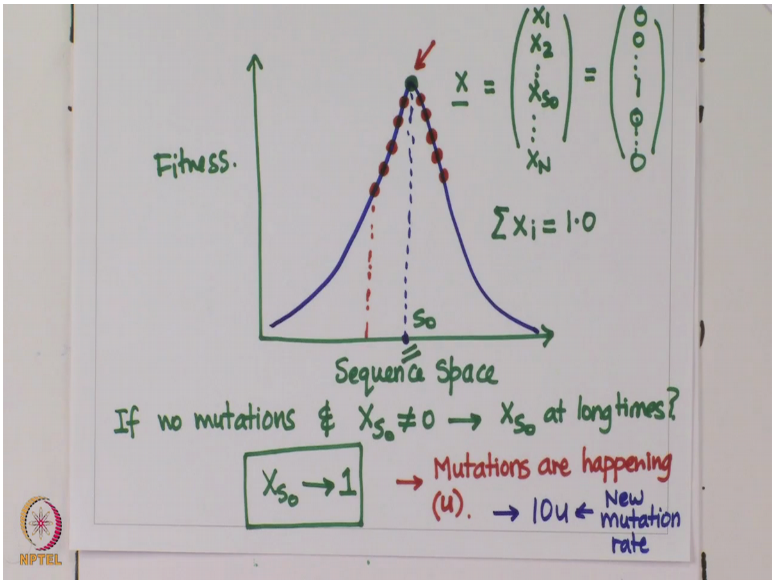
Now, q ij probability that i upon rap replication gives sequence j is probability that L minus alpha nucleotides were copied correctly and probability that alpha nucleotides were copied incorrectly, this is where the binary nature of the DNA string that we have assumed comes to help us because sequence i and j differ at alpha number of positions and if DNA was not being treated as a binary and suppose at a particular position sequence i had and sequence j had a t, then if a was copied incorrectly in sequence i then the incorrectly copied progeny could have had t g or c.

But because we are interested in only generating sequence j as a progeny; that means, we want a to be copied to t then; that means, we are we are interested in computing the probability that of all the errors that could have happened which could have happened three base a going to t a going to g and a going to c. We are only interested in one third of the errors that happened if DNA was treated as it exists in nature, but what the binary nature of this DNA in our treatment allows us to do is that if 0 is copied incorrectly it automatically means that the progeny has a one there because there is no other option available to it.

But even if DNA; if DNA was treated as comprised of 4 different type of basis all it all that would change in our analysis is we would introduce that factor of three associated with it, but the rest of the analysis remain the same and hence it is not, it is not too far a leap of faith to treat DNA as a binary sequence and do all of this analysis. So, L l minus alpha nucleotides need to be copied correctly the probability that that one nucleotide is correct copied correctly is 1 minus u the probability that L minus alpha are copied correctly will be L minus alpha 1 minus u raise to power L minus alpha and probability that alpha nucleotides are incorrectly copied would just be u to the power alpha. That is the expression that we have for q ij what this assumes is the there are there are some assumptions that have gone in to development of this expression the first one is that all positions all nucleotide positions and their replication are independent events.

So, whether if this is the length of the DNA this is position k this is position k plus 1 whether while copying k plus 1 and error is made or not this probability is independent of whether an error was made here or not these are independent events a, the second assumption that we have made here is that the mutation rate is constant throughout all. So, that that gives us some idea of how to compute hamming distances between these binary DNA sequences and this is something that we will come back to in our analysis in a little bit we next want to develop result which I think is very very curious and very very; it tells us something very fundamental about a how natural selection and property and organisms have evolved in the context of fitness landscapes.

(Refer Slide Time: 08:53)


Fitness.
S_0
Sequence Space
$\underline{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_{S_0} \\ \vdots \\ X_N \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$
$\sum x_i = 1.0$

If no mutations & $X_{S_0} \neq 0 \rightarrow X_{S_0}$ at long times?

$\boxed{X_{S_0} \rightarrow 1}$ → Mutations are happening (u). → 10u ← New mutation rate

For that let us let us revisit our fitness landscape example. So, again in this highly; highly simplified one dimensional sequence space if y axis is the fitness and I have a single peaked fitness landscape this is the sequence which corresponds to the maximum fitness on this landscape the question is that if we were to start anywhere the question is that if there were no mutations. And let us call this sequence s naught and the x corresponding to s naught was not equal to 0 what would be X s naught at long times that is the question that that we want to start this discussion with just is pause the video think about this for 15 seconds and see if you can come up with an answer for that.

So, because we are saying no mutations X of s naught is not equal to 0 the purpose of doing this is to make sure that the initial number of individuals which belong to this particular genotype which corresponds to maximum fitness on my landscape this number of individuals is not equal to 0. Hence, there is a finite number of individuals to start the experiment which correspond to maximum fitness and because there are no mutations genotypes do not go from one to another each grows at their respective; respective speed and eventually selection will decide that the fittest individuals replace all other types of genotypes which are present in the environment and eventually X of s naught would approach one. If we are looking at the vector which defines the makeup of the population at long times this was again something that we did other couple of lectures back we can define the makeup of the population at any given time by this vector x then at long times what we are saying that ins in a scenario as defined by this vector would take the value 0 0 1 0 zeros all the way.

Because since x is represent fraction of the population which belong to a particular genotype sigma x i has to be equal to 1 and selection in the case of no mutations selection decides that all individuals belonging to the population are at genotype s naught at long times. So, the entire population is right here from here I want to change the setting a little bit and say that now mutations are happening at rate u and the probability that a mutation is going to take place when one base is copied in a sequence is u this is the same u that we had defined in the last slide that probability that nucleotide is copied incorrectly is u.

So, now, from our 0 mutation rate we move to a nonzero mutation rate and I want you to imagine that what could be the steady state population in such a system again take a few seconds and think about this and what should what should really come to your mind is now the population will not be, obviously not be at the maximum fitness the genotype corresponding to the maximum fitness only. But there will be a spread around the peak perhaps the majority of individuals would still be at a genotype corresponding to the maximum fitness, but because of this mutation rate this there will be a steady state spread about the peak of this fitness landscape this is what we had done in quasi species analysis that we did a couple of lectures back.

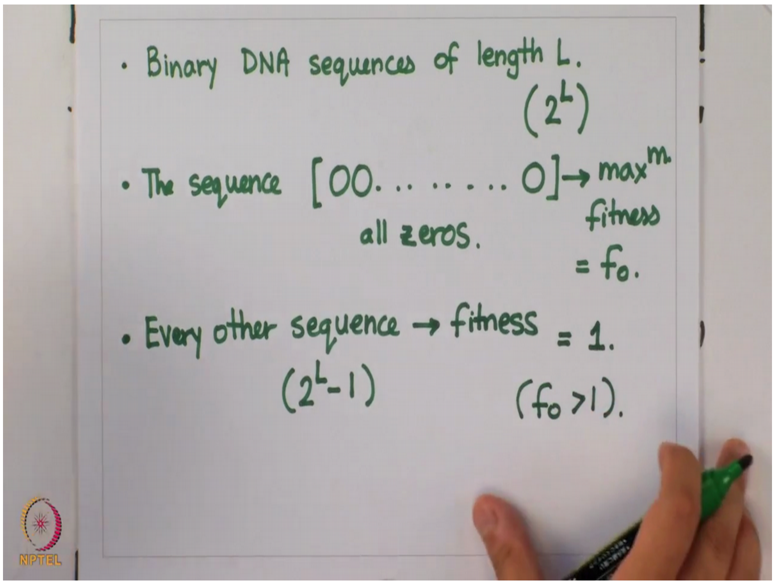
Now we want to go even further what happens if mu mutation rate are now 10 times mu this is my new mutation rate what would happen at the steady state in a scenario such as this where if this was the steady state corresponding to mutation rate new. Now the new mutation rate is 10 times mu what would be the corresponding steady state again think about this and see if you can develop an intuitive picture as to what should happen at steady state.

Now, what should come to mind is that the distribution of the population around the peak as mu increases is gets widened and now you have more genotypes available at steady state which are coexisting with the fittest genotype that such that is defined by the sequence s naught. So, in this case of mutation rate equal to 10 mu the spread at steady state would be even wider. And perhaps you would see individuals corresponding to genotypes as far away from s naught as depicted here and the idea being that as mutation rate increases this peak where individuals are present at steady state would keep getting wider and eventually we want to understand that what happens if mutation rate keeps on increasing like this keeps on increasing like this.

What is the eventual fate of the population, would it always remain centered around the peak of the fitness landscape or does something dramatic happen to the steady state population distribution in a fitness landscape such as this. And to develop that we are going to we are going to define a framework using the things that we have discussed today binary DNA sequences where error rates are happening with probability mu per base and we have hamming distances of h ij between sequences i and j.

So, we will spend the next few minutes trying to develop this framework as to how can we answer this question that what happens at large mutation rates whether the whether the structure of this population of being centered around the peak. And some mutants coexisting with the fittest sequence available in the fitness landscape remain order something catastrophic happen and a qualitative change in the nature of distribution of these individuals at steady state take place at some value of mutation rate.

(Refer Slide Time: 16:47)

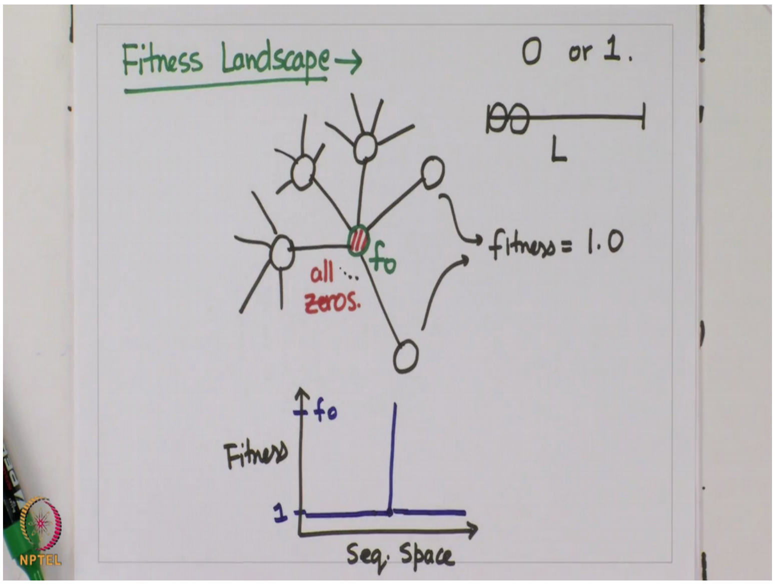


- Binary DNA sequences of length L.
  (2^L)
- The sequence $[00.\,.\,.\,.\,.\,.\,.\,0] \rightarrow$ max^m. fitness $= f_0$.
  all zeros.
- Every other sequence $\rightarrow$ fitness $= 1$.
  $(2^L - 1)$ $(f_0 > 1)$.

Let us develop this framework and how we are going to start with this is we are going to we are going to consider binary sequences of length. So, these are the definitions which give me the ground work towards formulating this problem quantitatively binary DNA sequences of length L how many such sequence would exist because this is binary for every position, I have 2 options the total number of sequences will be 2 to the power L. Now let us consider that the sequence. One of these 2 to the power L sequences would be all 0s. So, this is all zeros let us say that this sequence has among all 2 to the power L sequences maximum fitness.

So, this is the sequence which has maximum fitness let us call that is fitness equal to f naught we want to define the exact nature of our fitness landscape and we say that every other sequence and their number will; obviously, be 2 to the power L minus 1 except for the all zeros sequence has equal fitness and let us say that fitness is equal to 1. So, every other sequence other than the all 0 sequence has a fitness equal to 1 which implies that f naught is bigger than 1. So, what we have done by these definitions is essentially defined the fitness landscape associated with this problem what would this fitness landscape look like. So, let us try and imagine this fitness landscape.

(Refer Slide Time: 18:57)



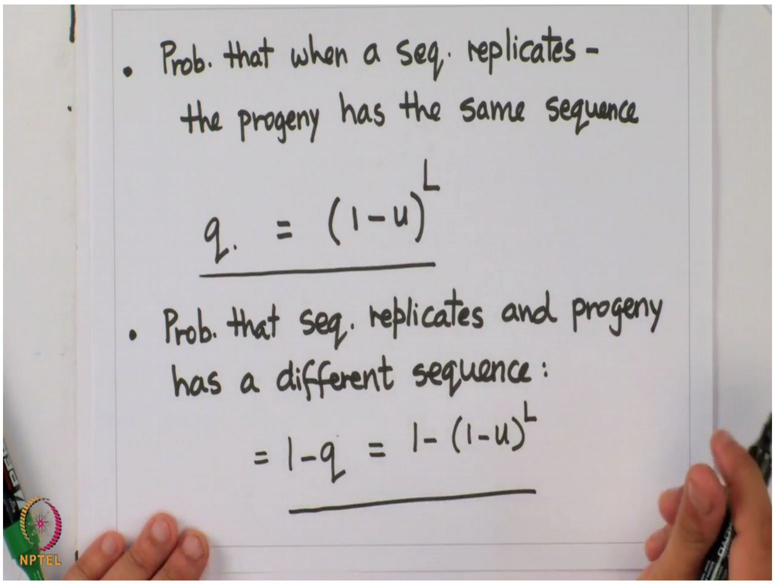
Let us draw at the center the fittest. So, we want to understand the fitness landscape corresponding to these definition let us say this is the node which represents to all zeros hence the fitness is f naught. So, this is my all zeros naught how many neighbors will this node have again may be take a few seconds pause the video and think about the number of neighbors that this node is going to have on my fitness landscape.

The number of neighbors for this particular node is going to be equal to exactly L if length if the length of the DNA that we are interested in is L the number of neighbours that this sequence will have or any other sequence will have we will also be equal to exactly L and the reason is because DNA is binary at every position you either have 0 or you have one by definition is defined as one mutant neighbor such that this acquires a single mutation and converts to its neighbor the neighbor could similarly acquire one mutation and convert to this sequence. So, these transitions are possible with the help of one mutation only and that one mutation could happen anywhere along the length of the DNA that we are interested in.

So, if the length is L that one difference in the nucleotide sequence could be at position one which is this neighbor could be at position 2 which is this neighbor and so on and so forth. So, this has L number of neighbors. So, let us represent these neighbors and all of them have fitness equal to 1 according to our definition and of course, these guys will have their these nodes will have their own neighbors and so on and so forth. And the graph would spread around the peak like this, but the point may and now you can sort of imagine this structure as this peak where there is one point here which has a fitness value of f naught and we know that f naught is more than one and all the other nodes have a fitness value equal to 1. So, this is just a plateau here with this one node which is centered which is centered on this depiction and has the value of f naught associated with the fitness for that sequence. So, that is the landscape that we have a plateau with every node having a value one and a sharp peak with a value f naught; f naught bigger than one for the sake of simplicity we can just represent this like this.

Sequence space and fitness just for the sake of simplicity we can just say that everything has fitness equal to 1, but there is this one node which has fitness equal to f naught.

(Refer Slide Time: 22:47)



- Prob. that when a seq. replicates - the progeny has the same sequence

$$q. = (1-u)^L$$

- Prob. that seq. replicates and progeny has a different sequence :

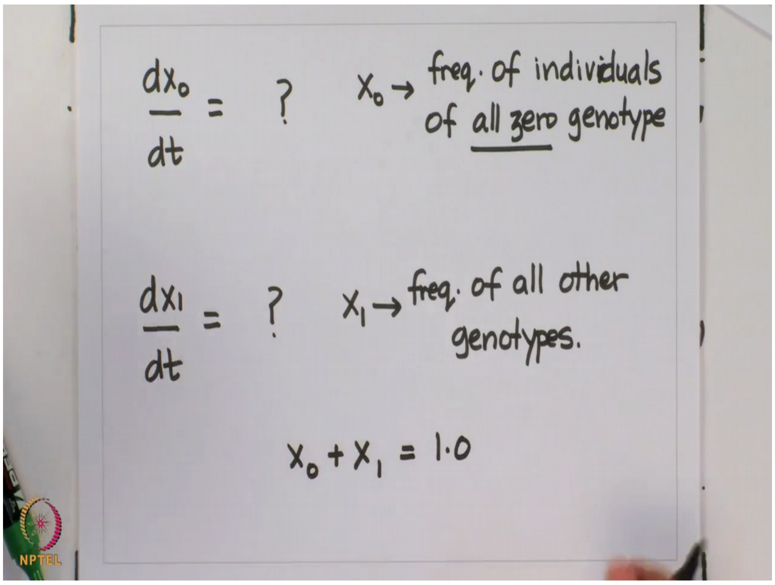
$$= 1-q = 1-(1-u)^L$$

Now, if we have this what is the probability that when a sequence any one of these sequences replicates the progeny is the exact sequence as the parent DNA. So, what is the probability that when a sequence replicates the progeny has the same sequence. That means, during the process of replication no error occurred and this is equal to 1 minus u which means that no error happened during replication for a particular base, but that needs to happen for all L basis which is the length of the DNA sequences that we are talking about. So, the answer for this is 1 minus u to the power L that is the probability that replication was identical and mutation did not happen.

Let us call this probability q what is the second quantity that we are interested in is what is the probability that sequence replicates and progeny has a different sequence and that probability if this is the relationship for the fact that progeny has the exact same sequence as the parent the probability that the progeny has a different sequence is just equal to 1 minus the probability that the progeny and the parent have the same sequence.

So, this probability is just going to be given by 1 minus q which is 1 minus 1 minus u to the power L. You should convince yourself that these are all positive numbers. Now we have the probabilities associated with the fact that given the mutation rate when division happens what is the probability that I am creating the progeny which is identical to my sequence and what is the probability that I am creating a progeny which is dissimilar from my sequence and is the mutant progeny.

(Refer Slide Time: 25:26)



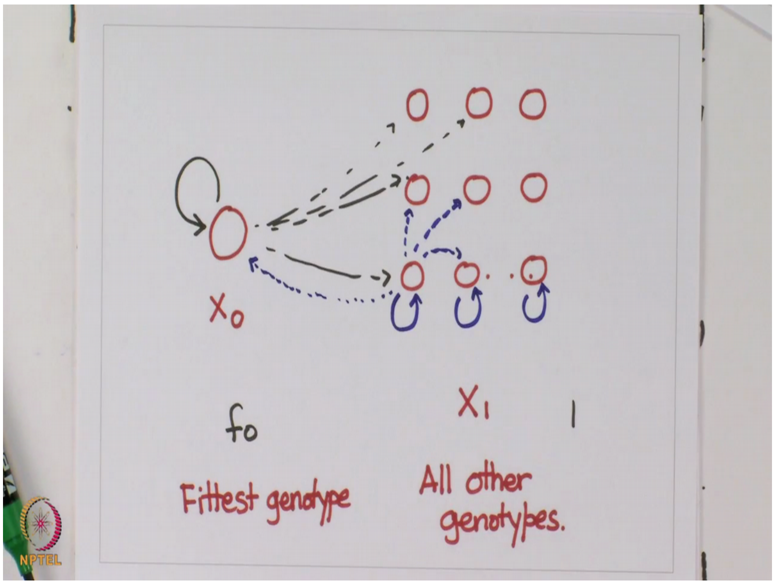
$$\frac{dx_0}{dt} = ? \quad x_0 \rightarrow \text{freq. of individuals of } \underline{\text{all zero}} \text{ genotype}$$

$$\frac{dx_1}{dt} = ? \quad x_1 \rightarrow \text{freq. of all other genotypes.}$$

$$x_0 + x_1 = 1.0$$

And what we are interested in here is coming up with expressions for d x naught by d t and d X 1 by d t what are x naught and x 1. So, these are the relationships that we are interested and deriving x naught is the frequency of individuals; individuals of all 0 genotype which means what fraction of the population belongs to the fittest genotype which is growing at rate f naught and X 1 is frequency of all other genotypes because these 2 quantities X naught and X 1 constitute all the possible genotypes associated with sequence of length L x naught plus X 1 has to be equal to 1. Very quickly we want to understand that can we understand these relationships can we understand these transitions from one particular genotype to another graphically and then come up with quantitative expressions which define these transitions.

(Refer Slide Time: 27:16)



So, let us let us depict this as my x naught pole and this is all the unique genotypes which constitute X 1 right these are X 1 this is the fittest genotype and each node here represents a genotype and together this is all other genotypes. Now how do we think about this what is going to happen as division is happening X naught divides whether enhanced fitness f naught. So, fitness here is f naught fitness here is one very often this is going to divide faster than everybody else and it is going to lead to its own progeny being identical to the parent genotype, but also they are going to be mutational events and it is going to lead to progeny of different genotypes as well.

That is going to happen, but at the same time all of these other nodes are going to replicate and they are going to produce progenies which are identical to themselves, but they will also replicate and produce and make errors during the replication process and lead to progenies of other genotypes in addition it is also possible that a division of replication of this particular genotype leads to a progeny which actually converts to the fittest genotype and goes over to the other pole. And we have one; what we want to do when we continue this in the next lecture is understand these transitions and develop equations which help us quantify these relationships and analyze the steady state associated with the system and in particular what we had started with.

Remember the objective that we started with this whole exercise with is understanding that as mu keeps on increases as mutation rate associated with the fidelity of the DNA polymerize keeps on increases does it always lead to just widening of the fitness widening of the quasi species steady state distribution around the center of the peak or does it lead to a qualitative nature in the distribution of individuals at steady state across various genotypes. So, that is something that we are going to pick up and continue within the next lecture.

Thank you.